Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors

Jan S. Erkelens, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen

***Abstract*—This paper considers techniques for single-channel speech enhancement based on the discrete Fourier transform (DFT). Specifically, we derive minimum mean-square error (MMSE) estimators of speech DFT coefficient magnitudes as well as of complex-valued DFT coefficients based on two classes of generalized gamma distributions, under an additive Gaussian noise assumption. The resulting generalized DFT magnitude estimator has as a special case the existing scheme based on a Rayleigh speech prior, while the complex DFT estimators generalize existing schemes based on Gaussian, Laplacian, and Gamma speech priors. Extensive simulation experiments with speech signals degraded by various additive noise sources verify that significant improvements are possible with the more recent estimators based on super-Gaussian priors. The increase in perceptual evaluation of speech quality (PESQ) over the noisy signals is about 0.5 points for street noise and about 1 point for white noise, nearly independent of input signal-to-noise ratio (SNR). The assumptions made for deriving the complex DFT estimators are less accurate than those for the magnitude estimators, leading to a higher maximum achievable speech quality with the magnitude estimators.**

***Index Terms*—Discrete Fourier transform (DFT)-based speech enhancement, generalized gamma speech priors, minimum mean-square error (MMSE) estimation.**

I. 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., [3]–[9]. Assuming that the noise process is additive and that noise and speech processes are independent, these methods generally attempt to estimate either the underlying noise-free magnitudes of the DFT coefficients, e.g., [4], [5], [9], or the complex-valued DFT coefficients, e.g., [8]. These methods differ in the statistical assumptions with respect to the clean speech DFT coefficients as well as to the noise DFTs; the speech has traditionally been assumed Gaussian, e.g., [4], [5], but recently estimators based on super-Gaussian distribution assumptions such as Laplacian or Gamma distributions have been derived [8]. A similar development has been seen for the noise assumptions; most often the noise is assumed Gaussian, but estimators exist which assume the noise to obey a Laplacian distribution [8]. Finally, existing methods differ in the objective they optimize for; most methods rely on the minimum-mean square error (MMSE) criterion, e.g., [4], [5], but sometimes computationally simpler estimators can be found with the maximum *a posteriori* (MAP) criterion, e.g., [7], [10].

In this paper, we focus on MMSE estimators of the clean speech DFT coefficient magnitudes as well as the complex-valued DFT coefficients. We assume that the noise DFT coefficients obey a (complex) Gaussian distribution as in [8]. For estimation of the speech DFT magnitudes, we assume a one-sided prior of the form

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

where $\Gamma(.)$ is the gamma function. The random variable A represents the DFT magnitude. Fig. 1 shows example densities for $\gamma = 1$ and $\gamma = 2$, respectively. For $\gamma = 2$ and $\nu = 1$, the Rayleigh distribution occurs as a special case. For this prior, the short-time spectral amplitude (STSA) estimator derived in [4] is the MMSE estimator. Also, for the Rayleigh prior, a MAP estimator was derived in [11]. For $\gamma = 2$, a generalized MMSE estimator was derived in [12]. Furthermore, [13] presented a generalized MAP estimator and an online algorithm for estimating the parameters of the generalized prior. For $\gamma = 1$, no generalized MMSE estimator is known in closed form, but a numerical approximation was presented in [12]. An approximate generalized MAP estimator for the case $\gamma = 1$ was derived in [7].

For estimation of the complex-valued speech DFT coefficients, we assume that the real and imaginary parts of the coefficients are statistically independent. We will derive MMSE estimators for a two-sided generalized gamma prior density of the following form:

$$f_{S_R}(s_R) = \frac{\gamma\beta^{\nu}}{2\Gamma(\nu)}|s_R|^{\gamma\nu-1}\exp\left(-\beta|s_R|^{\gamma}\right)$$
$$\beta > 0, \gamma > 0, \nu > 0, -\infty < s_R < \infty \tag{2}$$

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 consider the cases of $\gamma = 1$ and $\gamma = 2$. Examples of the resulting prior densities are shown in Fig. 2. These densities (parameterized by β and ν) contain a number of special cases for which estimators are already known. Specifically, for $\gamma = 2$, the prior parameterizes the Gaussian density ($\nu = 1/2$) for which the Wiener estimator is the MMSE estimator [14]. For $\gamma = 1$, (2) has the Gamma and the Laplacian density as special cases ($\nu = 1/2$ and $\nu = 1$, respectively); MMSE estimators

Manuscript received June 21, 2006; revised March 11, 2007. This work was supported in part by Philips Research, MultimediaN, the Technology Foundation STW (applied Science Division of NWO), and the Technology Programme of the Ministry of Economics Affairs, The Netherlands. Parts of this work have been published in IWAENC 2006 [1] and Interspeech 2006 [2]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yariv Ephraim.

J. Erkelens is with the Information and Communication Theory Group, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: j.s.erkelens@tudelft.nl).

R. C. Hendriks and R. Heusdens are with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands.

J. Jensen is with Oticon A/S, Kongebakken 9, 2765 Smørum, Denmark.

Digital Object Identifier 10.1109/TASL.2007.899233

1558-7916/$25.00 © 2007 IEEE

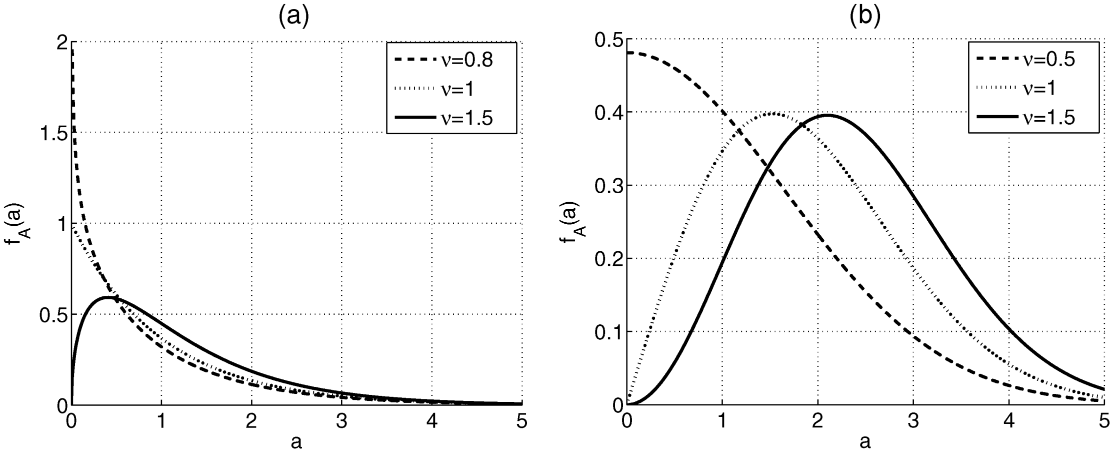


Fig. 1. Prior densities $f_A(a)$ for (a) $\gamma = 1$ with $\nu = \{0.8, 1, 1.5\}$, and for (b) $\gamma = 2$ with $\nu = \{0.5, 1, 1.5\}$. The densities have been normalized to unit variance.

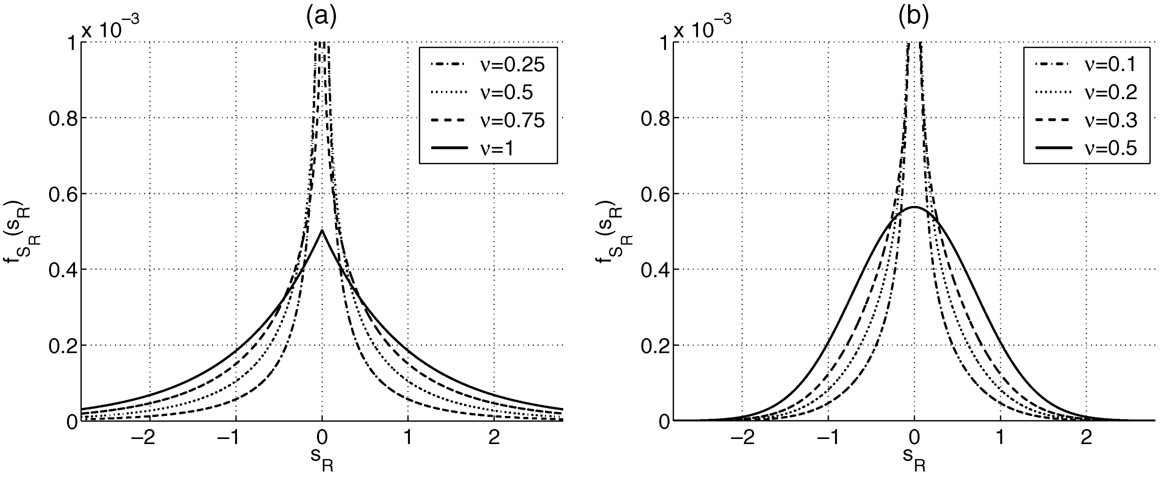


Fig. 2. Prior densities $f_{S_R}(s_R)$ for (a) $\gamma = 1$ with $\nu = \{0.25, 0.5, 0.75, 1\}$, and for (b) $\gamma = 2$ with $\nu = \{0.1, 0.2, 0.3, 0.5\}$. The densities have been normalized to unit variance.

TABLE I

SPECIAL CASES OF THE GENERALIZED PRIORS IN (2) AND (1) FOR WHICH ESTIMATORS ARE KNOWN. FOR ALL ESTIMATORS , THE NOISE IS ASSUMED TO BE ADDITIVE AND GAUSSIAN. * INDICATE ESTIMATORS FOR WHICH NO EXACT CLOSED-FORM SOLUTIONS EXIST

	Complex DFTs ($f_{S_R}(s_R)$)	DFT magnitudes ($f_A(a)$)
$\gamma = 1$	Laplacian (MMSE) [8]	Generalized gamma (MMSE) [12]*
	Gamma (MMSE) [8]	Generalized gamma (MAP) [7]*, [12]*
$\gamma = 2$	Gaussian (MMSE) [14]	Rayleigh (MMSE) [4]
		Rayleigh (MAP) [11]
		Generalized gamma (MMSE/MAP) [12] [13]

for these are derived in [8]. Choosing $\nu > 1.0$ with $\gamma = 1$ or $\nu > 0.5$ with $\gamma = 2$ leads to bimodal priors. Although the estimators derived below remain valid for bimodal priors, we have chosen to restrict ν in our evaluations to the range $0 < \nu \leq 1.0$ for $\gamma = 1$ and $0 < \nu \leq 0.5$ for $\gamma = 2$ to have unimodal priors and thus be better in line with observed speech data, e.g., [8].

Table I summarizes the special cases of (1) and (2) for which estimators (MAP or MMSE) have been documented in the literature.

In this paper, we derive MMSE estimators for the DFT magnitudes assuming that these quantities are drawn from the single-sided generalized gamma prior in (1). We will consider two approximations for the case $\gamma = 1$ for which analytical expressions can be derived, and will combine those approximations into one estimator, which is accurate everywhere and is computationally simpler than approaches based on numerical integration. We also derive MMSE estimators of the complex clean speech DFT coefficients assuming the real and imaginary parts follow the two-sided generalized gamma prior in (2). As mentioned, specific choices of ν and β lead to special cases for which MMSE estimators already exist. However, the derived estimators are more general and therefore cover all possible MMSE estimators (including the ones shown) in each of the quadrants of Table I.

This paper is structured as follows. In Section II, we discuss the validity of the assumptions made in the introduction and look into the consistency between the models of complex DFT coefficients and magnitudes. In Section III, we introduce the signal model and the notation used throughout the paper. Section IV treats MMSE estimation of DFT coefficient magnitudes, while Section V considers MMSE estimators of complex DFT coefficients. Filter characteristics corresponding to the derived estimators are shown in Section VI. In Section VII, we present experimental results. Section VIII concludes the paper.

II. Discussion of the Modeling Assumptions

As outlined in the Introduction, existing DFT coefficient estimators, as well as the ones derived here, rely on a number of assumptions with respect to speech DFT coefficients. In this section, we discuss the consistency and validity of these assumptions. Let us first discuss the probability density functions (pdfs) in (1) and (2). The pdfs contain the parameter β which is related to the speech spectral variance σ_S^2 (See Appendix I). Since in practice σ_S^2 is unknown, it is estimated from the noisy data, e.g., using the decision-directed approach introduced by Ephraim and Malah in [4]. Consequently, the pdfs in (1) and (2) are actually models for the priors faced in practice that are *conditioned* on the *estimated* speech spectral variance, rather than on the true underlying (but unknown) value of σ_S^2. Martin [8] and Lotter and Vary [7] showed that super-Gaussian models of the real and imaginary parts as well as magnitudes of DFT coefficients, conditioned on speech spectral variances estimated by the decision-directed approach, offer a better fit than a Gaussian model. Hence, it is important to notice that the appropriate distributional assumption is related to the speech variance estimator used. For example, Cohen [15] suggests that for a different *a priori* SNR estimator based on GARCH models, the Gaussian speech model is superior. A slight preference for complex Gaussian distributions has also been found for the DFT-coefficients from short analysis frames of individual speech sound classes (vowels, plosives, fricatives, etc.) [16].

The second point concerns the consistency between the models in the complex domain (real and imaginary parts) and the polar domain (amplitude[1] and phase). It is well known that independent and identically distributed (i.i.d.) Gaussian real and imaginary parts correspond to a Rayleigh distribution for the amplitudes that is independent of the uniformly distributed phase. Do i.i.d. generalized-gamma distributed real and imaginary parts lead to generalized-gamma distributed amplitudes? It is not difficult to show that this is indeed the case for the $\gamma = 2$ class of distributions. The value of the parameter ν in the polar domain is then twice as large as the corresponding value in the complex domain. Furthermore, amplitude and phase are also independent, but the phase is not uniformly distributed (except for the Gaussian case, of course). For the $\gamma = 1$ case, these results do not hold in an exact mathematical sense. However, an accurate fit can be made to the resulting amplitude distribution with about the same ratio of two between the ν-parameters in both domains. As in the $\gamma = 2$ case, the resulting phase distribution is generally nonuniform. When we start with a generalized-gamma model in the polar domain, and assume uniformly distributed phase, the corresponding real and imaginary parts are not independent, except for the Gaussian case. However, simulations showed that a fairly accurate fit of the pdf in (2) can still be made to their marginal distributions.

This bring us to perhaps the most important issue: how well do the assumed distributions match actual speech data? Martin [8] and Lotter and Vary [7] have measured the distributions of speech DFT coefficients conditional on a certain narrow range of high values of estimated *a priori* SNR. Contour lines of the measured joint pdf of real an imaginary parts of the DFT coefficients are very nearly circular. A circularly symmetric joint pdf means that the real and imaginary parts are uncorrelated (but, as we shall see, they are not independent), and that the phase distribution is uniform and independent from the amplitude distribution. Therefore, the noisy phase is an optimal estimator for the clean phase [17].

In order to gain further insights, we performed a similar experiment leading to the contour plots of measured histograms of real and imaginary parts shown in Fig. 3. As in Lotter [7], only DFT coefficients have been taken into account for which estimated *a priori* SNR was between 19 and 21 dB. Ephraim and Malah's [4] decision-directed estimator was used for *a priori* SNR estimation. The entire TIMIT-TRAIN database provided the speech material, limited to telephone bandwidth, to which white Gaussian noise at an SNR of 30 dB was added. The noise variance was estimated for each sentence from noise-only segment of 0.64 s, preceding each sentence. Fig. 3(a) shows the contours for the joint distribution; this distribution is very similar to the ones in [7] and [8]. Fig. 3(b) shows the contours for the product of the marginal distributions. This plot is different from Fig. 3(a), and therefore, even though real and imaginary parts may be uncorrelated, there is clearly some dependency between them.

In the derivation of the complex DFT estimators, the real and imaginary parts are assumed independent for mathematical tractability, but this is not entirely in line with measured speech data. Still, we cannot predict beforehand in which domain we will get the best speech enhancement performance, because of the following reason: the parametric distributions of (1) and (2) are only *models* of the actual conditional speech distributions. The fits to measured data are not perfect in either domain, and the derived estimators in each domain may not be equally sensitive to the modeling errors. In this paper, we will investigate the performance of the generalized estimators, and will show experimentally that the amplitude estimators perform slightly better than the complex DFT estimators under the assumed models.

III. Signal Model and Notation

We consider a signal model of the form

$$X(k,m) = S(k,m) + W(k,m) \tag{3}$$

where $X(k,m)$, $S(k,m)$, and $W(k,m)$ are complex-valued random variables representing the DFT coefficients obtained at frequency index k in signal frame m from the noisy speech, clean speech, and noise process, respectively. Applying the

[1] We will use the words *magnitude* and *amplitude* interchangeably. They mean the same, namely the absolute value of a complex DFT coefficient.

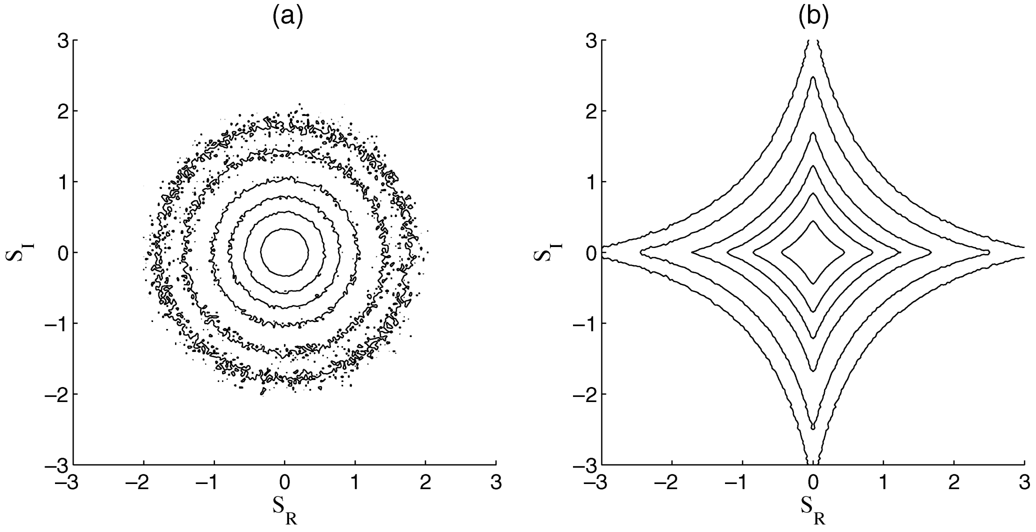


Fig. 3. Contour lines of measured distributions of real and imaginary parts normalized to unit variance. (a) Joint distribution. (b) Product of marginal distributions.

standard assumption that $S(k,m)$ and $W(k,m)$ are statistically independent across time and frequency as well as from each other leads to estimators that are independent of time and frequency. This allows us to increase readability by dropping the time/frequency indices, i.e.,

$$X = S + W \tag{4}$$

and introduce the following notation with respect to real and imaginary parts as well as the modulus of the random variables in question

$$X = X_R + jX_I, \quad |X| = R \tag{5}$$

$$S = S_R + jS_I, \quad |S| = A \tag{6}$$

and

$$W = W_R + jW_I \tag{7}$$

where $j = \sqrt{-1}$. We will use uppercase letters for random variables and the corresponding lowercase letters for their realizations. We assume in this paper that the noise coefficients W obey a Gaussian distribution, with independent and identically distributed real and imaginary parts with

$$\sigma_W^2 = \sigma_{W_R}^2 + \sigma_{W_I}^2, \quad \text{and} \quad \sigma_{W_R}^2 = \sigma_{W_I}^2. \tag{8}$$

We introduce the following SNRs [4]: *a priori* SNR $\xi = \sigma_S^2/\sigma_W^2$ and *a posteriori* SNR $\zeta = |x|^2/\sigma_W^2$, where we used lower-case x to represent a realization of the random variable X. We also define $\zeta_R = x_R^2/\sigma_{W_R}^2$, and $\zeta_I = x_I^2/\sigma_{W_I}^2$. From (8) we have $\zeta = (\zeta_R + \zeta_I)/2$.

IV. MMSE Estimation of Magnitudes of DFT Coefficients

In this section, we derive MMSE estimators of the magnitude of the clean speech DFT coefficients. The MMSE estimator is identical to the conditional mean [18] given by

$$E\{A|r\} = \frac{\int_0^\infty a f_{R|A}(r|a) f_A(a) da}{\int_0^\infty f_{R|A}(r|a) f_A(a) da}. \tag{9}$$

Further, since the noise is assumed to be Gaussian, $f_{R|A}(r|a)$ can be written as [19]

$$f_{R|A}(r|a) = \frac{2r}{\sigma_W^2} \exp\left(-\frac{r^2+a^2}{\sigma_W^2}\right) I_0\left(\frac{2ar}{\sigma_W^2}\right) \tag{10}$$

where I_0 is the 0th-order modified Bessel function of the first kind. No assumption about the clean speech phase distribution has to be made to derive this expression.

We will consider the case $\gamma = 2$ first, because the corresponding estimator can be derived without any approximations.

A. DFT Magnitudes, $\gamma = 2$

Inserting (1) with $\gamma = 2$ and (10) into (9) gives

$$\hat{A}^{(2)} = \frac{\int_0^\infty a^{2\nu} \exp\left(-\frac{a^2}{\sigma_W^2} - \beta a^2\right) I_0\left(\frac{2ar}{\sigma_W^2}\right) da}{\int_0^\infty a^{2\nu-1} \exp\left(-\frac{a^2}{\sigma_W^2} - \beta a^2\right) I_0\left(\frac{2ar}{\sigma_W^2}\right) da} \tag{11}$$

where the superscript (2) emphasizes that $\gamma = 2$.

Using [20, Eqs. 6.643.2, 9.210.1, and 9.220.2] we can solve the integrals for $\nu > 0$ and find

$$\hat{A}^{(2)} = \frac{\Gamma(\nu+0.5)}{\Gamma(\nu)} \sqrt{\frac{\xi}{\zeta(\nu+\xi)}} \frac{{}_1F_1\left(\nu+0.5; 1; \frac{\zeta\xi}{\nu+\xi}\right)}{{}_1F_1\left(\nu; 1; \frac{\zeta\xi}{\nu+\xi}\right)} r \tag{12}$$

where ${}_1F_1(a;b;x)$ is the confluent hypergeometric function [21, Ch. 13], and where we have inserted in (12) the relation between β and $E\{A^2\}$ given in (32). This result has also recently been derived in [12].

B. DFT Magnitudes, $\gamma = 1$

For $\gamma = 1$ the expression for the amplitude estimator becomes

$$E\{A|r\} = \frac{\int_0^\infty a^{\nu} \exp\left(-\frac{a^2}{\sigma_W^2} - \beta a\right) I_0\left(\frac{2ar}{\sigma_W^2}\right) da}{\int_0^\infty a^{\nu-1} \exp\left(-\frac{a^2}{\sigma_W^2} - \beta a\right) I_0\left(\frac{2ar}{\sigma_W^2}\right) da}. \tag{13}$$

Unfortunately the integrals do not appear to have closed-form solutions, which leads us to introduce two approximations of (13), one of which is most accurate under low SNR conditions,

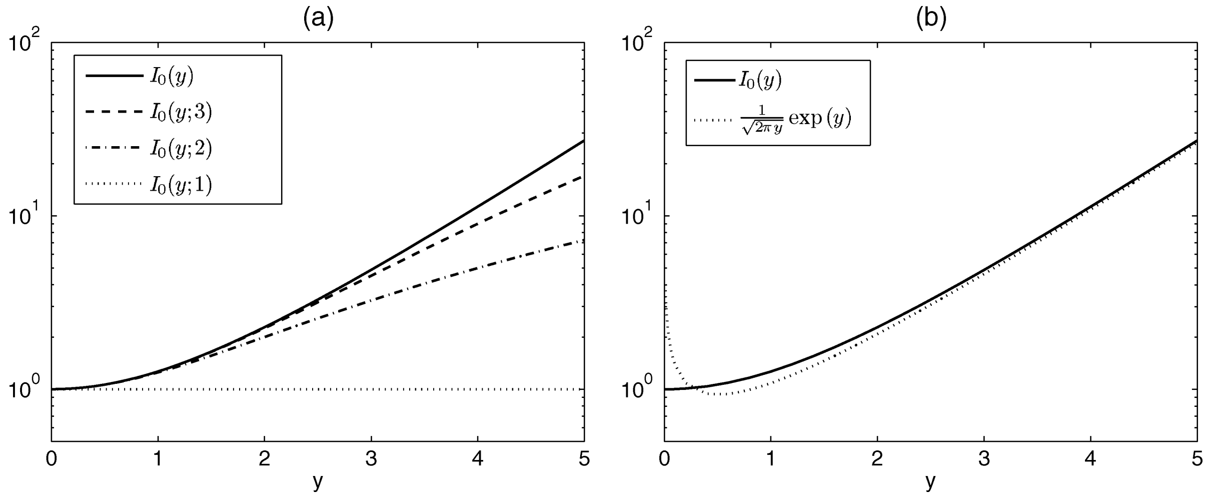


Fig. 4. I_0 and its approximations. (a) Taylor series expansion for small arguments. (b) Approximation for large arguments.

while the other is closer to (13) for high SNRs. With these approximations, the integrals can be solved in closed-form. Before discussing the approximations, we introduce a change of variable that makes it clearer to see under what conditions the various approximations are expected to be accurate.

1) Change of Variable: For convenience the following transformation is made: $y = 2ar/\sigma_W^2$. In addition, we make use of the relation in (31) between β and the second moment of A, $\beta = \sqrt{\nu(\nu+1)}/\sigma_S$. The expression for $\hat{A}^{(1)}$ now becomes

$$\hat{A}^{(1)} = \frac{\sigma_W^2}{2r} \frac{\int_0^\infty y^\nu \exp\left[-\frac{y^2}{4\zeta} - \frac{\mu y}{2\sqrt{\zeta\xi}}\right] I_0(y) dy}{\int_0^\infty y^{\nu-1} \exp\left[-\frac{y^2}{4\zeta} - \frac{\mu y}{2\sqrt{\zeta\xi}}\right] I_0(y) dy} \qquad (14)$$

where we have introduced $\mu = \sqrt{\nu(\nu+1)}$. The approximations of (14) discussed below concern the Bessel function. The function $y^\nu \exp[-(y^2/4\zeta) - (\mu y/2\sqrt{\zeta\xi})]$ attains its maximum at a small value of y when the exponentials decay fast and y^ν rises slowly. In this case, it is especially important to approximate the Bessel function well at small arguments. This happens when ζ and/or $\sqrt{\zeta\xi}$ are small, i.e., at low SNRs. Note that ζ is the more dominant parameter of ζ and ξ, because ξ is not present in the quadratic term in the exponentials. In other cases, namely high SNR conditions, the Bessel function should be accurately approximated for large arguments.

2) Approximation of $E\{A|r\}$, Low SNR Conditions: For low SNR conditions, we approximate I_0 by a Taylor series expansion around $y = 0$. The Taylor series of I_0, truncated after K terms, is given by [21, Eq. 9.6.10]

$$I_0(y;K) = \sum_{k=0}^{K-1} \left(\frac{y}{2}\right)^{2k} \frac{1}{(k!)^2}. \qquad (15)$$

Fig. 4(a) shows I_0 and several truncated Taylor series expansions. We see that for small arguments, I_0 is approximated well by only a few terms. Substituting (15) into (14) and using [20, Eq. 3.462.1] gives us an estimator, $\hat{A}^{(1)}_{\ll,K}$, which is most accurate for low SNRs

$$\hat{A}^{(1)}_{\ll,K} = \frac{1}{\sqrt{2\zeta}} \frac{\sum\limits_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{\zeta}{2}\right)^k \Gamma(\nu+2k+1) D_{-(\nu+2k+1)}\left(\frac{\mu}{\sqrt{2\xi}}\right)}{\sum\limits_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{\zeta}{2}\right)^k \Gamma(\nu+2k) D_{-(\nu+2k)}\left(\frac{\mu}{\sqrt{2\xi}}\right)} r \qquad (16)$$

where $D_\nu(\cdot)$ is a special function, called the parabolic cylinder function of order ν [21, Ch. 19]. The superscript (1) stresses that $\gamma = 1$.

For $K \to \infty$, $\hat{A}^{(1)}_{\ll,K}$ converges to $\hat{A}^{(1)}$. This is because the Taylor expansion in (15) converges $\forall y$ and because changing the order of integration and summation as is used in the derivation of (16) is allowed for $K \to \infty$ according to Fubini's theorem [22].

3) Approximation of $E\{A|r\}$, High SNR Conditions: Using the approximate estimator $\hat{A}^{(1)}_{\ll,K}$ under high SNR conditions requires K to be large for an accurate result. This leads to a high computational load and numerical problems may result when evaluating $(1/k!)^2$, the Γ-functions, and the parabolic cylinder functions. In order to avoid these complications, we investigate an approximation of (14) that is more accurate under high SNR conditions. We apply the following well-known large-argument approximation of I_0 [21, Eq. 9.7.1]:

$$I_0(y) \sim \frac{1}{\sqrt{2\pi y}} \exp(y). \qquad (17)$$

Fig. 4(b) shows I_0 and its approximation for large arguments. Substituting this approximation in (14) and using [20, Eq. 3.462.1], we find for $\nu > 0.5$

$$\hat{A}^{(1)}_{\gg} = \frac{(\nu - 1/2)}{\sqrt{2\zeta}} \frac{D_{-(\nu+1/2)}\left(\frac{\mu}{\sqrt{2\xi}} - \sqrt{2\zeta}\right)}{D_{-(\nu-1/2)}\left(\frac{\mu}{\sqrt{2\xi}} - \sqrt{2\zeta}\right)} r. \qquad (18)$$

The approximation in (18) is most accurate when ζ and $\sqrt{\zeta\xi}$ are large and ν is large too.

C. Combining Estimators

To get the best performance with these approximations a procedure is needed that decides which approximation to use under what circumstances. One could evaluate (14) numerically whenever possible, and use $\hat{A}^{(1)}_{\gg}$ for the largest values of ζ. This is a computationally complex procedure. An alternative is to look for a simple strategy for picking one of the two estimators, depending on the values of the parameters.

The faster the exponential term in the integrals in (14) decreases, the less important it becomes how well the Bessel function is approximated for large values of y. So generally speaking, the approximation for small arguments is most accurate for low SNRs. The approximation (18) is more accurate for high SNRs and large ν. Fortunately, the behavior of the approximations is such that a simple binary decision strategy can be found that leads to good results. Specifically, it turns out that the maximum of (16) and (18) is generally a good approximation of $\hat{A}^{(1)}$ for, say, $K > 4$. This procedure is motivated as follows. It can be proven (see Appendix II) that the approximation for low SNRs $\hat{A}^{(1)}_{\ll,K}$ is always smaller than $\hat{A}^{(1)}$ for all K. The approximation for high SNRs $\hat{A}^{(1)}_{\gg}$ can be both smaller and larger than $\hat{A}^{(1)}$ depending on the values of the parameters. As we will show by simulation experiments, however, it can be only slightly larger than $\hat{A}^{(1)}$. It can be much smaller though, but that happens for low SNRs. It turns out that the combined estimator $\hat{A}^{(1)}_{C,K} = \max[\hat{A}^{(1)}_{\ll,K}, \hat{A}^{(1)}_{\gg}]$ obtained from a simple binary decision leads to an accurate approximation of $\hat{A}^{(1)}$ for all values of ζ, ξ, and ν. To illustrate the practical use of the decision rule $G^{(1)}_{C,K} = \max[G^{(1)}_{\ll,K}, G^{(1)}_{\gg}]$, we refer to Fig. 5, where gain curves versus the *a posteriori* SNR ζ are shown for $\nu = 0.6$ and for several values of ξ. In each plot, we show $G^{(1)}_{\ll,5}$, $G^{(1)}_{\gg}$, and $G^{(1)}_{\mathrm{MMSE}}$ and see that taking the maximum of $G^{(1)}_{\ll,5}$ and $G^{(1)}_{\gg}$ leads to a gain $G^{(1)}_{C,5}$ close to $G^{(1)}_{\mathrm{MMSE}}$. $G^{(1)}_{\mathrm{MMSE}}$ was calculated by numerical integration of (14).

D. Experimental Analysis of Errors Due to Approximations

The errors in the combined estimator have been investigated for the range $0.6 \leqslant \nu \leqslant 3.2$, -20 dB $\leqslant \xi \leqslant +20$ dB, -20 dB $\leqslant \zeta \leqslant +14$ dB. For this range, $G^{(1)}_{\mathrm{MMSE}}$ could be evaluated numerically. For larger ζ, the accuracy of $\hat{A}^{(1)}_{\gg}$ only increases, so its error will be smaller. For the binary decision $\max[G^{(1)}_{\ll,5}, G^{(1)}_{\gg}]$, the maximum positive error was $+3.7$ dB, and the maximum negative error was -0.2 dB. A positive error means that $G^{(1)}_{\mathrm{MMSE}}$ was larger than the approximate gain function. Using $\max[G^{(1)}_{\ll,20}, G^{(1)}_{\gg}]$ the maximum positive error decreases to $+0.1$ dB. The largest positive errors occur for the lowest values of ν, and, for a given value of *a priori* SNR, only for a small range of *a posteriori* SNRs, as can be seen in Fig. 5. These results show that the simple binary decision procedure generally works well. Furthermore, although $G^{(1)}_{\gg}$ can be larger than $G^{(1)}_{\mathrm{MMSE}}$, it will not exceed it by more than 0.2 dB for the parameter range of interest. We will evaluate the combined gain function $G^{(1)}_{C,5} = \max[G^{(1)}_{\ll,5}, G^{(1)}_{\gg}]$ in the following sections.

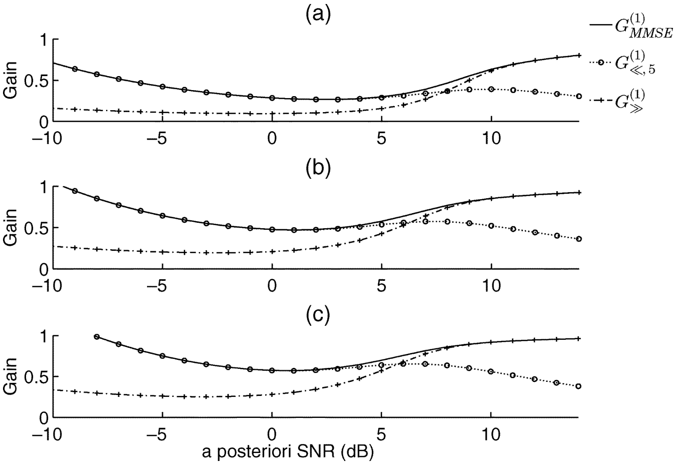


Fig. 5. Comparison of gain functions for amplitude estimators $\hat{A}^{(1)}_{\ll,5}$ Eq. (16), $A^{(1)}_{\gg}$ Eq. (18), and $\hat{A}^{(1)}$ for $\nu = 0.6$ and (a) $\xi = -5$ dB, (b) $\xi = +5$ dB, and (c) $\xi = +15$ dB.

E. Computational Complexity

MATLAB implementations of the algorithms in [23] have been used to evaluate the special functions (available at http://ceta.mit.edu/comp_spec_func/). We have adapted these programs so that they can handle vector arguments. The estimator for $\gamma = 2$ can thus be evaluated in real-time on a modern PC. The combined estimator for $\gamma = 1$ is more complex, because two estimators have to be evaluated and because of the sums in (16). The sums can be efficiently computed by making use of recursive relations [21, Eq. 19.6.4]. In this way, the combined estimator can be evaluated in 2-3 times real-time. In a practical system, for a fixed value of ν, all gain functions can be evaluated offline for the relevant range of the parameters and stored in a table. Computational complexity is not an issue then.

V. MMSE Estimation of Complex DFT Coefficients

In this section, we derive the MMSE estimator of the clean speech DFT coefficient S. Assuming as in [8] that the real and imaginary parts of S, S_R, and S_I are statistically independent, it follows

$$E\{S|x\} = E\{S_R|x_R\} + jE\{S_I|x_I\}. \quad (19)$$

We now consider estimation of S_R; a similar procedure can be followed for S_I. We have

$$E\{S_R|x_R\} = \frac{\int_{s_R} s_R f_{X_R|s_R}(x_R|s_R) f_{S_R}(s_R) ds_R}{\int_{s_R} f_{X_R|s_R}(x_R|s_R) f_{S_R}(s_R) ds_R}. \quad (20)$$

Using the Gaussian noise assumption it follows

$$f_{X_R|s_R}(x_R|s_R) = \left(2\pi\sigma^2_{W_R}\right)^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2\sigma^2_{W_R}}\left(x_R^2 + s_R^2 - 2x_R s_R\right)\right). \quad (21)$$

A. Complex DFTS, $\gamma = 1$

Using (2) with $\gamma = 1$, [20, Eq. 3.462.1] and the relation between β and $\sigma^2_{S_R}$ given by (33) in Appendix I, we find the following expression for the conditional mean [2], [24]

$$E\{S_R|x_R\} = \sigma_{W_R}\nu \frac{\exp\left(\frac{1}{4}x_-^2\right)D_{-(\nu+1)}(x_-) - \exp\left(\frac{1}{4}x_+^2\right)D_{-(\nu+1)}(x_+)}{\exp\left(\frac{1}{4}x_-^2\right)D_{-\nu}(x_-) + \exp\left(\frac{1}{4}x_+^2\right)D_{-\nu}(x_+)} \qquad (22)$$

where

$$x_\pm = \sqrt{\frac{\nu(\nu+1)}{\xi}} \pm \frac{x_R}{\sigma_{W_R}}. \qquad (23)$$

B. Complex DFTS: $\gamma = 2$

We now consider the MMSE estimator for $\gamma = 2$. Maintaining the Gaussian noise assumption, the conditional density $f_{X_R|s_R}(x_R|s_R)$ given in (21) remains valid. Using (2) with $\gamma = 2$ and [20, Eq. 3.462.1], it can be shown that the conditional mean estimator can be written as

$$E\{S_R|x_R\} = 2\nu \frac{\sigma_{W_R}}{\sqrt{1+2\nu\xi^{-1}}} \times \frac{D_{-(2\nu+1)}(x_-) - D_{-(2\nu+1)}(-x_-)}{D_{-2\nu}(x_-) + D_{-2\nu}(-x_-)} \qquad (24)$$

where

$$x_- = -\frac{x_R}{\sigma_{W_R}}\left(1 + 2\nu\xi^{-1}\right)^{-1/2}. \qquad (25)$$

It is easy to see that when ν/ξ is small and ζ_R is not, the estimators for $\gamma = 1$ (22) and $\gamma = 2$ (24) are approximately equal when the quantity $\gamma\nu$ has the same value for both estimators.

VI. Filter Characteristics

In this section, we study the input–output characteristics for the DFT magnitude estimators as well as for the complex DFT coefficient estimators.

A. Magnitudes of DFT Coefficients

Fig. 6 shows examples of input–output characteristics for the magnitude estimators. In Fig. 6(a), we consider the case $\gamma = 1$ and $\nu \in \{0.8, 1, 1.5\}$ for the combined estimator $\hat{A}^{(1)}_C = \max[\hat{A}^{(1)}_{\ll,5}, \hat{A}^{(1)}_{\gg}]$. In Fig. 6(b), we consider the case $\gamma = 2$ for $\nu \in \{0.5, 1, 1.5\}$. Further, the constraint $\sigma^2_S + \sigma^2_W = 2$ is used, and we consider *a priori* SNRs $\xi = -5$ dB and $\xi = 5$ dB. The input–output characteristics are more sensitive to ν values than for the $\gamma = 1$ case, and a smaller ν value clearly leads to less suppression at higher input values and to more suppression for lower input values.

B. Complex DFT Coefficients

Fig. 7(a) shows examples of input–output characteristics of the complex DFT estimators for the case of $\gamma = 1$ and $\nu \in \{0.25, 0.50, 0.75, 1\}$. For $\nu = 1.0$, we recognize the input–output characteristic of the Laplacian (two-sided exponential) prior and for $\nu = 0.5$ we get the input–output characteristics of the two-sided Gamma distribution. 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 nonlinear, especially for small values of ν, i.e., more peaked priors.

For the $\gamma = 2$ case, we consider $\nu \in \{0.1, 0.2, 0.3, 0.5\}$. Choosing $\nu = 0.5$ gives a Gaussian prior, while lower values of ν correspond to more peaked distributions. Fig. 7(b) shows input–output characteristics for the resulting MMSE estimators. For $\nu = 0.5$, the Wiener estimator occurs [solid line in Fig. 7(b)]. For all other choices of ν, the estimators are nonlinear in the noisy observation x_R.

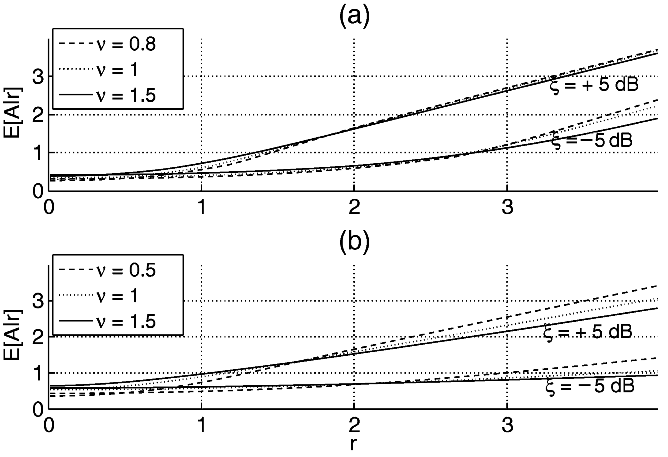


Fig. 6. Input–output characteristics for DFT magnitude estimators ($\sigma^2_S + \sigma^2_W = 2$). (a) $\gamma = 1$. (b) $\gamma = 2$.

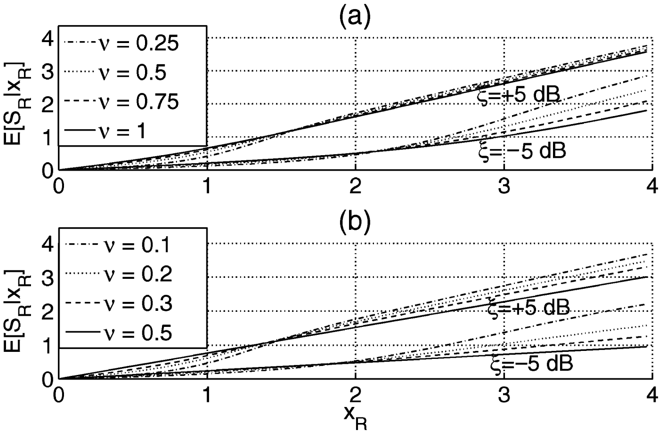


Fig. 7. Input–output characteristics for complex DFT estimators ($\sigma^2_S + \sigma^2_W = 2$). (a) $\gamma = 1$. (b) $\gamma = 2$.

VII. Experimental Results

In this section, we present experimental results obtained with the complex DFT and magnitude estimators. For the experiments, we use the Noizeus database [25], which consists of 30 IRS-filtered speech signals sampled at 8 kHz, contaminated by various additive noise sources. We added computer-generated telephone-bandwidth white Gaussian noise as an extra noise source, since it is not present in the data base. The frame size is

256 samples, with an overlap of 50%. The decision-directed approach with a smoothing factor $\alpha = 0.98$ was used to estimate ξ [4]. The noise variance was estimated with the minimum statistics approach [26]. Further, in all experiments the maximum suppression was limited to 0.1, for perceptual reasons [8]. Experiments with a lower limit did not change the conclusions.

A. Objective Quality Measures

We measure the performance of the proposed estimators using a range of objective speech quality measures. First, we introduce the squared error distortion measures

$$D_{\mathrm{ampl}} = \sum_{(k,m)\in\mathcal{Q}} \left(A(k,m) - \hat{A}(k,m)\right)^2 \tag{26}$$

and

$$D_{\mathrm{DFT}} = \sum_{(k,m)\in\mathcal{Q}} \left|S(k,m) - \hat{S}(k,m)\right|^2 \tag{27}$$

for the magnitude and complex DFT estimators, respectively. Our estimators assume speech presence. In order to avoid contamination of our experimental results by noise-only regions, we discard nonspeech frequency bins by using an index set $\mathcal{Q}$ denoting the DFT bins with energy no less than 50 dB below the maximum bin energy in the particular speech signal. These distortion measures evaluate the quantities for which the estimators are optimized.

In an attempt to express the objective performance of the estimators in terms of speech distortion and noise reduction separately, we follow the approach in [7] and define segmental speech SNR as

$$\mathrm{SNR{-}S} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} 10\log_{10}\left(\frac{||\mathbf{s}_p||_2^2}{||\mathbf{s}_p - \tilde{\mathbf{s}}_p||_2^2}\right) \tag{28}$$

where the vector $\mathbf{s}_p$ represents a clean speech (time-domain) frame, and $\tilde{\mathbf{s}}_p$ is the result of applying the gain functions to the clean speech frame. To discard nonspeech frames, an index set $\mathcal{P}$ is used of all clean speech frames with energy within 30 dB of the maximum frame energy in a particular speech signal. $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$. Similarly, noise reduction is measured as

$$\mathrm{SNR{-}N} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} 10\log_{10}\left(\frac{||\mathbf{w}_p||_2^2}{||\tilde{\mathbf{w}}_p||_2^2}\right) \tag{29}$$

where $\mathbf{w}_p$ is a noise frame, and $\tilde{\mathbf{w}}_p$ is the residual noise frame resulting from applying the noise suppression filter to $\mathbf{w}_p$.

Further, we use segmental SNR defined as

$$\mathrm{SNR_{seg}} = \frac{1}{M}\sum_{i=1}^{M} \mathcal{T}\left(10\log_{10}\frac{||\mathbf{s}_p||^2}{||\mathbf{s}_p - \hat{\mathbf{s}}_p||^2}\right) \tag{30}$$

where $\hat{\mathbf{s}}_p$ denotes an enhanced signal frame, M is the total number of frames, and $\mathcal{T}[y] = \max(\min(y, 35), -10)$, confining the local SNR to a perceptual meaningful range [27]. Finally, we apply the PESQ speech quality measure [28].

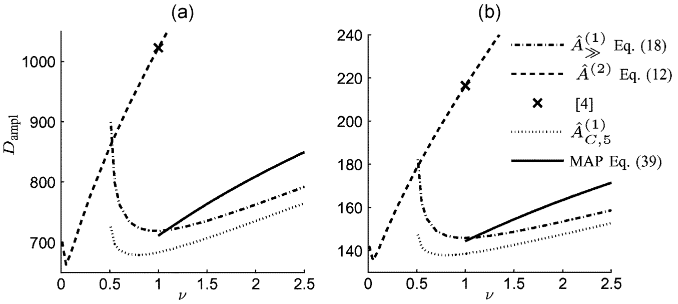


Fig. 8. Measured squared error D_{ampl} for white noise with (a) SNR = 0 dB. (b) SNR = 10 dB.

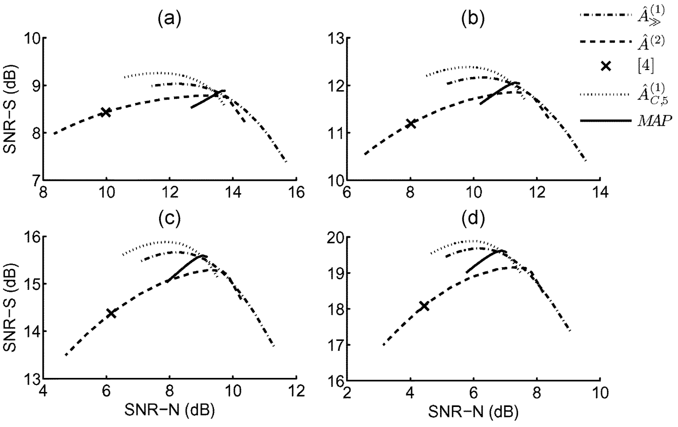


Fig. 9. SNR-S plotted versus SNR-N while varying ν for $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, the MAP estimator and the estimator of [4] for (a) input SNR = 0 dB, (b) SNR = 5 dB, (c) SNR = 10 dB, (d) SNR = 15 dB. ν decreases along the curves from the left to the right. The range of ν values is the same as for Fig. 8.

B. Magnitude Estimators

We evaluate the performance of the MMSE amplitude estimator $\hat{A}^{(2)}$ and two approximations of $\hat{A}^{(1)}$, namely $\hat{A}^{(1)}_{\gg}$ and $\hat{A}^{(1)}_{C,5} = \max[\hat{A}^{(1)}_{\ll,5}, \hat{A}^{(1)}_{\gg}]$; we included $\hat{A}^{(1)}_{\gg}$ in this comparison to show that the combined estimator $\hat{A}^{(1)}_{C,K}$ has clear advantages over using just the well-known high SNR approximation used for $\hat{A}^{(1)}_{\gg}$. Further, we make a comparison to a modification of the MAP amplitude estimator as presented in [7], which is in fact a MAP estimator under the generalized gamma distribution (1) with $\gamma = 1$. Details on the modified MAP estimator, which we refer to as $\hat{A}^{(1)}_{\mathrm{MAP}}$, can be found in Appendix III.

Fig. 8 plots D_{ampl} versus ν. We see that $\hat{A}^{(1)}_{C,5}$ improves over $\hat{A}^{(1)}_{\gg}$ and $\hat{A}^{(1)}_{\mathrm{MAP}}$, and that $\hat{A}^{(2)}$ does very well for $\nu \approx 0.1$.

Fig. 9 shows performance in terms of SNR-S versus SNR-N for several values of ν and speech signals degraded by white noise. It is shown that for a fixed SNR-N performance, $\hat{A}^{(1)}_{C,5}$ often leads to the best speech quality. Furthermore, we see that $\hat{A}^{(2)}$ has the worst SNR-S versus SNR-N tradeoff.

In Fig. 10, an evaluation in terms of segmental SNR versus ν is shown for the input SNRs of 5 and 15 dB and speech signals degraded with street noise and white noise. The estimators $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, and $\hat{A}^{(1)}_{\mathrm{MAP}}$ have a comparable performance and are relatively insensitive to ν. The estimator $\hat{A}^{(2)}$ is much more sensitive to ν and shows maximum performance at $\nu \approx 0.1$.

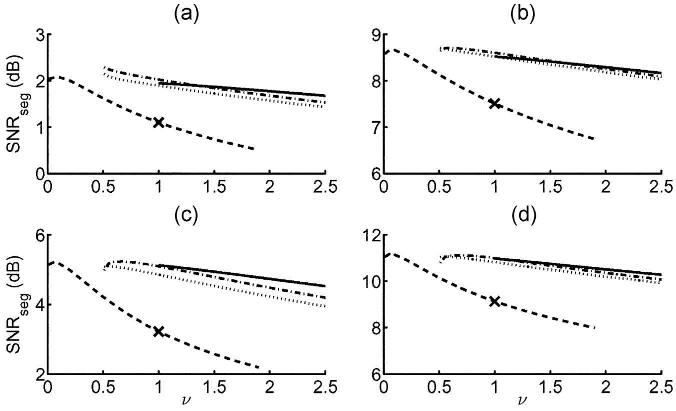


Fig. 10. Performance in terms of SNR$_{seg}$ versus ν for the MAP estimator, $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, and the estimator of [4] for (a) street noise at input SNR $=$ 5 dB, (b) street noise, SNR $= 15$ dB, (c) white noise, SNR $= 5$ dB, and (d) white noise, SNR $= 15$ dB. The legend of Fig. 8 applies.

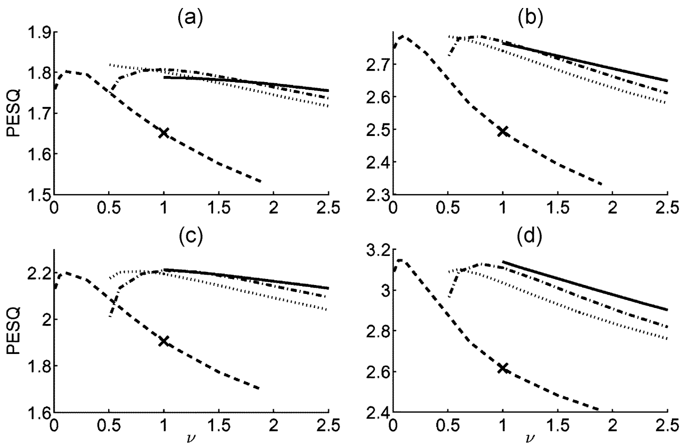


Fig. 11. Performance in terms of PESQ (MOS) versus ν for the MAP estimator, $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, and the estimator of [4] for (a) street noise at input SNR $=$ 5 dB, (b) street noise, SNR $= 15$ dB, (c) white noise, SNR $= 5$ dB, (d) white noise, SNR $= 15$ dB. The legend of Fig. 8 applies.

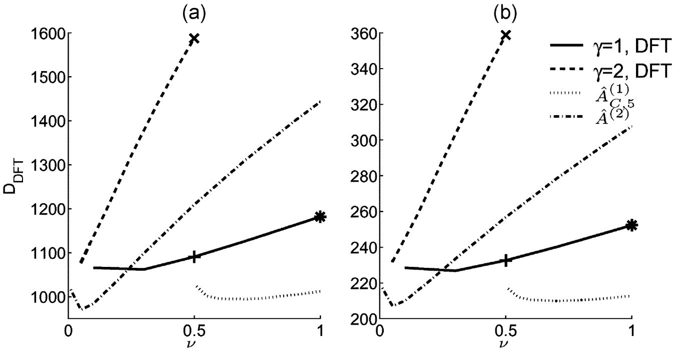


Fig. 12. D_{DFT} evaluated on complex DFT estimators with $\gamma = 1$ and $\gamma = 2$ and the amplitude estimators $\hat{A}^{(1)}_{C,5}$ and $\hat{A}^{(2)}$ for white noise with (a) SNR $=$ 0 dB, and (b) SNR $= 10$ dB. The special cases that correspond to the Gamma, Laplace, and Gaussian priors are indicated by $+$, $*$, and $\times$, respectively.

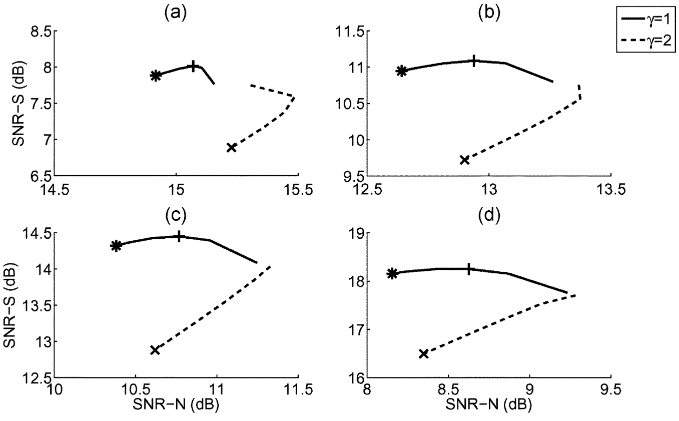


Fig. 13. SNR-S plotted versus SNR-N while varying ν for the complex DFT estimators for $\gamma = 1$ and $\gamma = 2$ for white noise. (a) Input SNR $= 0$ dB. (b) SNR $= 5$ dB. (c) SNR $= 10$ dB. (d) SNR $= 15$ dB. ν decreases along the curves from the left to the right. The range of ν values is the same as for Fig. 12.

The maximum performance of all four estimators $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(1)}_{\mathrm{MAP}}$, and $\hat{A}^{(2)}$ is approximately the same.

Fig. 11 plots PESQ versus ν for the input SNRs of 5 and 15 dB and speech signals degraded with street noise and white noise.

C. Complex DFT Estimators

We first consider the squared error distortion measure D_{DFT}. Fig. 12 plots D_{DFT} versus ν. The estimator based on a Gamma prior $(+)$ performs well, but choosing $\nu \approx 0.3$ leads to slightly better performance. In Fig. 12 the amplitude estimators $\hat{A}^{(1)}_{C,5}$ and $\hat{A}^{(2)}$ are evaluated with D_{DFT} as well. Although counter-intuitive, it shows that the amplitude estimators perform better as measured by the D_{DFT} distortion measure than the complex DFT estimators. This indicates that the underlying model assumptions for the complex DFT estimators are less valid for natural speech than those of the amplitude estimators.

Fig. 13 shows performance in terms of SNR-S versus SNR-N as a function of ν for the complex DFT estimators and speech signals degraded with 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 (x) provides the weakest SNR-S versus SNR-N tradeoff; as discussed in Section II, this suggests that the speech distribution *conditional* on the estimated *a priori* SNR is not well described by a Gaussian model. The Gaussian model and thus the Wiener estimator may perform better for a different *a priori* SNR estimator [15]. Also, rather simple modifications of the Wiener estimator have been proposed which significantly boost its performance (see, e.g., [29] and [30, Ch. 2]). As expected [see comment after (25)], choosing low ν values leads to similar performance of both classes $\gamma = 1$ and $\gamma = 2$. Comparing with Fig. 9, we see that the maximum achievable speech quality in terms of SNR-S is lower than for the amplitude estimators.

D. Subjective Evaluation

Andrianakis and White [12] report that the MMSE estimators generally perform better than their MAP counterparts, in the sense that weaker speech spectral components are better preserved, while the residual noise has a much more broadband character. The better preservation of weak speech components is confirmed by our informal listening, although the differences between the various estimators are generally small, partly because the maximum suppression was limited to 0.1 for all methods. The combined MMSE estimator $\hat{A}^{(1)}_{C,5}$ introduces slightly less speech distortions than the MAP estimator, $\hat{A}^{(1)}_{\gg}$, and $\hat{A}^{(2)}$. The complex DFT estimators appear to give better noise suppression and seem to introduce slightly less noise artifacts than the amplitude estimators, albeit at the cost of somewhat higher speech distortions (see also Figs. 9 and 13). Informal listening tests are in line with the objective results presented above concerning the influence of the parameter ν. For $\gamma = 1$, the perceived quality was rather insensitive to adjustments of ν, while for $\gamma = 2$ changes in ν had a bigger effect: for large values of ν, more residual noise but less musical noise was present as compared to the smallest values of ν.

VIII. Concluding Remarks

This paper considered DFT-based techniques for single-channel speech enhancement. In the first part, we extended existing MMSE estimators of the *magnitude* estimators for DFT-based noise suppression. The optimal estimators are found under a one-sided generalized Gamma distribution, which takes as special cases (different parameter settings) all priors used in known noise suppression schemes so far. Deriving the MMSE estimators involves integration of (weighted) Bessel functions. In order to find analytical solutions, for some parameter settings approximations were necessary. Ultimately, we combined two types of Bessel function approximations using a simple binary decision between the two. We showed by computer simulations that the estimator thus obtained is very close to the exact MMSE estimator for all SNR conditions. The presented estimators lead to improved performance compared to the suppression rule proposed by Ephraim and Malah [4]. Furthermore, the maximum possible performance is slightly better than that of state-of-the-art MAP amplitude estimators.

The second part of the paper dealt with MMSE estimators of *complex* DFT coefficients by deriving two classes of estimators based on generalized gamma prior pdfs. Estimators from the class $\gamma = 1$ typically perform better than the $\gamma = 2$ class, except for small values of the parameter ν, where the estimators are very similar. Applying a complex Gaussian model assumption for the complex speech DFT coefficients clearly leads to suboptimal results. The amplitude estimators performed better than the complex DFT estimators, even under the DFT distortion measure because the modeling assumptions in the complex domain are less accurate than those in the polar domain.

Super-Gaussian priors have been proposed in the literature because they fit better to measured distributions than the Gaussian/Rayleigh priors. These measured distributions are conditional on the estimated spectral variance parameters. However, the super-Gaussian priors still do not perfectly match the measured distributions. The independency assumption in the complex domain is also inconsistent with the data. We think further improvements in speech enhancement performance are still possible by considering more sophisticated pdf models and better spectral variance estimators [15], [31], [32] simultaneously. For certain types of distortions, other methods may be more appropriate. For example, for very nonstationary or impulsive noise sources, it may be very hard to estimate the signal variance parameters. The signal-to-noise ratio may also be very low, so that the signal is essentially lost in small time intervals. For these kind of disturbances, signal reconstruction techniques based on, e.g., time-frequency interpolation [10], may possibly lead to more satisfying results.

Appendix I
Second Moments

In this appendix, we derive expressions for the second moments of the random variables with densities $f_A(a)$ and $f_{S_R}(s_R)$ as given by (1) and (2) for the cases $\gamma = 1$ and $\gamma = 2$.

A. Single-Sided Prior $f_A(a)$

1) The Second Moment of A for the Case $\gamma = 1$: With [20, Eq. 3.381.4], it is straight forward to verify that $E\{A^2\}$ is given by

$$\sigma_S^2 = \frac{\nu(\nu+1)}{\beta^2}. \tag{31}$$

2) Second Moment of A for the Case $\gamma = 2$: One can show with the substitution $a = \sqrt{t}$ and [20, Eq. 3.381.4] that $E\{A^2\}$ is given by

$$\sigma_S^2 = \frac{\nu}{\beta}. \tag{32}$$

B. Two-Sided Prior $f_{S_R}(s_R)$

1) Second Moment of S_R for $\gamma = 1$: Clearly, the mean of the random variable S_R in question is zero, $E\{S_R\} = 0$, since the distribution $f_{S_R}(s_R)$ is symmetric around 0. The variance then equals the second moment. Using the definition of the gamma integral [20, Eq. 3.462.9], it can be shown that

$$\sigma_{S_R}^2 = \int_{-\infty}^{\infty} s_R^2 f_{S_R}(s_R) ds_R = \frac{\nu(\nu+1)}{\beta^2}. \tag{33}$$

2) Second Moment of S_R for $\gamma = 2$: The variance of S_R is in this case given by

$$\sigma_{S_R}^2 = \frac{\nu}{\beta} \tag{34}$$

where [20, Eq. 3.462.9] has been used.

Appendix II

Investigation of the Sign of the Errors in Approximate Estimators

A. Approximated pdf

Notice from (14) that $\hat{A}^{(1)}$ is proportional to the mean value of the following pdf:

$$f(y) = \frac{y^{\nu-1}\exp\left[-\frac{y^2}{4\zeta} - \frac{\mu y}{2\sqrt{\zeta\xi}}\right] I_0(y)}{\int_0^\infty y^{\nu-1}\exp\left[-\frac{y^2}{4\zeta} - \frac{\mu y}{2\sqrt{\zeta\xi}}\right] I_0(y)dy}. \tag{35}$$

The integral in the denominator provides for the proper normalization such that $f(y)$ integrates to one. When one of the approximations of the Bessel function discussed in Section IV-B is made, it is used both in the numerator and the denominator of (35), and the resulting estimate of the amplitude $\hat{A}^{(1)}_{\ll,K}$ or $\hat{A}^{(1)}_{\gg}$ can be interpreted as the mean of a *different* pdf that approximates the pdf in (35). Now it will be investigated under what conditions the mean of the pdf corresponding to a certain approximation is less than or equal to the mean of the original pdf.

B. Condition for a Smaller Mean

We have just seen that for the approximations of (35), the resulting amplitude estimate is the mean of an approximate pdf. Roughly speaking: this approximate pdf will certainly have a smaller mean than the original pdf when it has less probability mass everywhere in the tail of the distribution. A sufficient (but not necessary) condition for the resulting approximate pdf $\hat{f}(y)$ to have a mean that is less than or equal to the mean of $f(y)$ is

$$\frac{\hat{f}(y+\Delta y)}{\hat{f}(y)} \leqslant \frac{f(y+\Delta y)}{f(y)} \tag{36}$$

for all y and $\Delta y > 0$. We look at ratios of function values in (36) instead of differences between function values, because for different approximations, the normalization in the numerator of (35) is different. The normalization is automatically taken into account by considering ratios. In the limit of $\Delta y \to 0$, the condition becomes

$$\frac{d\ln \hat{f}}{dy} - \frac{d\ln f}{dy} \leqslant 0, \quad \forall y. \tag{37}$$

The condition says that $\hat{f}$ should decay faster on a logarithmic scale. A logarithmic scale is needed to take care of the normalization. It will be shown next that this condition is met by the approximation for low SNR conditions (16) for any $K > 0$.

1) Approximation $\hat{A}^{(1)}_{\ll,K}$ for Low SNRs: It will be shown that $\hat{A}^{(1)}_{\ll,K} < \hat{A}^{(1)}$, for all finite K. The Taylor expansion $I_0(y;K)$ in (15) becomes a more accurate approximation of $I_0(y)$ the smaller y becomes. Its derivative, denoted by $I_1(y;K)$, is given by

$$I_1(y;K) = \sum_{k=1}^{K-1} (2k)c_k y^{2k-1}, \quad c_k = \frac{2^{-2k}}{(k!)^2}. \tag{38}$$

When we let the number of terms K go to infinity, we arrive at $I_0(y)$ and $I_1(y)$, respectively [21, Eq. 9.6.10]. Define $Q(y;K) = I_1(y;K)/I_0(y;K)$. Since only the Bessel function is approximated here, the condition (37) is equivalent to $Q(y;K) \leq I_1(y)/I_0(y)$. For $y = 0$, $Q(y;K) = 0$ and also $I_1(y)/I_0(y) = 0$. The condition (37) is true for $y > 0$ if $Q(y;K)$ is monotonically increasing with K, since $Q(y;K)$ then approaches $I_1(y)/I_0(y)$ from below. Compared to $Q(y;K)$, $Q(y;K+1)$ contains an extra term $2(K+1)c_{K+1}y^{2K+1}$ in the numerator and an extra term $c_{K+1}y^{2K+2}$ in the denominator. For any $y > 0$, the ratio of those terms increases with K, since all c_k are positive. This ratio is therefore larger than $Q(y;K)$, and it follows that $Q(y;K+1) > Q(y;K)$.

2) Approximation $\hat{A}^{(1)}_{\gg}$: Equation (37) is not valid for all values of y for the approximation made in (18). This means that $\hat{A}^{(1)}_{\gg}$ may both be smaller and larger than $\hat{A}^{(1)}$. That it can be smaller is clearly visible in Fig. 5. This happens at low values of *a posteriori* SNR, where $\hat{A}^{(1)}_{\ll,K}$ is generally a much better approximation. It can also be larger than the true value, but extensive calculations have shown that when this happens, the error is not more than about 0.2 dB (see Section IV-D).

Appendix III
Modified MAP Estimator

The estimator originally proposed in [7] was computed as

$$\max_a \log f_A(a) f_{R|A}(r|a) \tag{39}$$

with $f_{R|A}(r|a)$ as in (10). However, in [7] $f_A(a)$ was not used in the form (1) with $\gamma = 1$, but in a slightly different form

$$f_A(a) = \frac{a^{\nu-1}}{\Gamma(\nu)}\left(\frac{\mu}{\sigma_S}\right)^{\nu} \exp\left\{-a\frac{\mu}{\sigma_S}\right\} \tag{40}$$

where μ and ν were treated as independent parameters although μ is in fact completely specified by ν, see (43). Since an analytical solution to (39) is hard to find, the approximation (17) for the Bessel function was made *before* taking the derivative with respect to a in (39). This led to the gain function

$$G^{(1)}_{\mathrm{MAP}} = u + \sqrt{u^2 + \frac{\nu' - 0.5}{2\zeta}}, \quad u = 1/2 - \frac{\mu}{4\sqrt{\zeta\xi}} \tag{41}$$

where $\nu' = \nu - 1$, and which is only valid for $\nu' > 0.5$. A joint amplitude and phase MAP estimator was proposed as well. The gain function $G^{(1)}_{\mathrm{JMAP}}$ of the joint MAP estimator is given by

$$G^{(1)}_{\mathrm{JMAP}} = u + \sqrt{u^2 + \frac{\nu'}{2\zeta}}, \quad u = 1/2 - \frac{\mu}{4\sqrt{\zeta\xi}} \tag{42}$$

which allows for a broader range of ν'-values, namely $\nu' > 0$. The parameters ν' and μ were estimated in [7] by fitting (40) to clean-speech amplitude distributions conditioned on a small range of high values of estimated *a priori* SNR.

The first of the modifications we make to this estimator concerns the number of free parameters in (40), (41), and (42). We see that μ and σ_S do not appear independently in (40), but only as the quotient μ/σ_S, and therefore only represent one degree of freedom. The parameter ν represents the second degree of freedom. Since $E\{A^2\}$ equals σ_S^2 by definition, it follows from (31) that

$$\mu = \sqrt{\nu(\nu+1)}. \tag{43}$$

The second modification concerns the order in which the approximation of the bessel function is used and the derivative of (39) is taken. More specifically, we compute the amplitude MAP estimator by *first* taking the derivative and *then* using the large-argument approximation $I_1/I_0 \approx 1$, where I_1 is the first-order modified Bessel function of the first kind. Interestingly, the resulting MAP gain function is identical to the joint MAP gain function in (42), with μ given by (43). We will refer to this amplitude MAP estimator as $\hat{A}^{(1)}_{\mathrm{MAP}}$.

References

[1] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, "Minimum mean-square error amplitude estimators for speech enhancement under the generalized gamma distribution," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2006.

[2] J. Jensen, R. C. Hendriks, J. S. Erkelens, and R. Heusdens, "MMSE estimation of complex-valued discrete fourier coefficients with generalized gamma priors," in *Proc. Int. Conf. Spoken Lang. Process.—Interspeech 2006*, Sep. 2006, pp. 257–260.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

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

[6] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[7] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 7, pp. 1110–1126, 2005.

[8] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[9] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[10] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularization for time-frequency surface estimation," *J. R. Statist. Soc. B*, vol. 66, pp. 575–589, 2004.

[11] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP J. Appl. Signal Process.*, no. 10, pp. 1043–1051, 2003.

[12] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. III, pp. 1068–1071.

[13] T. H. Dat, K. Takeda, and F. Itakura, "Generalized Gamma modeling of speech and its online estimation for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. IV, pp. 181–184.

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

[15] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

[16] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, "A study of the distribution of time-domain speech samples and discrete Fourier coefficients," in *Proc. IEEE 1st BENELUX/DSP Valley Signal Process. Symp.*, Apr. 2005, pp. 155–158.

[17] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1984, pp. 18A.2.1–18A.2.4.

[18] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[19] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

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

[21] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 1964, 9th dover printing, 10th gpo printing ed.

[22] W. Rudin, *Real and complex analysis*, 3rd ed. New York: McGraw-Hill, 1987.

[23] S. Zhang and J. Jin, *Computation of Special Functions*. New York: Wiley, 1996.

[24] J. Jensen, R. C. Hendriks, and J. S. Erkelens, "MMSE estimation of discrete fourier coefficients with a generalized gamma prior," Delft Univ. Technol., Delft, The Netherlands, 2006, Tech. Rep.

[25] Noizeus: A noisy speech corpus for evaluation of speech enhancement algorithms. [Online]. Available: http://www.utdallas.edu/ loizou/speech/noizeus/

[26] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

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

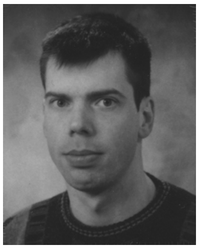
[28] J. G. Beerends, "PESQ for assessing speech intelligibility," Oct. 2004, Extending p.862, White Contribution COM 12-C2 to ITU-T Study Group 12.

[29] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[30] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. New York: Springer, 2005.

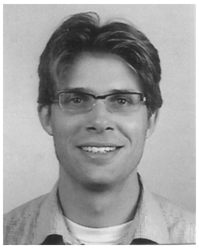
[31] J. S. Erkelens, J. Jensen, and R. Heusdens, "A general optimization procedure for spectral speech enhancement methods," in *Proc. Eur. Signal Process. Conf. EUSIPCO*, Florence, Italy, 2006.

[32] J. S. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun., Special Issue on Speech Enhancement*, to appear in 2007.



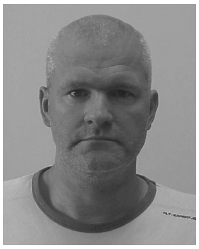
Jan S. Erkelens received the Ph.D. degree from the Applied Physics Department, Delft University of Technology, Delft, The Netherlands, in 1996. The subject of his thesis was low bit-rate speech coding. He also has experience in the atmospheric sciences and speech recognition.

He currently holds a Postdoctoral position in the Information and Communication Theory Group, Department of Mediamatics, Delft University of Technology, where he is working on speech enhancement.



Richard C. Hendriks received the B.Sc. and M.Sc. degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2001 and 2003, respectively. He is currently pursuing the Ph.D. degree at Delft University of Technology.

In September 2003, he joined the Department of Mediamatics, Delft University of Technology. From September 2005 to December 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. His main research interests are digital speech and audio processing, including acoustical noise reduction and speech enhancement.



Richard Heusdens received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, Delft University of Technology. In the spring of 1992, he joined the Digital Signal Processing Group, Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT group. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, and digital watermarking of audio.

Jesper Jensen received the M.Sc and Ph.D degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively.

From 1996 to 2001, he was with the Center for PersonKommunikation (CPK), Aalborg University, as a Researcher, Ph.D. student, and Assistant Research Professor. In 1999, he was a Visiting Researcher at the Center for Spoken Language Research, University of Colorado, Boulder. From 2000 to 2007, he was a Postdoctoral Researcher and Assistant Professor at Delft University of Technology, Delft, The Netherlands. He is currently with Oticon, Copenhagen, Denmark. His main research interests are digital speech and audio signal processing, including coding, synthesis, and enhancement.