Spectral Magnitude Minimum Mean-Square Error Estimation Using Binary and Continuous Gain Functions

Jesper Jensen and Richard C. Hendriks

***Abstract*—Recently, binary mask techniques have been proposed as a tool for retrieving a target speech signal from a noisy observation. A binary gain function is applied to time–frequency tiles of the noisy observation in order to suppress noise dominated and retain target dominated time–frequency regions. When implemented using discrete Fourier transform (DFT) techniques, the binary mask techniques can be seen as a special case of the broader class of DFT-based speech enhancement algorithms, for which the applied gain function is not constrained to be binary. In this context, we develop and compare binary mask techniques to state-of-the-art continuous gain techniques. We derive spectral magnitude minimum mean-square error binary gain estimators; the binary gain estimators turn out to be simple functions of the continuous gain estimators. We show that the optimal binary estimators are closely related to a range of existing, heuristically developed, binary gain estimators. The derived binary gain estimators perform better than existing binary gain estimators in simulation experiments with speech signals contaminated by several different noise sources as measured by speech quality and intelligibility measures. However, even the best binary mask method is significantly outperformed by state-of-the-art continuous gain estimators. The instrumental intelligibility results are confirmed in an intelligibility listening test.**

***Index Terms*—Ideal Binary Mask, Spectral Magnitude Estimation, Speech Enhancement, Speech Intelligibility, Speech Quality.**

I. Introduction

THE human auditory system is remarkable at processing speech in noise: humans are able to understand speech in severe noisy conditions, where the target speech signal has been contaminated by highly nonstationary noises and reverberance; this ability is perhaps best examplified by the so-called cocktail party problem [1]. The robustness of the human auditory system in this respect is unparalleled by any machine. The human auditory system is believed to process the acoustic input in two distinct stages, e.g., [1], [2]. An analysis stage, believed to mainly take place in the auditory periphery, where the input signal is decomposed in time–frequency units, and a grouping stage where time–frequency units dominated by the same source are grouped, using a combination of signal cues such as harmonicity, common onsets, etc., and learned characteristics of the target signal [1].

Recently, ideal binary mask (idbm) techniques have been proposed as a signal processing tool for simulating and studying the time–frequency analysis and grouping process of the auditory system, e.g., [3]–[8]. In the simplest setting, the idbm techniques assume that a target speech signal $s(n)$ has been contaminated by an additive noise source $w(n)$ such that the noisy mixture signal $x(n)$ is given by $x(n) = s(n) + w(n)$, where n denotes a discrete-time index. The idbm techniques decompose the signals $x(n)$, $s(n)$, and $w(n)$, e.g., by applying a discrete Fourier transform (DFT) in successive time frames, e.g., [5] and [6], or using gamma-tone filter banks, e.g., [7], [8], resulting in time–frequency units, $x(k,m)$, $s(k,m)$ and $w(k,m)$, respectively, where k and m are frequency and time indices. Generally speaking, the idbm techniques retain time–frequency units which are dominated by the target speech, and suppress time–frequency units which are dominated by the noise source. More specifically, an ideal binary mask value $g(k,m) \in \{g_{\min}, g_{\max}\}$ is computed for each time–frequency unit by comparing the local target-to-noise ratio $|s(k,m)|^2/|w(k,m)|^2$ to a threshold value $\rho(k,m)$

$$g(k,m) = \begin{cases} g_{\max}, & \text{if } \frac{|s(k,m)|^2}{|w(k,m)|^2} > \rho(k,m) \\ g_{\min}, & \text{otherwise} \end{cases} \tag{1}$$

where $g_{\max} > g_{\min} \geq 0$. Finally, the ideal binary mask values $g(k,m)$ are multiplied with the original noisy time–frequency units $x(k,m)$ and an idbm segregated signal can be synthesized by frequency-to-time transform, e.g., using short-time inverse DFTs and overlap-add techniques or a gamma-tone synthesis filter bank. The term "ideal" is appropriate here, since it is assumed that the local target-to-noise energy ratio in a given time–frequency unit is known with certainty. The term "binary" is used since the gain function $g(k,m)$ is one of two values; in the following we also refer to $g(k,m)$ as a *binary gain* (BG) function. Typically $g_{\max} = 1$, and often $g_{\min} = 0$ (see, e.g., [4]–[7]), although other values are used as well, see, e.g., [9]. The threshold $\rho(k,m)$ is often chosen as 0 dB (i.e., $\rho(k,m) = 1$) for all time and frequency indices, e.g., [6], although other choices are possible. Brungart [8] measured the impact of the threshold value on speech intelligibility and found that optimal thresholds were dependent on the global signal-to-noise ratio (SNR) measured across one or several speech sentences. Specifically, an SNR increase of Δ dB led to a threshold increase of

Manuscript received October 10, 2010; revised January 24, 2011; accepted May 09, 2011. Date of publication May 27, 2011; date of current version November 04, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Michael Seltzer.

J. Jensen is with Oticon A/S, 2765 Copenhagen, Denmark (e-mail: jsj@oticon.dk).

R. C. Hendriks is with the Multimedia Signal Processing Group, Delft University of Technology, Delft 2628 CD, The Netherlands (e-mail: r.c.hendriks@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2157685

approximately Δ dB; with this approach, a given high-intelligibility mask pattern would be roughly maintained for changing global SNRs. Later, this threshold-SNR relationship was confirmed in studies using other stimuli and intelligibility tasks [5], [7]. In fact, the relationship holds for noisy speech signals at global SNRs as low as -60 dB, which can be rendered essentially perfectly intelligible by applying an ideal binary mask [7].

The idbm framework was originally proposed as a signal processing tool for studying aspects of the auditory system. However, perhaps motivated by the significant intelligibility improvements achievable in this ideal setting, where local target-to-noise energy ratios are known with certainty, the idbm framework has more recently been adapted to the practical problem of retrieving a target speech signal from a noisy mixture in the *non-ideal* situation where the local target-to-noise energy ratios are unknown, but must be estimated from the noisy mixture signal $x(n)$, e.g., [4], [5], [10], [11]. For example, the binary gain method in [11] has demonstrated improvements of the intelligibility of noisy speech. This algorithm classifies time–frequency units as being noise or speech dominated using Gaussian mixture models of feature vectors derived from not only the noisy time–frequency unit in question, but also the units which are close in time and frequency. As with data driven methods in general, though, this method is somewhat dependent on *a priori* knowledge of the acoustic environment in which it operates (e.g., global SNR, noise type, transducer characteristics, etc.).

On the other hand, the general problem of extracting a target speech signal from a noisy mixture has been studied extensively in the speech enhancement community during the last three decades, see, e.g., [12] and [13] and the references therein. The class of DFT based enhancement methods, e.g., [14]–[17], is very similar to the idbm techniques described above, except that a real-valued, *continuous* rather than binary gain function is applied to short-time DFT coefficients, before the noise reduced signal is reconstructed through IDFT and overlap-add techniques. Often the continuous gain function is derived analytically by assuming a statistical model for the target and noise DFT coefficients, and minimizing a suitable distortion criterion, e.g., the spectral magnitude minimum mean-square error (MMSE). The main difference to the idbm techniques described above is that the gain values applied to the noise DFT coefficients are not constrained to be binary, but can in principle take on any real non-negative value; we refer to these as *continuous gain (CG)* functions. While single-channel CG DFT techniques are capable of increasing speech quality [18], increases in speech intelligibility remain modest at best [19], [20]. The CG techniques have primarily been applied to the practical situation of extracting a target speech signal when only a noisy realization is available. If CG techniques were applied to the ideal situation where the true target-to-noise ratio is available, as in (1), then performance similar to or better than that of BG techniques might be expected, as the CG techniques offer more degrees of freedom for solving the same problem [6].

In this paper, we compare single-channel DFT-based BG and CG techniques in the practical (non-ideal) situation where local target-to-noise ratios are unknown *a priori*. In light of the fact that the BG functions used in the idbm framework and the CG techniques have strong similarities and have co-existed for at least a decade, it is surprising to note that hardly any such comparison exists. The main goal of the paper is twofold. First, noting that existing BG estimators are heuristically motivated, we wish to derive analytical expressions for binary spectral MMSE estimators. Second, we wish to quantify the performance of these estimators in terms of target speech quality and intelligibility with reference to existing BG techniques and state-of-the-art CG techniques.

The remainder of the paper is organized as follows. In Section II, we present the basic statistical signal model used throughout, and introduce notation. Section III treats the problem of spectral magnitude MMSE estimation in the case where both target and noise spectral magnitudes are assumed sparsely distributed, in order to simulate a situation where BG functions have been hypothesized to be close to optimal. In Section IV, we derive binary gain functions which are optimal in an MMSE sense, and discuss the relationship between these and existing estimators. Section V presents an evaluation of the derived estimators in terms of speech quality and intelligibility. Finally, in Section VI we conclude the work.

II. Signal Model and Notation

We represent random variables by capital letters and realizations thereof as lowercase letters. We consider an additive signal model of the form

$$X(k,m) = S(k,m) + W(k,m) \tag{2}$$

where $X(k,m)$, $S(k,m)$, and $W(k,m)$ are zero-mean random variables representing DFT coefficients at frequency index k and frame index m for the noisy observation, the speech target and an additive noise term, respectively. We use the standard assumptions that $S(k,m)$ and $W(k,m)$ are statistically independent and that signal frames are sufficiently long and that their overlap is sufficiently small, such that DFT coefficients are approximately independent across time and frequency [21], [22]; for this reason, without loss of generality, we may drop time and frequency indices, and simply write

$$X = S + W. \tag{3}$$

Let $R = |X|, A = |S|$, and $N = |W|$ denote random variables representing the noisy, clean and noise spectral magnitude, respectively. Furthermore, we introduce the spectral variances, $\sigma_S^2 = E(A^2)$ and $\sigma_W^2 = E(N^2)$, and let $\xi = \sigma_S^2/\sigma_W^2$ and $\zeta = r^2/\sigma_W^2$ denote the *a priori* and *a posteriori* SNR [14], respectively.

To model the observation that speech spectral magnitudes are sparsely distributed, see, e.g., [23], we assume that speech DFT magnitudes $A \geq 0$ are distributed according to a probability density function (pdf) of the form

$$f_A(a;\gamma,\nu) = \frac{\gamma\beta^\nu}{\Gamma(\nu)} a^{\gamma\nu-1} \exp\left(-\beta a^\gamma\right), \ \gamma > 0, \nu > 0 \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function. For given parameters γ, ν, and spectral magnitude variance $\sigma_A^2 = E(A^2) - E(A)^2$, the parameter $\beta > 0$ is fully determined. For example, for $\gamma = 1$, $\beta = \sqrt{\nu(\nu+1)/\sigma_A^2}$ and for $\gamma = 2$, $\beta = \nu/\sigma_A^2$ [17]. In this

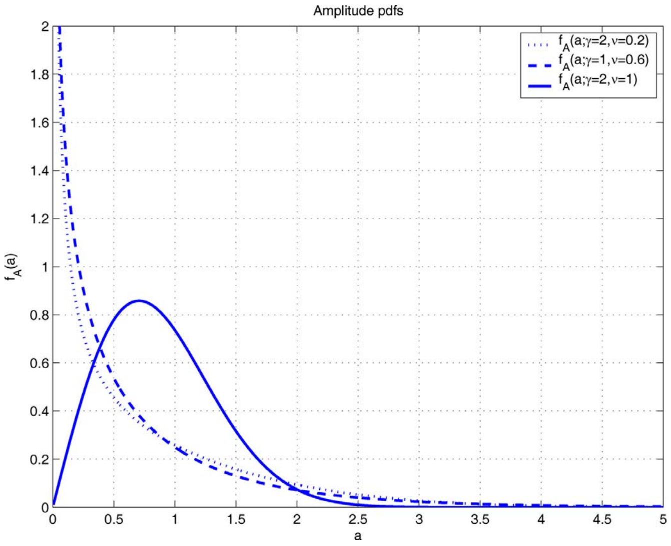

Fig. 1. Models of spectral magnitude pdfs ($\sigma_A^2 = 1$). The pdfs $f_A(a;\gamma = 2, \nu = 0.2)$ and $f_A(a;\gamma = 1, \nu = 0.6)$ are realistic models of DFT speech magnitude distributions. The Rayleigh pdf $f_A(a;\gamma = 2, \nu = 1)$ arises from a Gaussian speech DFT assumption.

paper, we consider as examples $f_A(a;\gamma = 1, \nu = 0.6)$ and $f_A(a;\gamma = 2, \nu = 0.2)$, see Fig. 1, which both provide good models of speech DFT magnitude distributions in the present context [17]. The corresponding phase variable is assumed to be uniformly distributed in $[0; 2\pi[$, and independent of A.

III. MMSE Gain Functions for Sparse Target and Noise

We start out by addressing one of the motivations often presented in support of the idbm framework. Specifically, it has been hypothesized that a binary gain rule would be close to optimal when the target and noise source are both sparse, see, e.g., [4] and the references therein. This situation arises e.g., for a single target speaker contaminated by a single competing speaker, because in this case any sufficiently small time–frequency unit of the mixture signal would be dominated by either the target or the noise source, but typically not both simultaneously; the optimal gain value is argued to be close to unity when the target dominates and close to zero when the noise source dominates. While this argument may be valid in the ideal case where local SNR *realizations* are available, i.e., the situation where (1) would be applicable (see [6] for a discussion), we show in this section that in the practical case where only statistical models of the target and noise source are available, spectral magnitude MMSE gain functions tend not to have binary characteristics.

To illustrate this, we compute spectral magnitude MMSE gain functions for sparse target and noise signals. More specifically, we assume that the spectral magnitudes of the target are distributed according to $f_A(a;\gamma = 2, \nu = 0.2)$, and, for the purpose of illustration, that the noise spectral magnitudes N follow an identical distribution $f_N(n) = f_A(n;\gamma = 2, \nu = 0.2)$, simulating a competing speaker. With these specific distributional assumptions it turns out that analytic expressions for MMSE gain functions do not exist, and therefore we estimate the gain functions numerically. This is done by constructing realizations of DFT coefficients by drawing spectral magnitudes according to $f_A(f)$ and $f_N(n)$, and appending uniformly and independently distributed phase realizations, to form realizations of S, W, $X = S + W$, A, and R. For a detailed description of the procedure, we refer to [24]; see also [25] for further analysis and applications of this procedure. Fig. 2(a) shows optimal gain functions for *a priori* SNR values of $\xi = -10, 0$ and 10 dB, respectively. The gain functions are plotted here as a function of *a posteriori* SNR $\zeta = r^2/\sigma_W^2$, and not the instantaneous SNR used in (1) in anticipation of the derivation of gain functions in Section IV for the practical case where target and noise realizations are not available; the instantaneous SNR is difficult to estimate reliably from noisy observations whereas the *a posteriori* SNR is easier. For comparison, Fig. 2(c) plots the gain functions for the analytically tractable case where target and noise magnitudes are Rayleigh distributed (corresponding to Gaussian DFT coefficients). Clearly, the gain functions in Fig. 2(a) do not have obvious binary characteristics; in fact, they are relatively constant with ξ, and taking into account the distribution of the *a posteriori* SNR ζ, Fig. 2(b), the applied gain will not differ much for varying values of ζ either. We conclude that binary gain functions are certainly suboptimal in the case of a sparse signal and noise. However, apart from the single competing speaker situation, many noise sources encountered in practice have a time-span of dependency [21] which is relatively short, see, e.g., [23], and it is often more reasonable to assume a Gaussian model for the noise DFT coefficients. This observation, combined with the fact that binary gain functions may have complexity and storage advantages, makes it of interest to determine the optimal performance of binary gain functions for Gaussian distributed noise DFT coefficients.

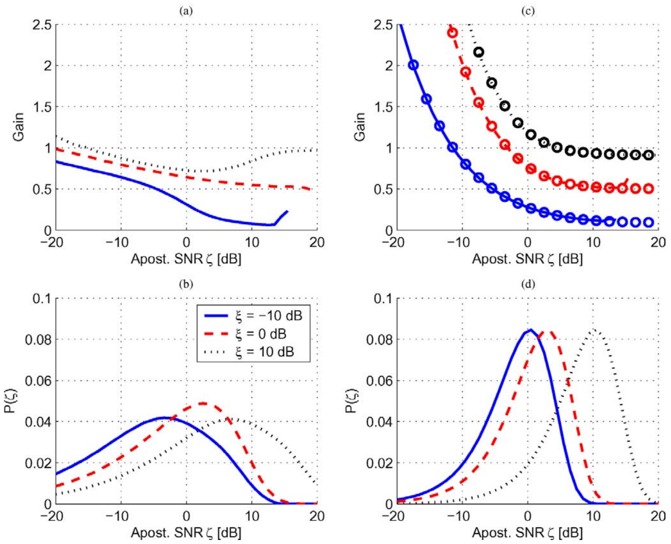

Fig. 2. (a) Spectral magnitude MMSE gain functions for sparse target and noise magnitudes for three *a priori* SNRs, $\xi = -10, 0,$ and 10 dB. Gain functions are only estimated for values of ζ for which more than five realizations of random variables (A, R) are available. (b) Distribution of *a posteriori* SNRs ζ. (c) MMSE gain functions for Gaussian target and noise. Circles indicate analytically derived gain functions (from (7) in [15]) for verification. (d) Distribution of *a posteriori* SNRs ζ.

IV. MMSE Binary Gain Functions

In this section, we derive spectral magnitude MMSE *binary* gain functions. Our approach is general and allows determining the MMSE binary gain function for any statistical signal model for which a continuous MMSE gain function exists. We consider two types of binary gain functions, Type 1 which is constrained to be non-decreasing in the *a posteriori* SNR ζ, and Type 2, where this constraint has been relaxed.

A. MMSE Binary Gain Function (Type 1)

Consider a binary gain function $g(k,m) \in \{\epsilon, 1\}, 0 \leq \epsilon < 1$, of the form

$$g(k,m) = \begin{cases} 1, & \text{for } \zeta(k,m) > \rho(k,m) \\ \epsilon, & \text{otherwise} \end{cases} \tag{5}$$

where, without loss of generality, we have chosen $g_{\max} = 1$, and $g_{\min} = \epsilon$. This type of BG function is typical for the idbm framework: for low (estimates of) SNR, the applied gain is low, $g = \epsilon < 1$, while for SNRs beyond a threshold, $\zeta > \rho$, the binary gain function is high, $g = 1$. Note the relationship $E(\zeta) = E(A^2)/E(N^2) + 1$ between the *a posteriori* SNR and the *expected* (i.e., *a priori*) SNR.

Using (5), the spectral magnitude mean-square error $J_1 = E(A - \hat{A})^2$ is given by

$$\begin{aligned} J_1 &= \int_R \int_A (a - g(r)r)^2 f_{A,R}(a,r) da dr \\ &= \int_0^{\tilde{\rho}} \int_A (a - \epsilon r)^2 f_{A|r}(a) f_R(r) da dr \\ &\quad + \int_{\tilde{\rho}}^{\infty} \int_A (a - r)^2 f_{A|r}(a) f_R(r) da dr \end{aligned} \tag{6}$$

where $\tilde{\rho}$ is the noisy magnitude corresponding to the threshold ρ, that is $\tilde{\rho} = \sqrt{\rho \sigma_W^2}$. We differentiate J_1 with respect to the threshold $\tilde{\rho}$ using Leibniz' rule [26, [0.410]],

$$\frac{\partial J_1}{\partial \tilde{\rho}} = 2\tilde{\rho} E(A|r = \tilde{\rho})(1 - \epsilon) - \tilde{\rho}^2(1 - \epsilon^2) \tag{7}$$

and solve $\partial J_1 / \partial \tilde{\rho} = 0$, to find that the optimal threshold value satisfies

$$\frac{E(A|r = \tilde{\rho})}{\tilde{\rho}} = \frac{1}{2}(1 + \epsilon). \tag{8}$$

Let $g_{\text{MMSE}}(r)$ denote the MMSE continuous gain function evaluated for the noisy magnitude r, and recall that the conditional mean $E(A|r)$ is identical to the MMSE estimator (e.g., [27]). It then follows that the left side of (8) is $g_{\text{MMSE}}(r = \tilde{\rho})$. We conclude that the optimal value of $\tilde{\rho}$ is simply the value of the noisy magnitude r for which the MMSE continuous gain function is equal to $(1/2)(1 + \epsilon)$.

More formally, let R_1 denote the set $R_1 = \{r : g_{\text{MMSE}}(r) = (1/2)(1 + \epsilon)\} \cup \{0\}$, and define the optimal threshold $\tilde{\rho}^* = \min_{r \in R_1} J_1$. Then, the MMSE BG function of Type 1 (BG1-MMSE) is given by

$$g(r) = \begin{cases} 1, & \text{for } r \geq \tilde{\rho}^* \\ \epsilon, & \text{otherwise.} \end{cases} \tag{9}$$

Fig. 3 shows examples of MMSE CG functions and the corresponding BG functions of Type 1 (BG1-MMSE) for different choices of *a priori* SNR ξ, for the case where noise DFT coefficients are (complex) Gaussian distributed, such that the noise magnitudes N are Rayleigh distributed; noise phases are independently and uniformly distributed [23].

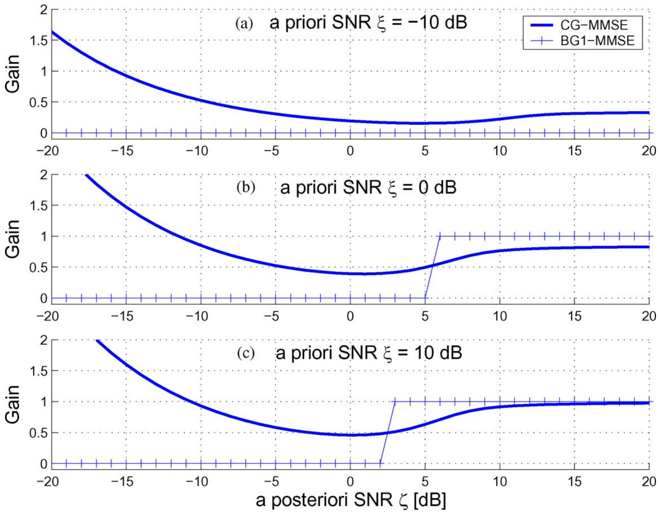

Fig. 3. Continuous and binary gain functions (Type 1) ($\epsilon = 0$) for target pdf $f_A(a; \gamma = 2, \nu = 0.2)$, for *a priori* SNRs (a) $\xi = -10$ dB, (b) 0 dB, and (c) 10 dB. The noise DFT magnitudes are assumed to follow a Rayleigh distribution (which implies that the noise DFT coefficients obey a complex Gaussian distribution).

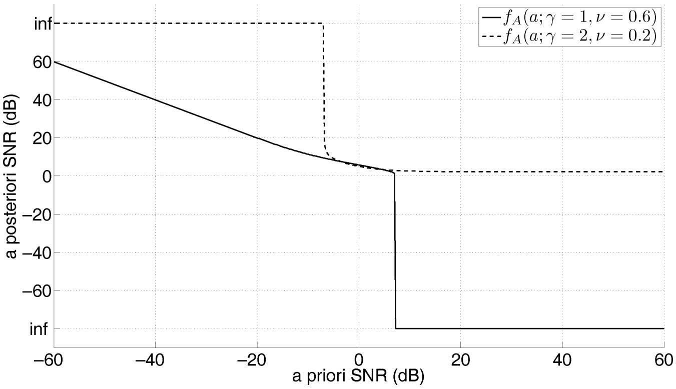

Fig. 4. Optimal threshold in terms of *a posteriori* SNR r^2/σ_W^2 as a function of *a priori* SNR ξ for the BG1-MMSE estimator ($\epsilon = 0$) for different statistical models $f_A(a)$. The noise magnitudes are assumed Rayleigh distributed.

The fact that the optimal threshold $\tilde{\rho}^*$ follows analytically is in contrast to the idbm schemes, which tend to choose the threshold more heuristically, e.g., [10]. To illustrate the behavior of the optimal threshold value, Fig. 4 plots ρ in terms of *a posteriori* SNR, as a function of *a priori* SNR. Not only is the optimal threshold value dependent on *a priori* SNR (and thus global SNR), the derived threshold *decreases* for increasing *a priori* SNR, leading to a less sparse mask pattern at higher global SNRs. Interestingly, this is in contrast to the idbm studies [5], [7], [8] which, in the ideal scenario where local SNR realizations are available, *increase* the threshold for increasing SNR, in order to maintain a given high-intelligibility mask pattern at all SNRs. Fig. 6(a), (c), and (e) summarize the BG1-MMSE estimators for various target distributions $f_A(a)$.

B. MMSE Binary Gain Function (Type 2)

The binary gain function in (5) is a non-decreasing function of the *a posteriori* SNR ζ, see Fig. 3, to be in line with the traditional ideal binary mask definition in (1); it is constant or shaped

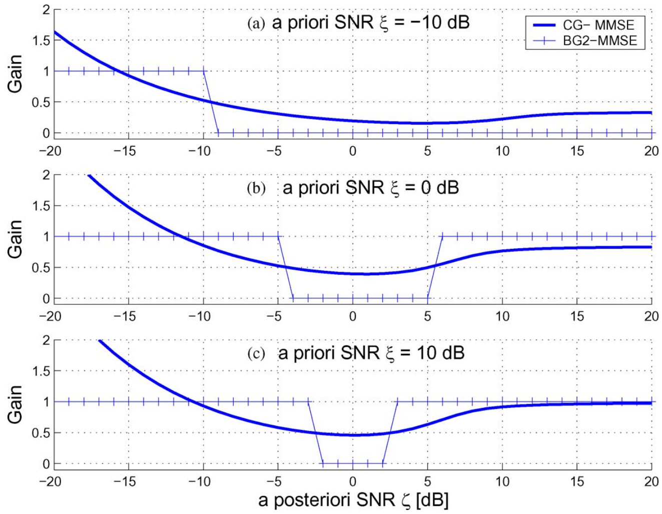

Fig. 5. Continuous and binary gain functions (Type 2) for target pdf $f_A(a;\gamma = 2, \nu = 0.2)$, for *a priori* SNRs (a) $\xi = -10$ dB, (b) 0 dB, and (c) 10 dB.

like a positive step function. In this section, we derive an MMSE binary gain function, $g(k,m) = \{\epsilon, 1\}$, which is not constrained to be non-decreasing. We show in the simulation examples in Section V, that this leads to a performance advantage.

Consider therefore minimization of the spectral magnitude mean-square error

$$J_2 = E(A - \hat{A})^2 = \int_R \int_A (a - g(r)r)^2 f_{A|r}(a|r) da f_R(r) dr$$
$$= \int_R \underbrace{\left(E(A^2|r) + g^2(r)r^2 - 2E(A|r)g(r)r\right)}_{T(r)} f_R(r) dr \quad (10)$$

where $T(r)$ is a function of the noisy magnitude realization r, but not of the clean magnitude realizations a. As the gain value is constrained to be binary, $g(r) \in \{\epsilon, 1\}, 0 \le \epsilon < 1$, there exist two possible values of $T(r)$ for a given r:

$$T_1(r) \triangleq E(A^2|r) + r^2 - 2E(A|r)r \quad (11)$$

for $g(r) = 1$, and

$$T_\epsilon(r) \triangleq E(A^2|r) + \epsilon^2 r^2 - 2\epsilon E(A|r)r \quad (12)$$

for $g(r) = \epsilon$.

In order to minimize J_2 in (10), one must choose the smaller of the two, that is use

$$g(r) = \begin{cases} 1, & \text{for } T_1(r) < T_\epsilon(r) \\ \epsilon, & \text{otherwise.} \end{cases} \quad (13)$$

Combining (11), (12), and (13), it follows that $T_1(r) < T_\epsilon(r)$ when $g_{\mathrm{MMSE}}(r) > (1/2)(1+\epsilon)$. We conclude that the MMSE binary gain function of type II (BG2-MMSE) can be described as follows. Let R_2 denote the noisy magnitude set $R_2 = \{r : g_{\mathrm{MMSE}}(r) > (1/2)(1+\epsilon)\}$. Then

$$g(r) = \begin{cases} 1, & \text{for } r \in R_2 \\ \epsilon, & \text{otherwise.} \end{cases} \quad (14)$$

Note that the BG2-MMSE gain function, (14), is simply a quantized version of $g_{\mathrm{MMSE}}(r)$. Fig. 5 shows examples of MMSE continuous gain functions and the corresponding BG2-MMSE functions. Fig. 6(b), (d), and (e) summarize the BG2-MMSE estimators for various target distributions.

C. Relations Between Derived Gain Functions

From Figs. 3 and 5 it is clear that the difference between the Type 1 and 2 MMSE BG functions occurs at low *a posteriori* SNRs ζ. For these low SNR values, the CG function attains values much higher than $(1/2)(1+\epsilon)$. Applying less suppression for low *a posteriori* SNR may seem surprising; however, recalling the definition of the *a posteriori* SNR $\zeta = r^2/\sigma_W^2$, it is clear that ζ can be low when relatively large target DFT coefficients are summed out-of-phase with noise DFT coefficients, producing a small noisy DFT magnitude r. Therefore, in order to restore the clean magnitude, the necessary gain can exceed one. As observed by Martin [28], these high gain values produce a magnitude estimate which is roughly constant, independent of the observed noisy magnitude r; this property is desirable as it helps reduce unnatural fluctuations in the estimated signal [28]. The Type 2 BG function approximates this behavior albeit coarsely, whereas the Type 1 BG function cannot, since it is constrained to be monotonic. Despite this difference between Type 1 and 2 estimators, informal listening tests suggests that it is rather small in practice, compared to the other and larger distortions that arise when using binary rather than continuous gain functions.

Theoretically, the binary gain functions are inferior to the continuous gain functions: the MMSE CG function $g_{\mathrm{MMSE}}(R)$ is determined by unconstrained minimization of the spectral magnitude mean-square error $J = E(A - g(R)R)^2$. For the BG2-MMSE function, a binary gain constraint $g(\cdot) = \{\epsilon, 1\}$ is imposed, leading to an MMSE of J_2. For the BG1-MMSE estimator, an *additional* constraint is imposed that $g(R)$ is non-decreasing in R. Due to this series of tighter constraints imposed on the solution space, it follows that

$$J \le J_2 \le J_1. \quad (15)$$

However, in practice, the case might be less clear because underlying distributional assumptions are not perfectly valid, and the spectral magnitude MMSE may not reflect well more complex signal aspects such as perceived quality or speech intelligibility.

D. Relations to Other Binary Gain Functions

Fig. 6(a)–(d) summarize the BG functions of Type 1 and 2 for the target magnitude model $f_A(a;\gamma = 1, \nu = 0.6)$, and $f_A(a;\gamma = 2, \nu = 0.2)$. We have added Fig. 6(e) which shows the MMSE binary gain function for the alternative distortion measure $D = E(|S - \hat{S}|^2)$; $\hat{S} = g(X)X$, for the case where both S and W are (complex) Gaussian distributed. For this case, the Type 1 and 2 binary gain functions are identical. It is interesting to compare the MMSE BG functions in Fig. 6 to BG functions proposed in the literature. Specifically, it turns out that some of heuristically motivated BG functions summarized in [10], are in fact MMSE optimal. In Table I, we have repeated (and slightly rewritten) some of the gain functions discussed in [10]. Method $C1$ is identical to the gain function in Fig. 6(e), and is therefore the binary MMSE gain function for the distortion measure $D = E(|S - \hat{S}|^2)$ under a Gaussian

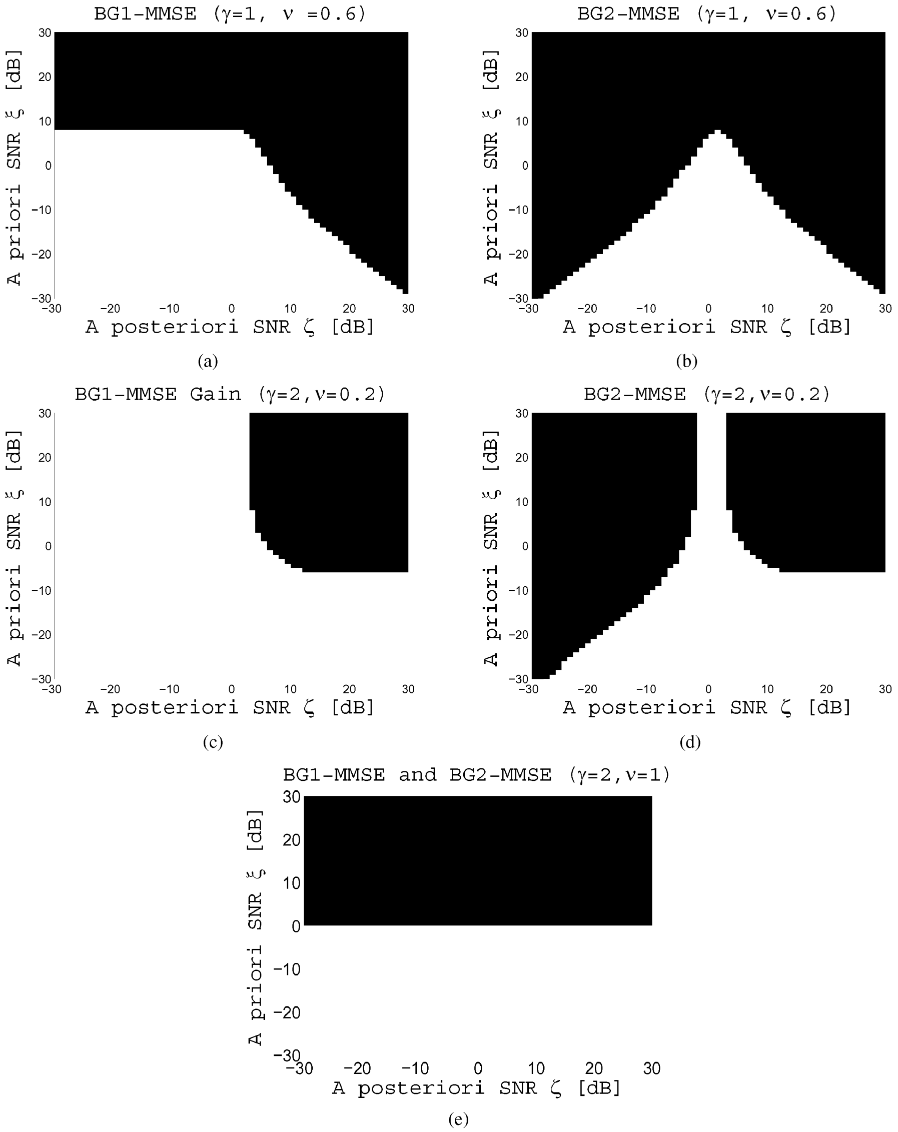

Fig. 6. Complete gain functions BG1-MMSE [(a) and (c)], and BG2-MMSE [(b) and (d)] for target source $f_A(a;\gamma=1,\nu=0.6)$ [(a) and (b)], and target source $f_A(a;\gamma=1,\nu=0.6)$ [(c) and (d)]. (e) MMSE binary gain function for distortion measure $E(|S-\hat{S}|^2)$ under Gaussian target and noise assumptions (the Type 1 and Type 2 BG estimators are identical in this case). $g=1$: black; $g=0$: white.

TABLE I
CRITERIA FOR CLASSIFYING TARGET-DOMINATED TIME–FREQUENCY UNITS (ADOPTED FROM TABLE I IN [10])

$C1$	$C2$	$C4$
$\xi > 1$	$\xi > 1,\ \zeta > 2$	$g_{ss}(r) > 1/\sqrt{2}$

target and noise assumption. Method $C2$ shows close resemblance to the BG1-MMSE function in Fig. 6(c) for target distribution $f_A(a;\gamma=2,\nu=0.2)$; both methods use $\zeta > 2$ (≈ -3 dB). The difference, apart from the soft corner, is that the BG1-MMSE function uses $\xi > 1/4$ (≈ -6 dB), rather than $\xi > 1$ used in $C2$. Method $C4$, which we have rewritten from [10], is similar to the derived Type 2 gain function, (14), provided that a minimum gain value of $\epsilon = 2g-1|_{g=1/\sqrt{2}} = 0.41$ is used, because in this case $R_2 = \{r : g_{\mathrm{MMSE}}(r) > 1/\sqrt{2}\}$; however, $C4$ deviates from the optimal estimator in that it is based on a spectral subtraction gain function $g_{ss}(r)$ and not the MMSE gain function $g_{\mathrm{MMSE}}(r)$, and uses a value of $\epsilon = 0$. Nevertheless, simulation experiments in [10] showed that $C4$ was the best among the methods considered there.

V. SIMULATION RESULTS

In this section, we present simulation results with the derived binary estimators, and compare them to existing methods. The speech material used in the simulation experiments is from the Noizeus data base [29] and consists of signals filtered at telephone bandwidth and sampled at 8 kHz. The additive noise signals are car noise and street noise from the Noizeus data base, and computer-generated telephone-bandwidth filtered white Gaussian noise. Speech and noise signals are added to form noisy speech at prescribed SNRs, with silence regions excluded when computing the SNR. The signals are processed in frames of 256 samples, with an overlap of 50%. The signal frames are weigthed with a square-root Hann analysis window, and a DFT is applied. Gain functions are computed and applied to the magnitude of the noisy DFT coefficient, before an IDFT

TABLE II
PROCESSING METHODS WHICH MAKE USE OF TRUE *A PRIORI* SNR

Abbreviation	Description
CG-MMSE-VAR	Continuous gain MMSE estimator. Computes conditional $E(A\|r)$ assuming target magnitude distribution $f_A(a;\gamma = 1$ $0.6)$, see Eq. (4). Target and noise variances used for computing estimated by smoothing of $\|a(k,m)\|^2$ and $\|n(k,m)\|^2$ across (see text for exact procedures).
BG-VAR	Ideal Binary Mask algorithm where gain values $\{0,1\}$ chosen on true ξ estimated as above, and using a threshold of $\rho = 1$ (Criterion $C1$).
BG1-MMSE-VAR	The Type 1 binary gain estimator derived in Sec. IV-A. Target noise variances for ξ estimated as above.
BG2-MMSE-VAR	The Type 2 binary gain estimator derived in Sec. IV-B. Target noise variances for ξ estimated as above.
Noisy	Unprocessed noisy speech, used for reference.

is performed. The resulting enhanced time domain frames are weighted with a square-root Hann synthesis window and overlap-added to form the enhanced signal.

A. Performance Measures

In order to quantify performance of various processing methods, we use a range of computational performance measures. First, as we have been considering MMSE magnitude estimators, we apply the squared error distortion measure

$$D_{\mathrm{magn}} = \sum_{(k,m)\in\mathcal{Q}} \left(a(k,m) - \hat{a}(k,m)\right)^2 \tag{16}$$

where $a(k,m)$ is the magnitude of a target DFT coefficient, which is available as the signals are mixed artificially, $\hat{a}(k,m)$ is the magnitude estimated by the method in question, and $\mathcal{Q}$ represents the set of all time and frequency indices. Although D_{magn} does not necessarily correlate well with speech quality or intelligibility, we choose to include it here to verify experimentally (15) for real speech data.

In addition, in order to quantify the quality of the enhanced signals, we apply the PESQ speech quality measure [30]; although originally developed for evaluation of speech codec distortions, PESQ has been used extensively for evaluating quality aspects of speech enhancement algorithms, see, e.g., [13].

Finally, to gain an indication of the intelligibility of the enhanced signals, we use the recently developed Short-Time Objective Intelligibility (STOI) measure [31]. The STOI measure outputs an average correlation coefficient $-1 \leq d_{\mathrm{STOI}} \leq 1$ which is monotonically related to the average intelligibility of the sentence in question, and has been successfully validated for noisy speech processed by binary and continuous gain functions, as well as unprocessed noisy speech, see [31] and [20], respectively.

B. True Spectral Variances Available

We first evaluate the *idealized* situation, where "true" *a priori* SNRs are available; this is possible since the noisy signals are artificially mixed from clean speech signals and noise signals. Recall that the *a priori* SNR is the ratio of target and noise spectral variances, respectively. The target spectral variance σ_S^2 is estimated using a variant of the decision directed approach [15], where, instead of using estimated spectral magnitudes from the previous frame, we simply use the noise-free magnitudes from the previous frame. The noise spectral variance σ_W^2 is estimated through recursive smoothing. Note that this approach is "less ideal" than the standard idbm setup, (1), which relies on SNR realizations.

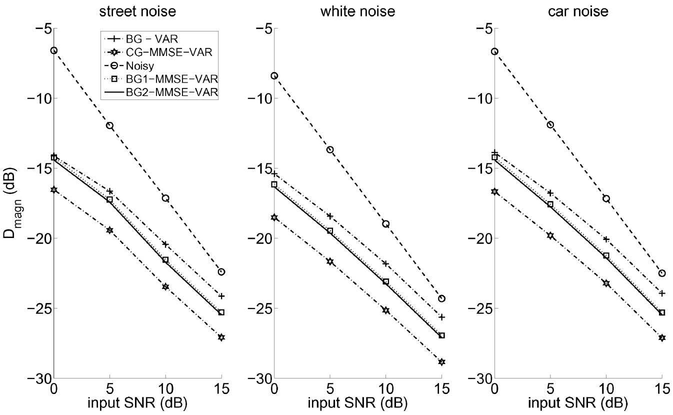

Fig. 7. Performance in terms of D_{magn} for the processing methods outlined in Table III (true *a priori* SNR known).

We consider the processing methods outlined in Table II. As the *a priori* SNRs are determined using separate target and noise signals, we expect to find upper performance bounds for the realistic case treated in Section V-C, where the *a priori* SNRs are estimated from noisy observations only.

Figs. 7–9 show the performance of the ideal processing conditions outlined in Table II. Fig. 7 shows that BG1-MMSE-VAR generally performs slightly better than BG-VAR: apparently it is advantageous to use a signal adaptive threshold, which is what BG1-MMSE-VAR does, compared to the fixed threshold ρ used in BG-VAR. However, applying a gain value of 1 rather than $\epsilon = 0$ for low *a posteriori* SNR, which is what BG2-MMSE-VAR does, improves performance even further. Finally, it is clear that the continuous gain function CG-MMSE-VAR outperforms any of the binary gain functions. The results in Fig. 7 are in line with (15). Fig. 8 shows that the binary gain functions lead to a rather large decrease in PESQ compared to the continuous gain functions; this degradation may be explained by large gain fluctuations from frame to frame. Clearly, it is possible to reduce the fluctuations, e.g., through temporal smoothing procedures; we did, however, not pursue this idea further, as the resulting estimators hardly would qualify as binary anymore. In terms of

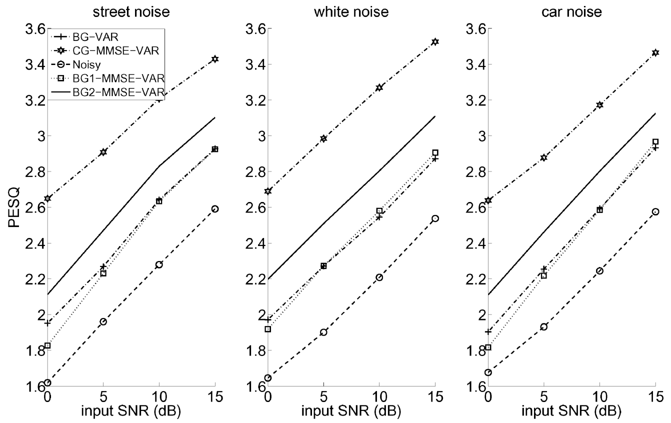

Fig. 8. Performance in terms of PESQ for the processing methods outlined in Table III (true *a priori* SNR known).

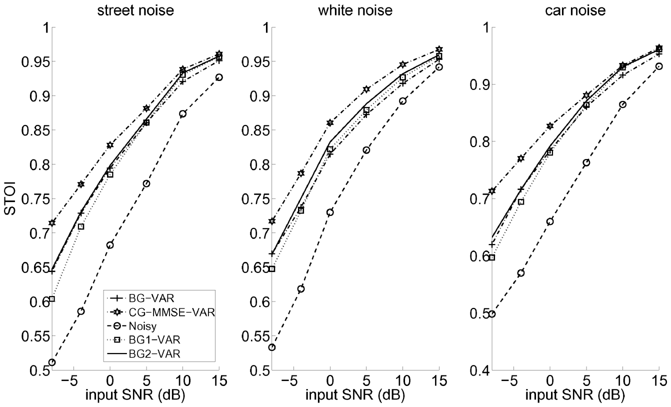

Fig. 9. Performance in terms of STOI for the processing methods outlined in Table III (true *a priori* SNR known).

STOI, Fig. 9, the binary gain functions give performance closer to that of continuous gain functions, especially for high SNRs.

We conclude:

- The Type 1 and 2 binary gain estimators generally improve over BG-VAR (criterion $C1$ in [10]), which is an example of a BG method that applies a *fixed a priori* SNR threshold to determine the binary gain value.
- The Type 2 binary gain estimator performs better than the Type 1 estimator, especially as measured by the PESQ quality measure. This is in line with informal subjective evaluations which indicate that BG2-MMSE produces slightly less musical artifacts than BG1-MMSE.
- Continuous gain methods outperform binary gain methods in the idealized scenarios where true *a priori* SNRs are available.

C. Only Noisy Realizations Available

We now turn to the realistic situation, where only the noisy signal is available. We compare the Type 1 and 2 binary MMSE estimators with other existing binary gain methods. We also compare them to a state-of-the-art continuous gain MMSE estimator to study the impact of using binary gain functions. The processing methods in this study are outlined in Table III. The methods BG-DD, BG1-MMSE, and BG2-MMSE rely on an estimate of the *a priori* SNR ξ. To this end, the decision-directed approach with a smoothing factor of $\alpha = 0.98$ was used [15]; note that this approach for estimating ξ requires a continuous gain function; using a binary gain function here degrades performance significantly. The method BG-HU is independent of *a priori* SNR but depends on the spectral noise variance only. The spectral noise variance was estimated using the noise tracker described in [32].

TABLE III
PROCESSING METHODS WHICH USE NOISY REALIZATIONS ONLY

Abbreviation	Description
CG-MMSE	Continuous gain MMSE estimator. Computes conditional mean $E(A\|r)$ assuming magnitude $f_A(a;\gamma = 1, \nu = 0.6)$, see Eq. (4) and [17].
BG-DD	Binary gain scheme proposed in [10] that thresholds the a priori SNR ξ to choose between a gain of 1 or 0. (Criterion $C1$)
BG-HU	The best of the binary gain schemes proposed in [10] (criterion $C4$).
BG1-MMSE	The Type 1 binary gain estimator derived in Sec. IV-A
BG2-MMSE	The Type 2 binary gain estimator derived in Sec. IV-B
Noisy	Unprocessed noisy speech to judge impact of estimators.

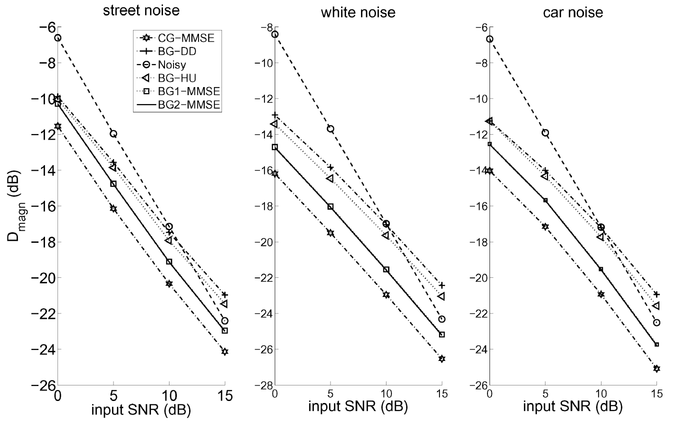

Fig. 10. Performance in terms of D_{magn} for the processing methods outlined in Table III.

Figs. 10–12 plot the performance of these processing methods as a function of input SNR as measured by D_{magn}, PESQ, and STOI for different noise types. Fig. 10 shows, as expected, that CG-MMSE is superior to any of the BG functions, while the Type 2 BG function is slightly better than the Type 1 function; this is in accordance with (15). The Type 1 and 2 BG functions perform better than BG-DD and BG-HU, because the latter were not designed to minimize D_{magn}. Notice that performance curves of the the Type 1 and 2 BG functions almost coincide in Fig. 10. The same ordering appears for PESQ and STOI. The difference between CG-MMSE and the best of the binary gain functions is as much as 0.4 PESQ points for the stationary noise sources, and slightly smaller for some input SNRs with street noise, see Fig. 11. Note that the STOI measure predicts that the intelligibility of the signals enhanced with CG-MMSE is similar to or slightly better than the unprocessed noisy speech; similar results have been reported in, e.g., [19],

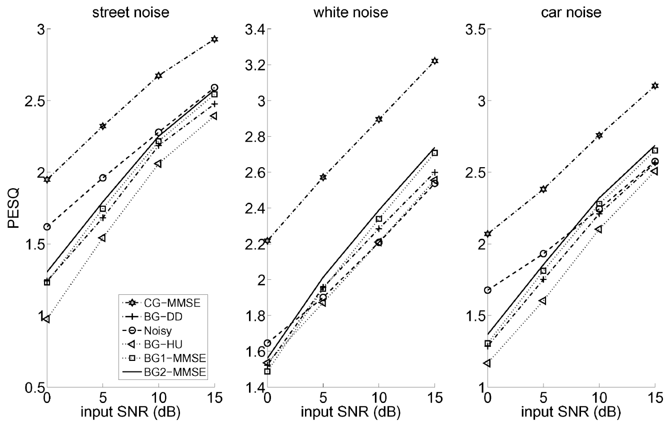

Fig. 11. Performance in terms of PESQ for the processing methods outlined in Table III.

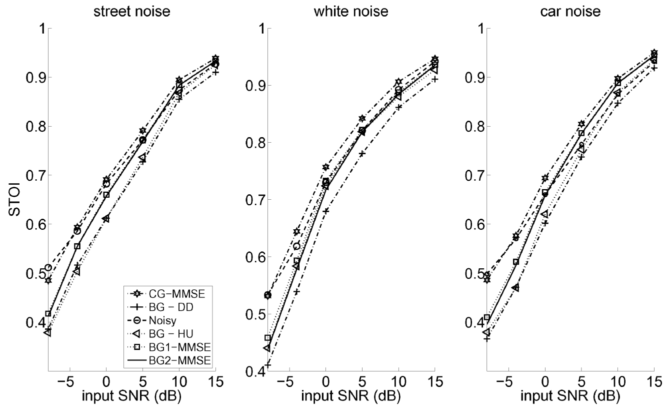

Fig. 12. Performance in terms of STOI for the processing methods outlined in Table III.

[33]. Also note the rather big difference in the STOI measure between the BG-MMSE variants and BG-DD and BG-HU.

We draw the following conclusions from the simulation results.

- The Type 1 and 2 BG estimators improve over state-of-the-art estimators BG-DD and BG-HU proposed in [10]. Specifically, PESQ improvements in the order of 0.1 are observed.
- The Type 1 and 2 BG estimators show similar performance.
- The CG-MMSE estimator is always better than any BG method, for all considered distortion measures. Improvements of PESQ scores as high as 0.4 are typical, while the distortion measure D_{magn} is reduced by approximately 1 dB.

Finally, in comparing Figs. 7–9 with Figs. 10–12, it is clear that the privileged, but practically unrealistic, situation of using correct *a priori* SNRs leads to significantly higher performance than the practical situation where the *a priori* SNR must be estimated from the available noisy data. Note also that with ideal SNR estimators, BG2-MMSE performs quite a lot better than BG1-MMSE, but that this difference vanishes in the practical situation. Thus, if better SNR estimators could be applied in the practical situation, we would expect BG2-MMSE to be the more preferable estimator.

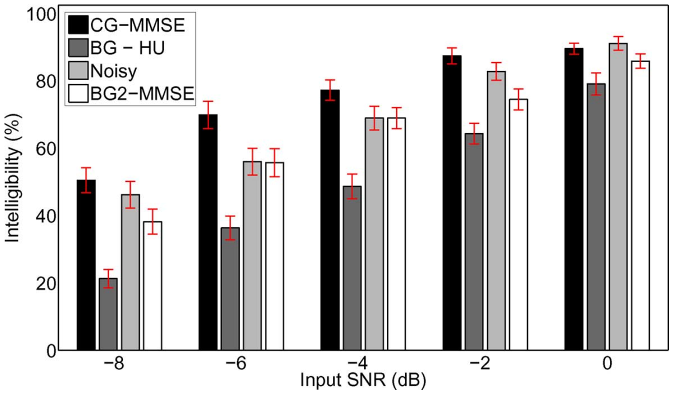

Fig. 13. Scores of intelligibility test for speech shaped noise as a function of global input SNR for noisy signal and processed variants.

D. Intelligibility Test

In this section, we compare the BG and CG functions by means of an intelligibility test. A subset of the algorithms from the previous section were evaluated: CG-MMSE, BG-HU and BG2-MMSE. The intelligibility of the noisy signal itself was also evaluated.

We conducted a closed Dutch speech-in-noise intelligibility test proposed in [34]. The sentences used in the test consist of five words with a correct grammatical structure, similar to the one proposed by Hagerman in [35]. The possible words are arranged in an 10-by-5 matrix on a computer screen, such that the ith column contains exactly the 10 possible alternatives for the ith word. The task of the listener is to select via a graphical user interface the words that she heard. For each test sentence, one word from each column must be selected. The signals are sampled at a sampling rate of 8 kHz, and degraded by speech-shaped noise at five different SNRs, namely -8, -6, -4, -2, and 0 dB. The noisy signals were processed with the three aforementioned algorithms. Thirteen native Dutch speaking, normal-hearing subjects participated in the test. Each processing condition was presented five times, and each sentence was used only once. The order of presenting the different algorithms and SNRs was randomized. The signals were presented diotically through head-phones (Sennheiser HD 600).

Fig. 13 shows average intelligibility scores including standard errors. The scores of the noisy signals are quantitatively in line with the results obtained in [7]. More specifically, Fig. 13 suggests a 50% speech reception threshold (SRT) between -6 and -8 dB, while [7] found an SRT of -7.7 dB (albeit for a different test paradigm, different speech material, and different sample rate). A t-test [36] was used to test whether for a certain input SNR intelligibility scores were significantly different between methods. Multiple paired t-tests with Bonferroni correction [36] were used to test whether for a certain input SNR, intelligibility scores were significantly different for the following five comparisons: 1) CG-MMSE and BG-HU; 2) the noisy signal and BG-HU; 3) BG2-MMSE and BG-HU; 4) CG-MMSE and BG2-MMSE; and 5) CG-MMSE and the noisy signal. The statistical significance level corrected by Bonferroni was set at $p < 0.01$ ($\alpha = 0.05$).

The main conclusions are: the noisy signals and the noisy signals processed by CG-MMSE had statistically significantly higher intelligibility than those processed by the BG-HU algorithm for all tested SNRs. The proposed algorithm BG2-MMSE leads to statistically significantly better intelligibility than the BG-HU algorithm for input SNRs from -8 to -2 dB. The intelligibility of CG-MMSE was statistically significantly better than the BG2-MMSE algorithm for input SNRs -8, -6 and -2 dB.

The average intelligibility scores for the CG-MMSE algorithm were higher than those of the noisy signal for input SNRs of -8 up to -2 dB. For an input SNR of -6 dB, the intelligibility of CG-MMSE was even statistically significantly better than the intelligibility of the noisy speech signal. This is a somewhat surprising result, as it is reported [19] that existing single-channel noise reduction algorithms generally do not improve the intelligibility. However, apart from results published in [19], [20], hardly any intelligibility test results for single-channel noise reduction have been published. The results published in [19] do not encompass the algorithm CG-MMSE tested in this paper. The intelligibility test in [20], on the other hand, *did* include the CG-MMSE algorithm, but had implementational differences. Specifically, [20] limited the maximum suppression to 10 dB for signal quality reasons, whereas the CG-MMSE version used here had no such limit. Moreover, the specific intelligibility test conducted in this article is a closed test, whereas the test performed in [20] is a semi-closed test. We do not have strong reasons to be believe that the performance difference is especially high at an input SNR of -6 dB, over any of the other tested SNRs. It might be, though, that for higher SNRs it is harder to improve the noisy signals because they are already quite intelligible, while at lower SNRs, the conditions under which the enhancement algorithms operate are so harsh that they do in fact not work well. A closer study of these hypotheses requires additional intelligibility tests and is a topic for future work.

VI. Conclusion

Ideal binary mask (idbm) techniques originate as a simple model for simulating the time–frequency analysis and grouping processes of the human auditory system. Given true local SNR values for various time–frequency tiles, these techniques are capable of recovering speech signals in severe noise conditions. Recently, idbm techniques have been adopted to the non-ideal situation where local SNR values must be estimated from the noisy observation, in the hope that the promising performance would remain in this new domain.

In the non-ideal domain, single-channel idbm techniques bear strong similarities to the class of DFT-based speech enhancement algorithms. Initially we demonstrate in this single-channel DFT framework that even for a sparse target signal contaminated by a sparse noise signal, a situation which has been hypothesized to be particularly favorable for the idbm techniques, the gain function which is optimal in spectral magnitude MMSE sense is far from binary; in this practical situation, making extreme decisions (such as applying a binary gain function) tends to be sub-optimal on average, because errors in decisions are inevitable, and the cost of making these errors is high.

Despite this discouraging initial result, we derive binary mask estimators which are optimal in MMSE sense. For these estimators, the SNR thresholds used for classifying a time–frequency region as target or noise dominated follow analytically, which is in contrast to the existing estimators where the threshold choice is often heuristically motivated. We also point out several links to existing binary mask estimators. Simulation experiments with noisy speech signals show that the derived MMSE binary gain functions generally perform better than existing binary estimators, as measured by objective quality (PESQ) and speech intelligibility (STOI) predictors. However, even the best of our binary gain functions was significantly outperformed by an MMSE estimator which was not constrained to be binary. The STOI results are supported by an intelligibility test with speech-shaped noise where the derived BG-MMSE estimators performed statistically significantly better than an existing binary estimator, but the CG-MMSE functions were statistically significantly better than any binary gain function for a majority of the tested SNRs.

We conclude that while there may be good reasons for applying binary gain techniques in the non-ideal situation where only a noisy observation is available, e.g., reduced computational and storage complexity, there is a potential significant performance loss associated with these methods, at least in the DFT framework considered here. Although we have focused here on the spectral magnitude MMSE criterion, for other, perhaps perceptually more meaningful, criteria such as spectral magnitude log-MMSE, these conclusions remain. As is the case for continuous gain functions, accurate estimation of the local SNR in a given time–frequency region is of critical importance. While improved SNR estimators are worth pursuing for both binary and continuous gain functions, it may be noted that for the binary gain functions, this SNR estimation could easily dominate the total complexity of the algorithm such that the minor complexity reduction realized by applying binary gain techniques might not be justified by the potential performance loss in doing so.

Acknowledgment

The authors would like to thank the three anonymous reviewers for their comments which helped improve the presentation of this work. J. Jensen would like to thank his family for their support and for providing the conditions to conduct this research.

References

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[2] D. L. Wang and G. J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications Hoboken, NJ, Wiley/IEEE Press, 2006.

[3] G. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. Workshop Applicat. Signal Process. Audio Acoust.*, 2001, pp. 79–82.

[4] D. L. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplificat.*, vol. 12, no. 4, pp. 332–353, Dec. 2008.

[5] N. Li and C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[6] Y. Li and D. L. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2008.

[7] U. Kjems *et al.*, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1415–1426, Sep. 2009.

[8] D. Brungart *et al.*, "Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, Dec. 2006.

[9] M. C. Anzalone, "Determination of the potential benefit of time–frequency gain manipulation," *Ear Hear.*, vol. 27, pp. 480–492, 2006.

[10] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust., Echo, Noise Control*, 2008.

[11] G. Kim, Y. Lu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listener," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, Sep. 2009.

[12] P. Vary and R. Martin*, Digital Speech Transmission: Enhancement, Coding, and Error Concealment*. New York: Wiley, 2006.

[13] P. C. Loizou*, Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[14] 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.

[15] 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.

[16] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. I, pp. 253–256.

[17] J. 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, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[18] Y. Hu and P. Loizou, "Subjective comparison and evaluations of speech enhancement methods," in *Speech Commun.*, 2007, vol. 49, Special Iss. on Speech Enhancement, pp. 588–601.

[19] Y. Hu and P. C. Loizou, "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 4, pp. 561–564.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Proc. Int. Workshop, Acoust. Echo Noise Control*, 2010.

[21] D. R. Brillinger*, Time Series – Data Analysis and Theory*. Philadelphia, PA: Soc. for Ind. and Appl. Math., 1981.

[22] I. Cohen, "Relaxed statistical model for speech enhancement and *a priori* SNR estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 870–881, Sep. 2005.

[23] 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.

[24] J. Jensen and R. Heusdens, "A numerical approach for estimating optimal gain functions in single-channel dft based speech enhancement," in *Proc. Eusipco*, 2006.

[25] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun.*, vol. 49, pp. 530–541, Jul./Aug. 2007.

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

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

[28] 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, Sep. 2005.

[29] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, vol. 1, pp. 153–156.

[30] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU, ITU-T Rec. P.862 (02/2001), 2001.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time–frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

[32] R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "Low complexity dft-domain noise psd tracking using high-resolution periodograms," *EURASIP J. Adv. in Signal Process.*, vol. 2009, 2009.

[33] C. H. Taal, "Predicting the effect of single-channel noise reduction on speech intelligibility," in *Proc. IEEE Benelux Signal Process. Symp.: Signal Process. for Digital Hear. Aids*, Delft, The Netherlands, 2010.

[34] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," in *Proc. 8th EFAS Congr., 10th DGA Congr.*, Heidelberg, Germany, Jun. 2007.

[35] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scand. Audiol.*, vol. 11, no. 2, pp. 79–87, 1982.

[36] D. J. Sheskin*, Parametric and Nonparametric Statistical Procedures*, 3rd ed. Boca Raton, FL: Chapman & Hall/CRC, 2004.

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

From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is with Oticon A/S, Copenhagen, Denmark. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, and perceptual aspects of signal processing.

Richard C. Hendriks received the B.Sc., M.Sc., (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003 and 2008, respectively.

From 2003 to 2007, he was a Ph.D. Researcher at the Delft University of Technology. From 2007 to 2010, he was a Post-Doctoral Researcher at Delft University of Technology. Since 2010, he has been an Assistant Professor in the Multimedia Signal Processing Group, Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology. In the autumn of 2005, he was a Visiting Researcher at the Institute of Communication Acoustics, Ruhr-University Bochum, Bochum, Germany. From March 2008 to March 2009, he was a Visiting Researcher at Oticon A/S, Copenhagen, Denmark. His main research interests are digital speech and audio processing, including single-channel and multi-channel acoustical noise reduction, speech enhancement, and intelligibility improvement.