SPECTRAL MAGNITUDE MINIMUM MEAN-SQUARE ERROR BINARY MASKS FOR DFT BASED SPEECH ENHANCEMENT

Jesper Jensen

Oticon A/S, Denmark
jsj@oticon.dk

Richard C. Hendriks[*]

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

ABSTRACT

Originally, ideal binary mask (idbm) techniques have been used as a tool for studying aspects of the auditory system. More recently, idbm techniques have been adapted to the practical problem of retrieving a target speech signal from a noisy observation. In this practical setting, the binary mask techniques show similarities with existing DFT based speech enhancement techniques. In this context, we derive single-channel, binary mask estimators which minimize the spectral magnitude mean-square error. We show in simulation experiments with natural speech and noise signals that the proposed estimators perform significantly better than existing binary mask estimators. However, even the best of the proposed estimators is clearly outperformed by non-binary estimators, both in terms of speech quality and intelligibility.

Index Terms— Speech enhancement, binary masks, minimum mean-square error, intelligibility.

1. INTRODUCTION

Originally, ideal binary mask (idbm) techniques were proposed as a signal processing tool for simulating and studying the time-frequency analysis and grouping process of the auditory system, e.g. [1, 2]. In the simplest setting, it is assumed that a target signal $s(n)$ is contaminated by an additive noise source $w(n)$, resulting in a noisy signal $x(n) = s(n) + w(n)$. The idbm techniques decompose the signals in a suitable time-frequency (TF) representation, e.g. through short-time discrete Fourier transforms (DFTs), resulting in TF units $x(k, m)$, $s(k, m)$, and $w(k, m)$, where k and m are frequency and time indices, respectively. To retain TF units with high signal-to-noise ratio (SNR), and suppress TF units with low SNR, a binary gain (BG) function $g(k, m)$ is multiplied onto the noisy TF units $x(k, m)$. The BG function is found by comparing the local SNR $|s(k, m)|^2/|w(k, m)|^2$ to a threshold $\rho(k, m)$, i.e.,

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

where $g_{max} > g_{min} \geq 0$. The gain modified TF units are transformed back to time domain e.g. using inverse DFTs and overlap-add techniques. Interestingly, idbm techniques can render noisy signals with an SNR as low as -60 dB essentially perfectly intelligible [3]. Note, however, that to achieve this, the local SNR realizations $|s(k, m)|^2/|w(k, m)|^2$ must be known with certainty.

Perhaps motivated by these impressive intelligibility improvements, the idbm frame work has more recently been adapted to the practical problem of retrieving a speech signal from a noisy observation, when local SNRs are *not* known, e.g. [4, 5]. This methodology bears strong similarities to the class of DFT based speech enhancement methods, e.g. [6], the main difference being that the speech enhancement methods apply a *continuous* gain (CG) function rather than a binary one to the noisy TF units.

Observing that existing BG functions tend to be heuristically motivated, we present in this paper BG functions which are optimal in minimum mean-square error (mmse) sense. The goal of the paper is twofold. First, we wish to derive theoretically optimal BG functions and compare their performance to existing BG estimators. Secondly, we wish to study the performance difference between these optimal BG estimators, and their continuous-valued counterparts.

2. SIGNAL MODEL AND NOTATION

We use capital letters to denote random variables and the corresponding lower case letter for their realizations. We consider an additive signal model of the form

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

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 target and the noise, respectively. We use the standard assumptions that $S(k, m)$ and $W(k, m)$ are statistically independent, and that DFT coefficients are approximately independent across time and frequency [7]; thus, without loss of generality, we may drop time and frequency indices, and simply write

$$X = S + W.$$

Let $R = |X|$, $A = |S|$, and $N = |W|$ represent the noisy, clean and noise spectral magnitude, respectively. Furthermore, let $\xi = \sigma_S^2/\sigma_W^2$ and $\zeta = r^2/\sigma_W^2$ denote the *a priori* and *a posteriori* SNR [6], respectively, with spectral variances given by $\sigma_S^2 = E(A^2)$ and $\sigma_W^2 = E(N^2)$.

We assume that speech spectral magnitudes $A \geq 0$ are distributed according to a probability density function (pdf) of the form

$$f_A(a; \gamma, \nu) = \frac{\gamma \beta^\nu}{\Gamma(\nu)} a^{\gamma\nu - 1} \exp\left(-\beta a^\gamma\right), \ \gamma > 0, \nu > 0, \qquad (2)$$

where $\Gamma(\cdot)$ is the gamma function. Specifically, we focus on $f_A(a; \gamma = 1, \nu = 0.6)$ which is a good model of the speech DFT magnitude distribution [8, 9] in our context. The corresponding phase variable is assumed to be uniformly distributed in $[0; 2\pi[$, and independent of A. For given parameters γ, ν, and magnitude variance σ_A^2, the parameter $\beta > 0$ is fully determined; for $\gamma = 1$, $\beta = \sqrt{\nu(\nu + 1)/\sigma_A^2}$ [9]. Furthermore, we assume that W obeys a

[*]The research is supported by the Dutch Technology Foundation STW.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

(complex) Gaussian distribution, leading to a Rayleigh distribution for N [8].

3. MMSE BINARY MASK SPECTRAL MAGNITUDE ESTIMATORS

In this section we derive spectral magnitude MMSE *binary* gain functions. We treat two types of binary gain functions: Type 1 which is constrained to be non-decreasing in the aposteriori SNR ζ, and Type 2, where this constraint has been relaxed.

3.1. MMSE Binary Gain Function (Type 1)

In the idbm frame work, a low gain is applied in low SNR regions and a high gain is applied in high SNR regions, Eq. (1). To be in line with this, we consider the Type 1 BG function $g(k,m)$ of the form

$$g(k,m) = \begin{cases} 1 & \text{for } \zeta(k,m) > \rho(k,m) \\ \epsilon & \text{otherwise,} \end{cases} \tag{3}$$

where the realization based SNR from Eq. (1), which is difficult to estimate reliably in practice, has been replaced with the *a posteriori* SNR $\zeta(k,m) = r^2(k,m)/\sigma_N^2(k,m)$, which is easier to estimate.

Using Eq. (3), the mean-square error $J_1 = E(A-\hat{A})^2$ is

$$J_1 = \int_0^{\tilde{\rho}} \int_A (a-\epsilon r)^2 f_{A|r}(a) f_R(r) da dr + \\ \int_{\tilde{\rho}}^{\infty} \int_A (a-r)^2 f_{A|r}(a) f_R(r) da dr,$$

where $\tilde{\rho} = \sqrt{\rho \sigma_N^2}$ is the noisy magnitude corresponding to the threshold ρ. To find the threshold that minimizes J_1, we use Leibniz' rule [10, 0.410] to compute

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

Solving $\frac{\partial J_1}{\partial \tilde{\rho}} = 0$, we find that the optimal threshold satisfies

$$\frac{E(A|r=\tilde{\rho})}{\tilde{\rho}} = \frac{1}{2}(1+\epsilon). \tag{4}$$

Recall that the conditional mean $E(A|r)$ is identical to the MMSE estimator (e.g. [6]), i.e., $E(A|r) = g_{MMSE}(r) \cdot r$, where $g_{MMSE}(\cdot)$ is the MMSE CG function. It then follows that the left-hand side of Eq. (4) is $g_{MMSE}(r = \tilde{\rho})$, and the optimal threshold $\tilde{\rho}^*$ is simply the value of r for which the MMSE CG function is equal to $\frac{1}{2}(1+\epsilon)$. Fig. 1 shows examples of MMSE CG functions (CG-MMSE) and the corresponding BG functions of Type 1 (BG1-MMSE) for different choices of *a priori* SNR ξ.

The optimal threshold $\tilde{\rho}^*$ follows analytically from Eq. (4) as a function of *a priori* SNR ξ, and g_{max} and g_{min}. This is in contrast to the idbm schemes, which tend to choose the threshold more heuristically, e.g. [5]. Fig. 2a summarizes the BG1-MMSE estimator.

3.2. MMSE Binary Gain Function (Type 2)

The Type 1 binary gain function in Eq. (3) is a non-decreasing function of the *a posteriori* SNR ζ, see Fig. 1. 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 Sec. 4, that this leads to a performance advantage.

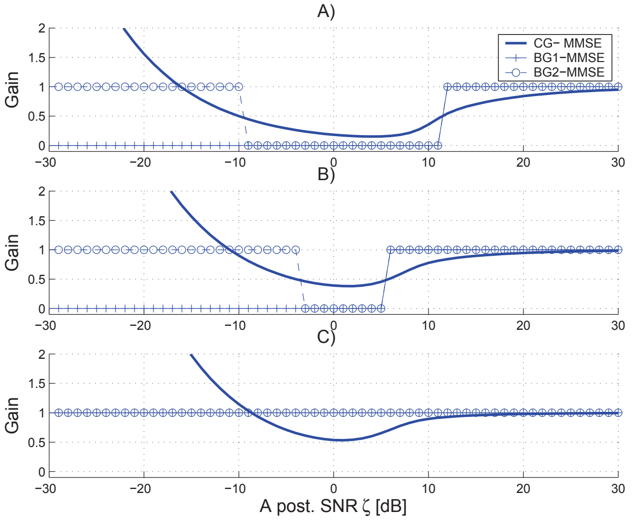


Fig. 1. Continuous and binary MMSE gain functions ($\epsilon = 0$) for target pdf $f_A(a;\gamma = 1, \nu = 0.6)$, for *a priori* SNRs (A) $\xi = -10$ dB, (B) $\xi = 0$ dB, and (C) $\xi = 10$ dB. The noise DFT magnitudes are assumed to follow a Rayleigh distribution.

Consider therefore the mean-square error $J_2 = E(A-\hat{A})^2$,

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

where $T(r)$ is a function of the noisy magnitude realization r. As the gain function is constrained to be binary, $g(r) \in \{\epsilon, 1\}, \epsilon < 1$, there exist two possible values of $T(r)$ for a given r, namely

$$T_1(r) \triangleq E(A^2|r) + r^2 - 2E(A|r)r, \text{ for } g(r) = 1, \text{ and} \tag{5}$$

$$T_\epsilon(r) \triangleq E(A^2|r) + \epsilon^2 r^2 - 2\epsilon E(A|r)r \text{ for } g(r) = \epsilon. \tag{6}$$

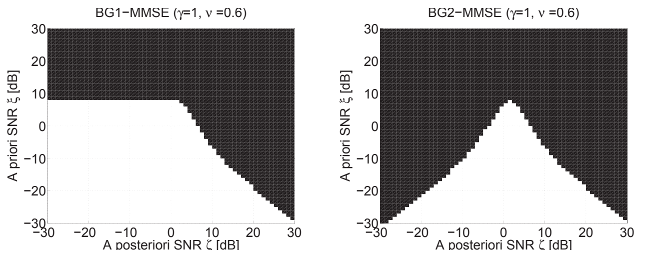
Thus, for $T_1(r) < T_\epsilon(r)$, the optimal gain value is $g(r) = 1$, and otherwise, $g(r) = \epsilon$. From Eqs. (5) and (6) it follows that $T_1(r) < T_\epsilon(r)$ when $g_{MMSE}(r) > \frac{1}{2}(1+\epsilon)$. We conclude that the MMSE BG function of Type 2 (BG2-MMSE) can be described as

$$g(r) = \begin{cases} 1 \text{ for } r \in \{r : g_{MMSE}(r) > \frac{1}{2}(1+\epsilon)\} \\ \epsilon \text{ otherwise.} \end{cases}$$

Note that BG2-MMSE is simply a quantized version of the MMSE CG function $g_{MMSE}(r)$. Fig. 1 shows examples of BG2-MMSE functions for $f_A(a;\gamma = 1, \nu = 0.6)$, and Fig. 2b summarizes the BG2-MMSE estimator.

4. SIMULATION RESULTS

We compare the derived binary estimators and existing methods in simulation experiments using speech and noise signals from the Noizeus data base [11]. The signals are sampled at a rate of 8 kHz. The car noise signal (which is roughly stationary) and the street noise signal (which is more non-stationary) were added to the



(A) (B)

Fig. 2. Complete gain functions BG1-MMSE (A), and BG2-MMSE (B) for target source $f_A(a; \gamma = 1, \nu = 0.6)$. Black: $g = 1$, White: $g = 0$.

speech signals to form noisy signals at SNRs 15, 10, 5, and 0 dB. The signals are processed in frames of 256 samples, with an overlap of 50 %. The signal frames are weighted with a square-root Hann window, and a DFT is applied. Gain functions are computed and applied to the noisy DFT coefficients, before an IDFT is performed. The resulting enhanced time domain frames are overlap-added with a square-root Hann window to form the enhanced signal.

4.1. Objective Evaluations

To quantify the quality of the enhanced signals, we apply the PESQ speech quality measure [12]. To gain an indication of the intelligibility of the enhanced signals, we use the recently developed Short-Time Objective Intelligibility (STOI) measure [13]. STOI outputs an average correlation coefficient $-1 \leq d_{STOI} \leq 1$ 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 [13] and [14], respectively.

We compare the derived BG estimators to existing BG methods, and to a state-of-the-art MMSE CG estimator to judge the impact of restricting the gain function to be binary. The processing methods in this study are outlined in Table 1. The BG methods in Table 1 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 [6]. Note that estimating ξ in this way implies using a continuous gain function; using a binary gain function here degrades performance significantly. The spectral noise variance was estimated using the noise tracker described in [15].

Abbreviation	Description
CG-MMSE	Continuous gain MMSE estimator. Computes conditional mean $E(A\|r)$ assuming magnitude pdf $f_A(a; \gamma = 1, \nu = 0.6)$, see Eq. (2) and [9].
BG-DD	Binary gain scheme proposed in [5] that thresholds the *a priori* SNR ξ to select gain of 1 or 0. (Criterion $C1$ in [5])
BG-HU	The best of the binary gain schemes proposed in [5] (criterion $C4$ in [5]).
BG1-MMSE	The Type 1 binary gain estimator from Sec. 3.1
BG2-MMSE	The Type 2 binary gain estimator from Sec. 3.2
Noisy	Unprocessed noisy speech.

Table 1. Processing methods used in the simulation study.

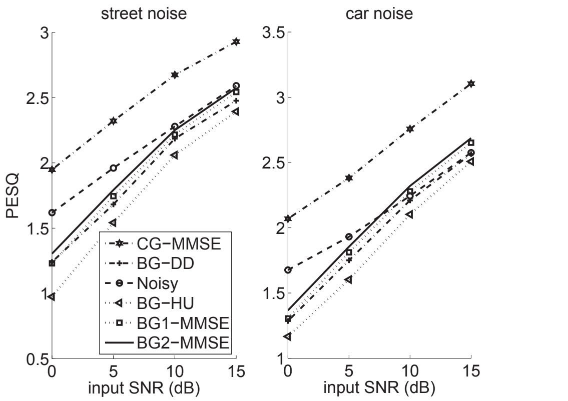


Fig. 3. Performance in terms of PESQ for the processing methods outlined in Table 1.

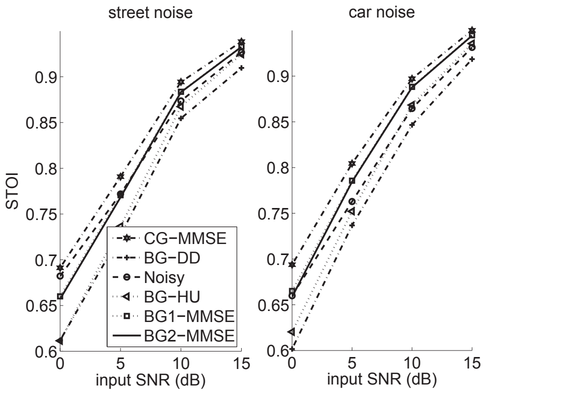


Fig. 4. Performance in terms of STOI for the processing methods outlined in Table 1.

Figs. 3 and 4 plot the performance of the processing methods as a function of input SNR as measured by PESQ and STOI. For both distortion measures, the derived MMSE BG estimators perform better than the existing methods BG-DD and BG-HU; Fig. 3 shows PESQ improvements in the order of 0.1. However, using a continuous gain function significantly outperforms any of the binary gain functions. The difference between CG-MMSE and the best of the BG functions, BG2-MMSE, is as much as 0.4 PESQ points for the stationary car noise, and slightly smaller for street noise. From Fig. 4, note that STOI predicts the intelligibility of the signals enhanced with CG-MMSE to be similar to or slighly better than the unprocessed noisy speech; similar results have been reported in e.g. [14]. Also note the rather big STOI difference between the BG-MMSE variants and BG-DD and BG-HU.

4.2. Intelligibility Test

The idbm framework has often been used to study the impact on intelligibility of various processing conditions. To study this aspect in the presented non-ideal context, an intelligibility test was conducted with a subset of the algorithms from the previous section: CG-MMSE, BG-HU and BG2-MMSE.

We use the closed Dutch speech-in-noise intelligibility test proposed in [16]. The test sentences consist of five words with a correct

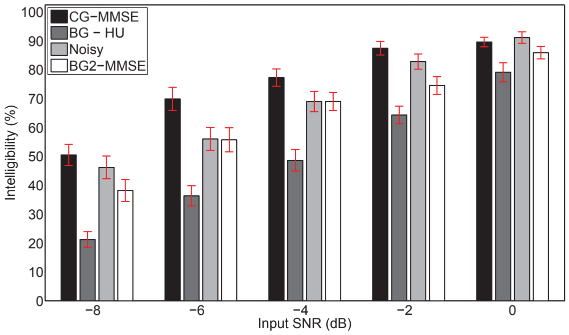


Fig. 5. Intelligibility test scores for speech shaped noise as a function of global input SNR for noisy signal and processed variants.

grammatical structure. The signals are sampled at a rate of 8 kHz, and degraded by speech-shaped noise at five fixed SNRs, -8, -6, -4, -2 and 0 dB. The noisy signals were processed with the three aforementioned algorithms.

Thirteen native Dutch speaking people participated in the test. Each processing condition was presented 5 times, and each sentence was used only once. The presentation order of the different algorithms and SNRs was randomized. The signals were presented diotically through head-phones (Sennheiser HD 600).

Fig. 5 shows the average intelligibility scores including the standard errors. Fig. 5 suggests a 50 % speech reception threshold (SRT) between -6 and -8 dB, which is in line with [3] where an SRT of 7.7 dB was found (albeit for a different test paradigm, different speech material, and different sample rate). A t-test [17] was used to judge the difference between methods for specific SNRs. The main conclusions (statistical significance level $\alpha = 0.05$) are: the noisy signals and signals processed with BG2-MMSE and CG-MMSE have statistically significantly higher intelligibility scores than signals processed with BG-HU, for all SNRs. Secondly, CG-MMSE is statistically significantly better than BG2-MMSE for SNRs -8 through -2 dB. Finally, CG-MMSE signals have statistically significantly higher intelligibility than the noisy signals for SNRs -6 and -4 dB. This last point is slightly surprising as it is generally reported that existing single-channel noise reduction algorithms are not able to improve the intelligibility [18]; this point is currently under further study.

5. CONCLUSIONS

We have derived binary mask MMSE estimators for speech spectral magnitudes. Compared to existing binary gain functions, the proposed estimators lead to better speech quality as measured by PESQ, and significantly better intelligibility as measured by objective measures and intelligibility listening tests. However, existing *non-binary* estimators significantly outperform the binary estimators, both in terms of speech quality and intelligibility. We conclude that while binary mask estimators may offer advantages in terms of storage and computational complexity, there is a potential significant performance loss associated with these methods.

6. REFERENCES

[1] G. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 79–82.

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

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

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

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

[6] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[7] D. R. Brillinger, *Time Series - Data Analysis and Theory*, Society for Industrial and Applied Mathematics, 1981.

[8] 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.

[9] J. S. Erkelens et al., "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients With Generalized Gamma Priors," *IEEE Trans. Audio., Speech, Language Processing*, vol. 15, no. 6, pp. 1741–1752, August 2007.

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

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

[12] ITU, *ITU-T Recommendation P.862 (02/2001) Perceptual Evaluation of Speech Quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.*, 2001.

[13] C. H. Taal et al., "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech,," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2010, pp. 4214–4217.

[14] C. H. Taal et al., "On predicting the difference in intelligibility before and after single-channel noise reduction," in *Int. Workshop, Acoustic Echo and Noise Control*, 2010.

[15] R. C. Hendriks et al., "Low Complexity DFT-Domain Noise PSD Tracking Using High-Resolution Periodograms," *Eurasip Journal on Advances in Signal Processing*, 2009.

[16] J. Koopman et al., "Development of a speech in noise test (matrix)," in *8th EFAS Congress, 10th DGA Congress*, Heidelberg, Germany, 2007.

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

[18] Y. Hu and P. C. Loizou, "A Comparative Intelligibility Study of Speech Enhancement Algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2007, vol. 4, pp. 561–564.