Estimators of the Magnitude-Squared Spectrum and Methods for Incorporating SNR Uncertainty

Yang Lu and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—Statistical estimators of the magnitude-squared spectrum are derived based on the assumption that the magnitude-squared spectrum of the noisy speech signal can be computed as the sum of the (clean) signal and noise magnitude-squared spectra. Maximum *a posterior* (MAP) and minimum mean square error (MMSE) estimators are derived based on a Gaussian statistical model. The gain function of the MAP estimator was found to be identical to the gain function used in the ideal binary mask (IdBM) that is widely used in computational auditory scene analysis (CASA). As such, it was binary and assumed the value of 1 if the local signal-to-noise ratio (SNR) exceeded 0 dB, and assumed the value of 0 otherwise. By modeling the local instantaneous SNR as an F-distributed random variable, soft masking methods were derived incorporating SNR uncertainty. The soft masking method, in particular, which weighted the noisy magnitude-squared spectrum by the *a priori* probability that the local SNR exceeds 0 dB was shown to be identical to the Wiener gain function. Results indicated that the proposed estimators yielded significantly better speech quality than the conventional minimum mean square error spectral power estimators, in terms of yielding lower residual noise and lower speech distortion.**

***Index Terms*—Binary mask, maximum *a posterior* (MAP) estimators, minimum mean square error (MMSE) estimators, soft mask, speech enhancement.**

I. Introduction

A NUMBER of estimators of the signal magnitude spectrum have been proposed for speech enhancement (see review in [1, Ch. 7]). The minimum mean square error (MMSE) estimators [2], [3] of the magnitude spectrum, in particular, have been found to perform consistently well, in terms of speech quality, in a number of noisy conditions [4]. Several MMSE estimators of the power spectrum [5]–[7] or more general the pth-power magnitude spectrum [8] have also been proposed. In some applications such as speech coding [6], where the autocorrelation coefficients might be needed, the optimal power-spectrum estimator might be more useful than the magnitude estimator. Some [9], [10] have also incorporated the power-spectrum estimator in the "decision-directed" approach used for the computation of the *a priori* signal-to-noise ratio (SNR). This was based on the justification that the MMSE estimator of the power-spectrum is not equivalent to the square of the MMSE estimator of the magnitude spectrum, which is often used in the implementation of the "decision-directed" approach.

Analysis of the attenuation curves of the MMSE estimators of the pth-power magnitude spectrum revealed that these estimators provide less attenuation than the linear and log-MMSE estimators, at least for $p \geq 2$ [8]. This in turn leads to substantial residual noise. In this paper, we derive estimators of the short-time power-spectrum, henceforth denoted as magnitude-squared spectrum, which markedly reduce the residual noise without introducing speech distortion. Maximum *a posteriori* (MAP) estimators and MMSE estimators of the magnitude-squared spectrum are derived. A number of MAP estimators of the magnitude spectrum have been proposed [11], [7], [12]–[14] in the literature, but no MAP estimators of the magnitude-squared spectrum have been reported. Furthermore, no closed form solutions of the MAP estimators of the magnitude spectrum were derived in prior studies without resorting to some approximations to the underlying density or the Bessel function. In contrast, no approximations are used in the derivation of the proposed MAP estimator of the magnitude-squared spectrum. The proposed MMSE and MAP estimators are derived using a Gaussian statistical model [2] and the assumption that the magnitude-square spectrum of the noisy speech signal can be computed as the sum of the (clean) signal and noise magnitude-squared spectra. This assumption has been used widely in spectral subtraction algorithms [15]–[20], as well as in statistical-model based speech enhancement algorithms [5], and is known to hold statistically assuming that the signal and noise are independent and zero mean. Under some conditions [21], this assumption also holds in the instantaneous case, i.e., for short-time magnitude-squared spectra.

Of particular interest in this paper is the derived gain function of the MAP estimator of the magnitude-square spectrum, which is shown to be the same as the ideal binary mask. The ideal binary mask is a simple technique which is widely used in the computational auditory scene analysis (CASA) field [22]. The ideal binary mask can be considered as a binary gain function which assumes the value of 1 if the local SNR at a particular time–frequency (T-F) unit is larger than a threshold, and assumes the value of 0 otherwise. When the ideal binary mask is applied to the spectrum (computed using either the FFT or a filterbank) of the noisy speech signal, it can synthesize a signal with high intelligibility even at extremely low SNR levels (-5, -10 dB) [23], [24]. The optimality of the ideal binary mask, in terms of maximizing the SNR, was analyzed in [25]. The concept of the ideal binary mask has been motivated by auditory

Manuscript received November 05, 2009; revised May 20, 2010, September 15, 2010; accepted September 16, 2010. Date of publication September 30, 2010; date of current version May 04, 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sharon Gannot.

Y. Lu was with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, 75080 USA. He is now with Cirrus Logic, Inc. Austin, TX 78746 USA (e-mail: luyang1980@gmail.com).

P. C. Loizou is with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: loizou@utdallas.edu).

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.2010.2082531

masking principles [26], but has not been derived thus far analytically using known statistical techniques. A theoretical formulation of the ideal binary mask is presented in this paper, along with some new techniques for estimating the binary mask. As the construction of the MAP gain function relies on estimates of the SNR at each frequency bin, new estimators are proposed that incorporate SNR uncertainty. The SNR thresholding rule used in the ideal binary mask bears resemblance to the "hard-thresholding" rule used in wavelet denoising [27]–[29]. The similarities and dissimilarities of the ideal binary mask with the wavelet shrinkage rules are discussed.

This paper is organized as follows. Section II presents the background information, and Section III presents the assumptions, and also derives the MMSE estimator that uses these assumptions. The derivation of MAP estimator is presented in Section III-C. Section IV presents the details of soft mask estimators incorporating SNR uncertainty, and also analyzed the relationship between these estimators and binary masking. Section V provides the implementation details, Section VI presents the experimental results, and finally Section VII gives the conclusions.

II. Background

Let $y(n) = x(n) + d(n)$ denote the noisy signal, with $x(n)$ and $d(n)$ representing the clean speech and noise signals, respectively. Taking the short-time Fourier transform of $y(n)$, we get

$$Y(\omega_k) = X(\omega_k) + D(\omega_k). \tag{1}$$

The above equation can also be expressed in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{2}$$

where $\{Y_k, X_k, D_k\}$ denote the magnitudes and $\{\theta_y(k), \theta_x(k), \theta_d(k)\}$ denote the phases at frequency bin k of the noisy speech, clean speech, and noise, respectively.

Wolfe and Godsill [7] proposed the following MMSE estimator of the short-time power spectrum (MMSE-SP):

$$\begin{aligned}\hat{X}_k^2 &= E\left\{X_k^2 \mid Y(\omega_k)\right\} \\ &= \int_0^\infty X_k^2 f_{X_k}(X_k \mid Y(\omega_k)) dX_k \\ &= \frac{\xi_k}{1+\xi_k}\left(\frac{1}{\gamma_k} + \frac{\xi_k}{1+\xi_k}\right) Y_k^2\end{aligned} \tag{3}$$

where

$$\xi_k \equiv \frac{\sigma_x^2(k)}{\sigma_d^2(k)}, \quad \gamma_k \equiv \frac{Y_k^2}{\sigma_d^2(k)} \tag{4}$$

$$\sigma_x^2(k) \equiv E\left\{X_k^2\right\}, \quad \sigma_d^2(k) \equiv E\left\{D_k^2\right\}. \tag{5}$$

and ξ_k and γ_k denote the *a priori* and *a posteriori* SNRs, respectively. The derivations of the above MMSE estimator as well as the MAP estimator were based on the following Rician posterior density $f_{X_k}(X_k \mid Y(\omega_k))$:

$$f_{X_k}(X_k \mid Y(\omega_k)) = \frac{X_k}{\sigma_k^2} \exp\left(-\frac{X_k^2 + s_k^2}{2\sigma_k^2}\right) I_0\left(\frac{X_k s_k}{\sigma_k^2}\right) \tag{6}$$

where

$$\frac{1}{\lambda'(k)} \equiv \frac{1}{\sigma_x^2(k)} + \frac{1}{\sigma_d^2(k)} \tag{7}$$

$$\upsilon_k \equiv \frac{\xi_k}{1+\xi_k}\gamma_k. \tag{8}$$

$$\sigma_k^2 \equiv \frac{\lambda'(k)}{2}, \quad s_k^2 \equiv \upsilon_k \lambda'(k) \tag{9}$$

and $I_0(\cdot)$ is the first kind modified Bessel function of zeroth order. Approximations of the $I_0(\cdot)$ Bessel function were found necessary in [7] and [14] in order to derive the MAP estimator of the magnitude spectrum. Analysis of the suppression curves in [7] revealed that the MMSE spectral power suppression rule of (3) follows that of the MMSE magnitude estimator [2] closely, but provides less suppression in regions of low *a priori* SNR.

The proposed estimators of the short-time power-spectrum will be compared against the above estimator.

III. Proposed Magnitude-Squared Estimators

A. Statistical Model and Assumptions

Assuming that $x(n)$ and $d(n)$ are uncorrelated stationary random processes, the power spectrum of the noise-corrupt signal, $P_y(\omega)$ is simply the sum of the power spectra of the clean speech and noise

$$P_y(\omega) = P_x(\omega) + P_d(\omega). \tag{10}$$

The above assumption is true only in the statistical sense. However, taking this assumption as a reasonable approximation for short-term (20 ms in this paper) spectra, its application can lead to simple noise reduction methods [16].

Two assumptions are used in the derivation of the proposed estimators. The first assumption used in this paper is based on (10) by approximating the power spectrum using the magnitude-squared spectrum, which is the sample estimate of the ensemble average. Therefore, we rewrite (10) as follows:

$$Y_k^2 \approx X_k^2 + D_k^2. \tag{11}$$

Note that X_k^2 is limited in $[0, Y_k^2]$ due to (11). The above approximation is in fact widely used in all spectral subtractive algorithms [16]–[20], as well as in statistical-model based speech enhancement algorithms [5]. Analysis in [21] indicated that in high or low SNR conditions, (11) still holds in the instantaneous sense.

In the rest of the paper, we will be referring to Y_k^2, X_k^2, and D_k^2 as the magnitude-squared spectra of the noisy, clean and noise signals, respectively.

The second assumption is that the real and imaginary parts of the discrete Fourier transform (DFT) coefficients are modeled as independent Gaussian random variables with equal variance [2], [30]. Consequently, the probability density of X_k^2 is exponential [31, p. 190], and is given by

$$f_{X_k^2}\left(X_k^2\right) = \frac{1}{\sigma_x^2(k)} e^{-\frac{X_k^2}{\sigma_x^2(k)}}. \tag{12}$$

Similarly, the density of D_k^2 is given by

$$f_{D_k^2}\left(D_k^2\right) = \frac{1}{\sigma_d^2(k)} e^{-\frac{D_k^2}{\sigma_d^2(k)}} \tag{13}$$

where $\sigma_x^2(k)$ and $\sigma_d^2(k)$ are given by (5).

The posterior probability density of the clean speech magnitude-squared spectrum can be obtained using the Bayes' rule as follows

$$\begin{aligned} f_{X_k^2}\left(X_k^2 \,|\, Y_k^2\right) &= \frac{f_{Y_k^2}\left(Y_k^2 \,|\, X_k^2\right) f_{X_k^2}\left(X_k^2\right)}{f_{Y_k^2}\left(Y_k^2\right)} \\ &= \begin{cases} \Psi_k e^{-\frac{X_k^2}{\lambda(k)}}, & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{Y_k^2}, & \text{if } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \end{aligned} \tag{14}$$

where $X_k^2 \in [0, Y_k^2]$ and $\lambda(k)$ is defined as

$$\frac{1}{\lambda(k)} \equiv \frac{1}{\sigma_x^2(k)} - \frac{1}{\sigma_d^2(k)}, \quad \text{if } \sigma_x^2(k) \neq \sigma_d^2(k), \tag{15}$$

and

$$\Psi_k \equiv \frac{1}{\lambda(k)\left\{1 - \exp\left[-\frac{Y_k^2}{\lambda(k)}\right]\right\}}. \tag{16}$$

Note that if $\sigma_x^2(k) > \sigma_d^2(k)$, then $1/(\lambda(k)) < 0$, and vice versa. Thus, Ψ_k in (14) is always positive.

B. *Minimum Mean Square Error Estimator*

Using (11)–(14), we can derive two different estimators of the magnitude-squared spectrum. The MMSE estimator is obtained by computing the mean of the posteriori density given in (14)

$$\begin{aligned} \hat{X}_k^2 &= E\left\{X_k^2 \,|\, Y_k^2\right\} \\ &= \int_0^{Y_k^2} X_k^2 f_{X_k^2}\left(X_k^2 \,|\, Y_k^2\right) dX_k^2 \\ &= \begin{cases} \left(\frac{1}{\nu_k} - \frac{1}{e^{\nu_k}-1}\right) Y_k^2 & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{2} Y_k^2 & \text{if } \sigma_x^2(k) = \sigma_d^2(k) \end{cases} \end{aligned} \tag{17}$$

where ν is defined as

$$\nu_k \equiv \frac{1 - \xi_k}{\xi_k} \gamma_k. \tag{18}$$

Note that the above MMSE estimator is derived by computing the mean of the *posteriori* density conditioned on the noise-corrupt magnitude-squared spectrum (Y_k^2), rather than the complex noisy spectrum $(Y(\omega_k))$. This differentiates the present MMSE estimator from that derived in (3) [6], [7].

The gain function of the above MMSE estimator is given by

$$G_{\mathrm{MMSE}}(\xi_k, \gamma_k) = \begin{cases} \left(\frac{1}{\nu_k} - \frac{1}{e^{\nu_k}-1}\right)^{\frac{1}{2}} & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \left(\frac{1}{2}\right)^{\frac{1}{2}} & \text{if } \sigma_x^2(k) = \sigma_d^2(k). \end{cases} \tag{19}$$

We will henceforth refer to the above estimator as the MMSE-SPZC estimator, where SPZC stands for Spectrum Power estimator based on Zero Cross-terms assumptions. Note that much like the gain function of MMSE-SP estimator (3), the above gain function depends on two parameters, ξ_k and γ_k. Figs. 1 and 2 show the gain function of the MMSE-SPZC estimator for fixed values of ξ and fixed values of γ, respectively. As can be seen from these two figures, the MMSE-SPZC estimator provides more suppression than the MMSE-SP estimator for small values of ξ ($\xi < 0$ dB) and large values of γ ($\gamma > 10$ dB). We thus expect the MMSE-SPZC estimator to reduce the residual noise commonly encountered in speech processed by the MMSE-SP estimator. It is interesting to note, that when $\xi = 0$ dB, the MMSE-SPZC estimator provides constant attenuation of -3 dB, independent of the value of γ. This is shown analytically in (17) and in Appendix A.

Note that Ding *et al.* [5] proposed this MMSE estimator incorporating a mixture of Gaussians for modeling the clean speech

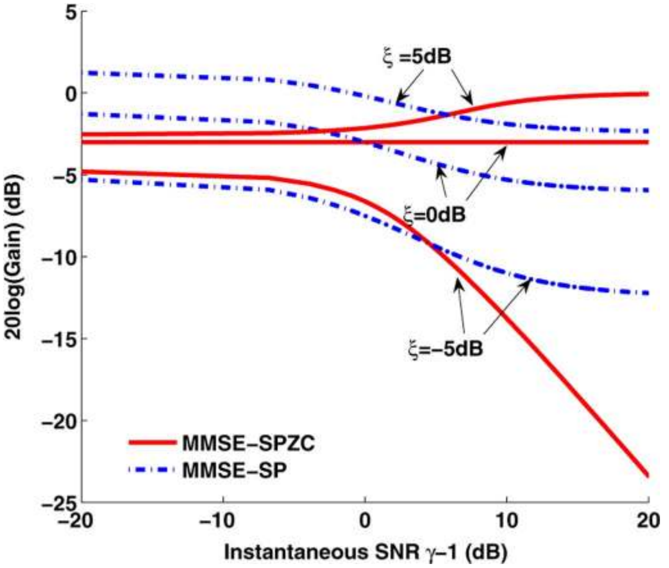

Fig. 1. Gain function of the proposed MMSE-SPZC estimator of the power spectrum plotted as a function of the instantaneous SNR $(\gamma_k - 1)$ for fixed values of ξ_k. The gain function of the MMSE-SP estimator [7] is superimposed for comparison.

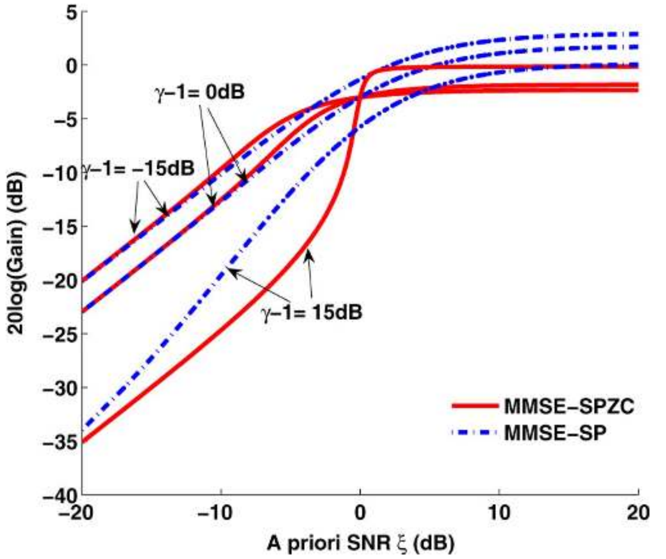

Fig. 2. Gain function of the proposed MMSE-SPZC estimator of the power spectrum plotted as a function of the *a priori* SNR (ξ_k) for fixed values of γ_k. The gain function of the MMSE-SP estimator [7] is superimposed for comparison.

variance. A mixture model, trained using data from a large database, was used for online estimation for the clean speech from the corrupted speech. Unlike [5], a single Gaussian was used in the present study for modeling the density of the real and imaginary parts of the DFT coefficients.

C. Maximum a Posterior (MAP) Estimator

The *a posterior* probability density (14) function is monotonic, and when ξ (expressed in dB) changes its sign, the density changes its direction (increasing versus decreasing). This simplifies the maximization a great deal. The MAP estimator is given as follows:

$$\begin{aligned}\hat{X}_k^2 &= \arg\max_{X_k^2} f_{X_k^2}\left(X_k^2 \mid Y_k^2\right) \\ &= \begin{cases} Y_k^2, & \text{if } \frac{1}{\lambda(k)} < 0 \\ 0, & \text{if } \frac{1}{\lambda(k)} > 0 \end{cases} \\ &= \begin{cases} Y_k^2, & \text{if } \sigma_x^2(k) \geq \sigma_d^2(k) \\ 0, & \text{if } \sigma_x^2(k) < \sigma_d^2(k). \end{cases} \end{aligned} \qquad (20)$$

Note that X_k^2 is limited in $[0, Y_k^2]$ due to (11). Based on (14), when $\sigma_x^2(k) = \sigma_d^2(k)$, the conditional density is uniformly $(1/Y_k^2)$ in the range of $[0, Y_k^2]$, and therefore the MAP estimate in this special case could be any value in the range of $[0, Y_k^2]$. In our case, we chose to use the noisy observation as in (20). The gain function of the MAP estimator is given by

$$G_{\mathrm{MAP}}(k) = \begin{cases} 1, & \text{if } \sigma_x^2(k) \geq \sigma_d^2(k) \\ 0, & \text{if } \sigma_x^2(k) < \sigma_d^2(k). \end{cases} \qquad (21)$$

Using (4), the above gain function can also be written as

$$G_{\mathrm{MAP}}(\xi_k) = \begin{cases} 1, & \text{if } \xi_k \geq 1 \\ 0, & \text{if } \xi_k < 1. \end{cases} \qquad (22)$$

Note that unlike the MMSE gain function (19), the MAP gain function is binary valued. In fact, it is nearly the same as the ideal binary mask widely used in CASA [22], [23]. In CASA, the binary mask assigns a binary weight for each time–frequency unit based on the value of the local, instantaneous, SNR. If the local SNR is greater than a pre-defined threshold (e.g., 0 dB), the binary mask takes the value of 1, and if it is less than the threshold, the binary mask takes the value of 0. Speech is synthesized by multiplying the binary mask with the noisy signal, and large gains in intelligibility were reported in [23], [24] with speech synthesized by the ideal binary mask. The gain function implicitly used in the ideal binary mask technique is nearly identical to that given by (22). The main difference between the ideal binary mask and the MAP gain function (22) is that the latter is based on the *a priori* SNR, whereas the ideal binary mask is based on the instantaneous SNR.

It is also interesting to note that this MAP estimator follows a so-called "hard-thresholding" rule often used in the wavelet shrinkage literature [32], [27], [28]. The hard-thresholding rule belongs to the class of diagonal linear projection estimators. These estimators [32] share the same rule as given in (22) in that they keep the observation when the signal is larger than the noise level, and "kill" the observation otherwise. According to [32] the ideal risk for our estimation problem at hand can be computed as $R(\hat{X}, X) = \sum_k \min(\sigma_{x_k}^4, \sigma_{d_k}^4)$. There are, however, a number of differences between the diagonal estimators used in the wavelet literature and the above MAP estimator. For one, the diagonal estimators operate on the wavelet coefficients, which possess a different distribution than the Fourier coefficients used in the present study. The wavelet transform produces a sparse signal and noise is typically spread out equally over all coefficients [29]. Second, most of the oracle risk bounds that were computed for different thresholding rules are not applicable here, as those bounds were derived under the assumption that the additive noise was Gaussian [33], [34]. In our case, the magnitude-squared spectrum of the noise in our model in (11) is assumed to have an exponential distribution, i.e., our additive noise model in (11) is based on an exponential distribution assumption and not a Gaussian assumption. In brief, while the proposed MAP estimator is similar to the hard-thresholding rule used in the wavelet shrinkage literature, the underlying assumptions and criteria are totally different.

As mentioned earlier, a number of MAP estimators of the *magnitude spectrum* have been proposed in the literature [35], [12]–[14], [11], [7] for speech enhancement, and these are summarized in Table I. There are however a number of distinct differences between the derived MAP estimator and the previous MAP estimators. For one, no MAP estimators of the magnitude-squared spectrum have been reported previously. Second, the *posteriori* density used in prior studies (except [14]) is different as it is conditioned on the complex spectrum of the noisy signal, rather than the magnitude-squared spectrum of the noisy signal (see Table I). As shown in (6), the *posteriori* density involved in the derivation of previous MAP estimators contains a Bessel function $(I_o(x))$, making it difficult to derive a closed form solution for the MAP estimator. In fact, a closed form solution was found in previous MAP estimators [11], [7], [12]–[14] only after approximating the Bessel function with a function of the form $\exp(x)/\sqrt{2\pi x}$. While this approximation is valid for large values of x, it becomes erroneous for small values of x. In contrast, the derived *posteriori* density [see (14)] in the present study has a much simpler form enabling us to derive a closed form solution without resorting to any approximations. Furthermore, based on the fact that $X_k^2 \leq Y_k^2$ [owing to (11)], the integration is simplified a great deal, as shown for instance in (17). In [14], the authors opted to approximate the Laplacian and Gamma distributions with parametric density functions. In brief, we derived in the present study a MAP estimator of the magnitude-squared spectrum, rather than a MAP estimator of the magnitude spectrum (already reported previously—see Table I), and this MAP estimator was derived in closed-form without making any approximations. Finally, and perhaps more importantly, we demonstrated that there exists a link between the proposed MAP estimator and the ideal binary mask used in CASA applications.

IV. Incorporating SNR Uncertainty and Proposed Soft Masks

We showed in the last section that the MAP estimator is similar to the binary mask technique used in CASA [22]. The ideal binary mask (IdBM) is often used as the computational goal in CASA [25], [22]. Use of IdBM has been shown to restore speech

TABLE I
MAP ESTIMATOR COMPARISONS

Method	Posterior Density	MAP Estimators
MAP spectral amplitude estimator [11][7]	$f(X\|Y(\omega)) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{X}{s}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\frac{X-s}{\sigma}\right]^2\right\}$	$\hat{X} = \arg\max_X f(X\|Y(\omega))$ $= \frac{\xi + \sqrt{\xi^2 + (1+\xi)(\xi/\gamma)}}{2(1+\xi)} Y$
Joint MAP spectral amplitude and phase estimator [11][7]	$f(X, \theta_x\|Y(\omega))$ $\propto \frac{X}{\pi^2 \sigma_x^2 \sigma_d^2} \exp\left(-\frac{\left\|Y(\omega) - Xe^{j\theta_x}\right\|^2}{\sigma_d^2} - \frac{X^2}{\sigma_x^2}\right)$	$\hat{X} = \arg\max_X f(X, \theta_x\|Y(\omega))$ $= \frac{\xi + \sqrt{\xi^2 + 2(1+\xi)(\xi/\gamma)}}{2(1+\xi)} Y$
MAP spectral amplitude estimator using super-Gaussian assumption [12][14]	$f(X\|Y) \propto X^{\kappa-1/2} \exp\left\{-\frac{X^2}{\sigma_d^2} - X\left(\frac{\mu}{\sigma_x} - \frac{2Y}{\sigma_d^2}\right)\right\}$ κ, μ are parameters for super-Gaussian approximation	$\hat{X} = \arg\max_X f(X\|Y)$ $= \left(u + \sqrt{u^2 + \frac{\kappa - 1/2}{2\gamma}}\right) Y,$ $u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}$
Joint MAP spectral amplitude and phase estimator using super-Gaussian assumption [13][14]	$\ln\left(f(X, \theta_x\|Y(\omega))\right)$ $\propto -\frac{\left\|Y(\omega) - Xe^{j\theta_x}\right\|^2}{\sigma_d^2} + \kappa \ln X - \mu \frac{X}{\sigma_x}$ κ, μ are parameters for super-Gaussian approximation	$\hat{X} = \arg\max_X f(X, \theta_x\|Y(\omega))$ $= \left(u + \sqrt{u^2 + \frac{\kappa}{2\gamma}}\right) Y,$ $u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma\xi}}$
Proposed MAP power spectrum estimator (22)	$f(X^2\|Y^2) = \begin{cases} \Psi e^{-\frac{X^2}{\lambda}} & \text{if } \sigma_x^2 \neq \sigma_d^2 \\ \frac{1}{Y^2} & \text{if } \sigma_x^2 = \sigma_d^2, \end{cases}$	$\hat{X}^2 = \arg\max_{X^2} f(X^2\|Y^2)$ $= \begin{cases} Y^2 & \text{if } \sigma_x^2 \geq \sigma_d^2 \\ 0 & \text{if } \sigma_x^2 < \sigma_d^2. \end{cases}$

The frequency subscript k is removed for convenience in this table. Note that $Y(\omega)$ is complex valued, while X and Y indicate the magnitude spectra (real-valued).

intelligibility even when speech is corrupted at extremely low SNR levels [23], [24], [36]. However, implementation of IdBM requires access to the true local (instantaneous) SNR rather than the *a priori* SNR. Estimation of the local SNR is difficult as it requires knowledge of the speech and noise magnitude-squared spectra, which we do not have. Furthermore, applying a binary gain to noisy speech spectra, could affect the quality of speech in that frequent zeroing of spectral components (when the local SNR < 1) could potentially produce musical noise. This is so because the zeroing of spectral components can create small, isolated peaks in the spectrum occurring at random frequency locations in each frame. Converted to the time domain, these peaks sound similar to tones with frequencies that change randomly from frame to frame, and produce musical noise. In brief, there exists an uncertainty in estimating the local and *a priori* SNR accurately and reliably at all SNR levels.

In this section, we propose soft masking methods which incorporate local SNR uncertainty, thereby making the gain function continuous (soft) rather than binary. Henceforth, we refer to these estimators as soft masking estimators. Methods for estimating reliably binary gain functions, as required for the IdBM technique, have been reported in [36] and [37].

In the rest of this section, we propose two soft masking methods that incorporate *a priori* and *a posteriori* SNR uncertainty, respectively.

A. Soft Mask Formulation

The variances of the speech and noise spectra are the key parameters in most statistical models. As neither speech or noise are stationary, their variances are time-varying. However, in short-time intervals (10–30 ms), the speech and noise signals can be assumed to be quasi-stationary processes. Their variances can be modeled as unknown but deterministic parameters. Thus, the *a priori* SNR ξ_k can also be assumed to be unknown but deterministic.[1] Given the *a priori* SNR, the probability density of the local (instantaneous) SNR can be computed. More precisely, after defining the instantaneous SNR, ξ_I, as follows:

$$\xi_{I,k} \equiv \frac{X_k^2}{D_k^2} \tag{23}$$

we express the ideal binary mask (IdBM) rule as

$$\hat{X}_k^2 = G_k \cdot Y_k^2 = \begin{cases} Y_k^2, & \text{if } \xi_{I,k} \geq 1, \text{ where } G_k = 1 \\ 0, & \text{if } \xi_{I,k} < 1, \text{ where } G_k = 0. \end{cases} \tag{24}$$

Following the approach in [40], we formulate the binary mask problem using the following binary hypothesis model:

$$\begin{aligned} &H_0 : \xi_{I,k} < \theta : \text{masker dominates} \\ &H_1 : \xi_{I,k} \geq \theta : \text{target signal dominates.} \end{aligned} \tag{25}$$

The gain function G in (24) can be considered to be a random variable as it depends on the instantaneous SNR, ξ_I. In the context of binary masking, G is a Bernoulli distributed random variable taking the value of 0 or 1, and its parameter p is the hypothesis probability $P(H_1)$. It is difficult to estimate G as it depends

[1] The noise variance is typically estimated using noise PSD estimation methods, such as the minimum statistics [38], and minimum controlled recursive average [39] algorithms. The *a priori* SNR is usually estimated by the "decision-directed" [2] method.

on accurate estimates of the instantaneous SNR. However, we can obtain G more reliably by taking its expectation. In doing so, we obtain the following weighted average estimate of the magnitude-square spectrum now incorporating the aforementioned two hypothesis:

$$\begin{aligned}\hat{X}_k^2 &= \mathbb{E}\{G_k\} \cdot Y_k^2 \\ &= [E[G_k \mid H_1] \cdot P(H_1) + E[G_k \mid H_0] \cdot P(H_0)] \cdot Y_k^2\end{aligned} \tag{26}$$

where $P(H_1)$ denotes the probability that hypothesis H_1 is true, $E[G_k \mid H_1]$ denotes the gain function assuming that hypothesis H_1 is true (i.e., target signal dominates) and $E[G_k \mid H_0]$ denotes the gain function assuming that hypothesis H_0 is true (i.e., masker dominates). From (24), $E[G_k \mid H_1] = 1$ and $E[G_k \mid H_0] = 0$. In practice, using a very small value for $E[G_k \mid H_0]$ results in better quality and with enhanced speech containing small amounts of residual noise. In our study, we used the value of $G_f = -20$ dB for $E[G_k \mid H_0]$ to minimize the residual noise. In the next two subsections, we derive the probability terms $P(H_1)$ and $P(H_0)$.

B. Soft Masking by Incorporating a Priori SNR Uncertainty

Assuming independence between the clean speech and noise magnitude-squared spectra, we can easily use (12) and (13) to model the hypothesis probability given the *a priori* SNR ξ. As we do not use any other constraint or assumption, we refer to this hypothesis probability as the *a priori* SNR uncertainty.

Using the exponential models for X_k^2 and D_k^2 [i.e., (12) and (13)] it is easy to derive (see Appendix B) the probability density of ξ_I as

$$f_{\xi_I}(\xi_I) = \frac{\xi}{(\xi + \xi_I)^2} u(\xi_I) \tag{27}$$

where $u(\cdot)$ is the step function. For an arbitrary SNR threshold θ, the hypothesis probability needed in (26) is computed as

$$P(H_1) = P(\xi_I > \theta) = \int_\theta^\infty f_{\xi_I}(z) dz = \frac{\xi}{\xi + \theta}. \tag{28}$$

Note that the above probability can only be assessed when the *a priori* SNR ξ is given. We refer to this probability as *priori* since it does not require information from the noise-corrupt observations and does not need the assumption of (11). As mentioned before, ξ can be estimated using the "decision-directed" approach in conjunction with noise PSD estimation algorithms.

Finally, by inserting (28) into (26), we get

$$\hat{X}_k^2 = \frac{\xi_k}{\xi_k + \theta} \cdot Y_k^2 \tag{29}$$

where ξ_k is the *a priori* SNR (4). It is interesting to note that when $\theta = 1$, the above estimator becomes identical to the Wiener filter. We will be referring to the above estimator as the soft mask estimator with *a priori* SNR uncertainty, and we denote it as SMPR.

Fig. 3 plots the gain function of the SMPR estimator for three different thresholds, $\theta = -5$, 0, and 5 dB. The gain function of the Wiener filter is superimposed for comparative purposes. As discussed, the Wiener gain is identical to the SMPR gain for $\theta = 0$ dB. For thresholds $\theta > 0$ dB, the SMPR gain function becomes steep and more aggressive, while for thresholds $\theta < 0$ dB, the SMPR gain function becomes shallow and less aggressive.

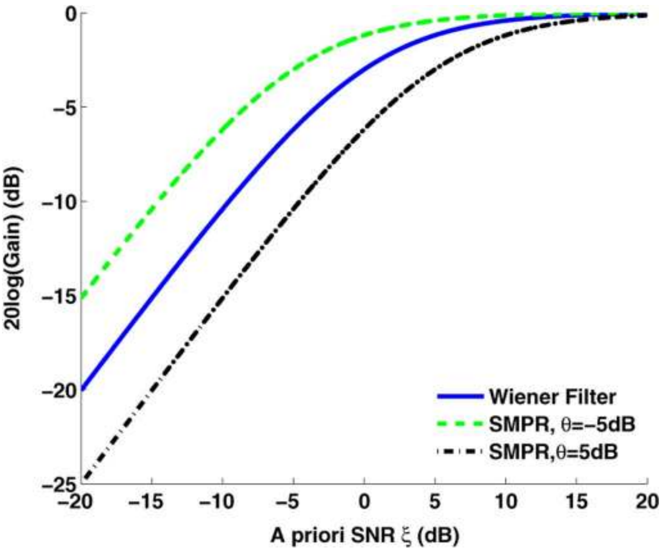

Fig. 3. Gain function of the SMPR estimator plotted as a function of the *a priori* SNR ξ_k and for different values of threshold θ. The Wiener gain function is superimposed for comparison.

There exists a large body of literature in wavelet denoising in terms of choosing the right threshold, and includes among others adaptive selection procedures such as the SURE [28] and cross-validation methods. These threshold selection techniques, however, are based on the Gaussian additive model assumption, which as discussed previously (see Section III-C) is not applicable to our study. Our choice of thresholds was based largely on perceptual studies. The study in [23], for instance, indicated that SNR threshold values in the range of $[-12, 5]$ dB produced large improvements in intelligibility. This range of SNR threshold values will be examined in the present study.

C. Soft Masking Based on Posteriori SNR Uncertainty

Clearly the above SMPR estimator did not incorporate information about the noisy observations, as it relied solely on *a priori* information about the instantaneous SNR ξ_I. It is reasonable to expect that a better estimator could be developed by incorporating *posteriori* information about the SNR at each frequency bin. In this case, we incorporate the assumption given in (11) to compute the hypothesis probability, which is referred to as *a posteriori* SNR uncertainty.

This hypothesis probability can be computed as the posteriori probability of $\xi_{I,k} \geq \theta$ as follows:

$$\begin{aligned}P(H_1) &= P\left(\xi_{I,k} > \theta \mid Y_k^2\right) \\ &= P\left(X_k^2 > \frac{\theta}{\theta+1} Y_k^2 \mid Y_k^2\right) \\ &= \int_{\frac{\theta}{\theta+1} Y_k^2}^{Y_k^2} f_{X_k^2}\left(z \mid Y_k^2\right) dz.\end{aligned} \tag{30}$$

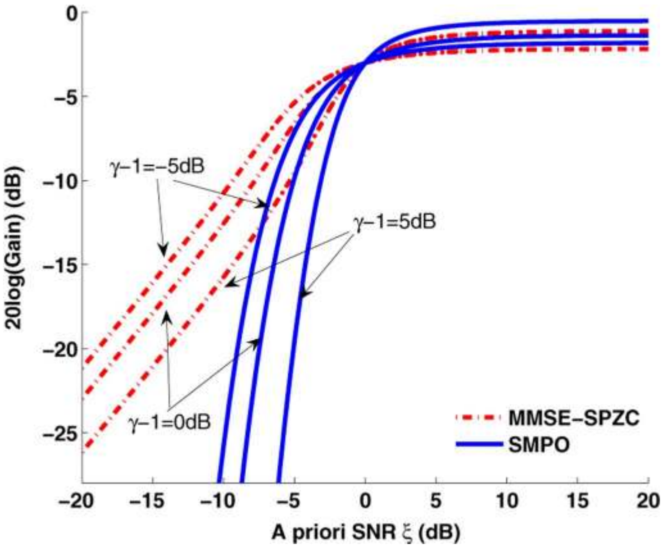

Fig. 4. Gain function of the SMPO estimator plotted as a function of the *a priori* SNR ξ_k and for different values of γ_k. The threshold θ was fixed at $\theta = 0$ dB. The gain function of the MMSE-SPZC estimator is superimposed for comparison.

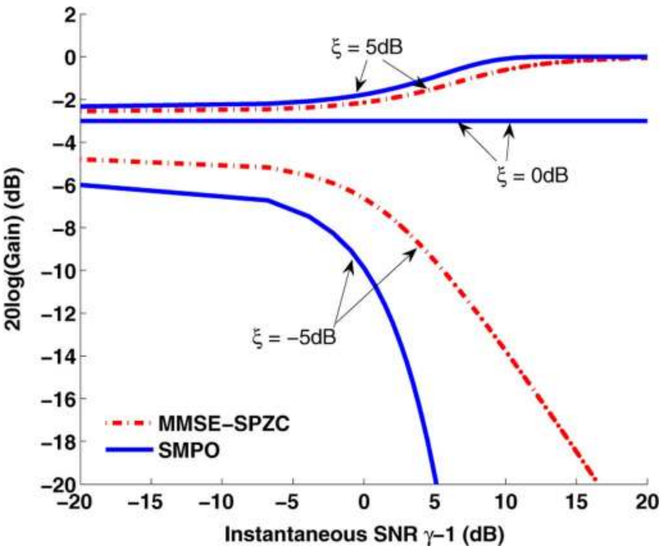

Fig. 5. Gain function of the SMPO estimator plotted as a function of the instantaneous SNR $(\gamma_k - 1)$ and for different values of ξ_k. The threshold θ was fixed at $\theta = 0$ dB, while the floor gain G_f was set to -20 dB. The gain function of the MMSE-SPZC estimator is superimposed for comparison.

Inserting (14) into (30), we get

$$P\left(\xi_{I,k} > \theta \,|\, Y_k^2\right) = \begin{cases} \frac{e^{\frac{\nu_k}{\theta+1}}-1}{e^{\nu_k}-1}, & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{\theta+1}, & \text{if } \sigma_x^2(k) = \sigma_d^2(k). \end{cases} \tag{31}$$

Finally, substituting (31) into (26), we obtain the following estimator:

$$\hat{X}_k^2 = \begin{cases} \frac{e^{\frac{\nu_k}{\theta+1}}-1}{e^{\nu_k}-1} \cdot Y_k^2, & \text{if } \sigma_x^2(k) \neq \sigma_d^2(k) \\ \frac{1}{\theta+1} \cdot Y_k^2, & \text{if } \sigma_x^2(k) = \sigma_d^2(k). \end{cases} \tag{32}$$

We will be referring to the above estimator as the soft mask estimator with *posteriori* SNR uncertainty, and will be denoted as SMPO.

The SMPO gain function (32) is dependent on both the ξ and the γ values. Figs. 4 and 5 plot the gain functions of SMPO as a function of ξ (for fixed values of γ) and as function of γ (for fixed values of ξ), respectively. For these plots the SNR threshold was fixed at $\theta = 0$ dB. The gain function of the MMSE-SPZC estimator (19) is plotted for comparison. As can be seen from both figures, the gain function of the SMPO estimator is more aggressive (i.e., provides more attenuation) than the MMSE-SPZC for low values of ξ ($\xi < -5$ dB). Fig. 6 plots the gain function of the SMPO estimator for different values of θ (with γ fixed at 0 dB). Overall, the gain functions are steep, resembling to some degree binary functions (at least for the value of γ chosen), with small values of θ ($\theta < 0$ dB) shifting the curve to the left and large values of θ ($\theta > 0$ dB) to the right, as expected. Unlike the binary gain function of the MAP estimator (22) which depends solely on the value of ξ, the gain function of the SMPO estimator depends on information collected from both the ξ and γ parameters. As shown in Fig. 4, the γ parameter can shift the gain function to the right (for large values of γ) and to the left (for smaller values of γ). For that reason, we expect the SMPO estimator to be more robust than the MAP estimator (22) to inaccuracies in the estimate of ξ.

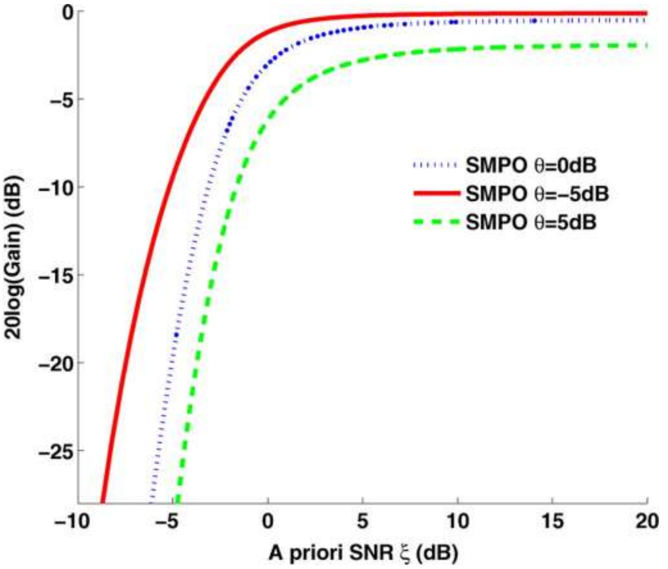

Fig. 6. Gain function of the SMPO estimator plotted as a function of the *a priori* SNR ξ_k ($\gamma = 5$ dB) and for different values of threshold θ.

V. Implementation

Estimates of the *a priori* SNR ξ are needed in the implementation of the MMSE-SPZC, SMPO and SMPR estimators. For that, we used the "decision-directed" [2] approach:

$$\xi_k(l) = \max\left\{ \alpha \frac{\hat{X}_k^2(l-1)}{\hat{\sigma}_d^2(k,l-1)} + (1-\alpha)\max\left[\gamma_k(l)-1, 0\right], \xi_{\min} \right\} \tag{33}$$

where $\xi_{\min} = -20$ dB, l denotes the frame index and $\hat{\sigma}_d^2(k,l)$ denotes the estimate of the noise variance.

The MAP estimator can be implemented by either using (21) or (24). Both implementations were considered. In order to estimate the instantaneous SNR $\hat{\xi}_{I,k}$ needed in (24), we used the

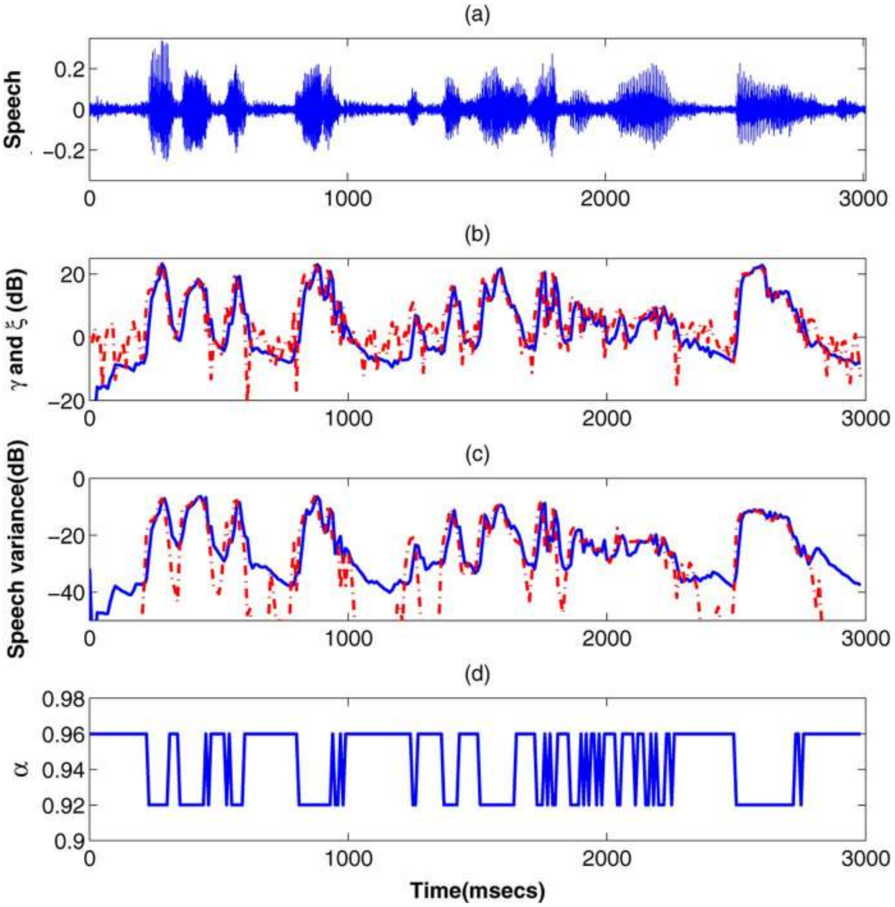

Fig. 7. Panel (d) shows example estimates of the smoothing constant α_k (at bin $f = 500$ Hz) used in the computation of the signal variance (34). Panel (a) shows the time waveform for a sentence corrupted by babble noise at 10 dB SNR. Panel (b) shows the *a priori* SNR ξ (solid) and the *a posteriori* SNR γ (dash-dotted) values. Panel (c) shows the estimated speech variance (solid), based on (34) and (37), and the true speech variance (dash-dotted).

MMSE estimator [2] to obtain the spectral amplitude estimate $\tilde{X}_k$ of the clean speech and thereafter computed the instantaneous SNR as $\hat{\xi}_{I,k} = (\tilde{X}_k^2)/(\hat{\sigma}_d^2(k))$. This method was noted as MAP-BM.

For the implementation of the MAP estimator given in (21), a method was needed to compute the signal variance (σ_x^2). More precisely, the following method was adopted for estimating the signal variance

$$\hat{\sigma}_x^2(k,l) = \alpha_k \bar{\sigma}_x^2(k,l-1) + (1-\alpha_k)\tilde{\sigma}_x^2(k,l) \tag{34}$$

where a_k is a smoothing constant (computed adaptively) and $\tilde{\sigma}_x^2(k,l)$ is estimated from the current frame as follows:

$$\tilde{\sigma}_x^2(k,l) = \hat{\sigma}_y^2(k,l) - \sigma_d^2(k,l) \tag{35}$$

and $\hat{\sigma}_y^2$ is computed using first-order recursive smoothing

$$\hat{\sigma}_y^2(k,l) = \eta\hat{\sigma}_y^2(k,l-1) + (1-\eta)Y_k^2(l) \tag{36}$$

where η is a smoothing constant. The signal variance $\bar{\sigma}_x^2(k,l-1)$ was computed using (3) as follows:

$$\bar{\sigma}_x^2(k,l-1) = \frac{\xi_k(l-1)}{\xi_k(l-1)+1} \times \left[\frac{1}{\gamma_k(l-1)} + \frac{\xi_k(l-1)}{\xi_k(l-1)+1}\right] Y_k^2(l-1). \tag{37}$$

A simple adaptive method was used to adjust the smoothing constant α_k in (34). The motivation behind the adaptive rule described below is to use a small value of α when γ is large, and a comparatively larger value when γ is small:

$$\alpha_k = \begin{cases} \alpha_1, & \text{if } \gamma_k < \zeta_k \\ \alpha_0, & \text{if } \gamma_k > \zeta_k, \end{cases} \tag{38}$$

where $\alpha_0 < \alpha_1 < 1$, and ζ_k are adaptive thresholds determined similarly by

$$\zeta_k(l) = \begin{cases} \zeta_0, & \text{if } \xi_k(l-1) < \delta \\ \zeta_1, & \text{if } \xi_k(l-1) > \delta \end{cases} \tag{39}$$

where ζ_0, ζ_1, and δ are constants. Fig. 7 shows example estimates of α for a sentence corrupted by babble at 10 dB SNR. The signal variance estimate is also shown in panel (c) based on (34) and (37). As can be seen, when γ is small, the value of α is large $(\alpha_1 = 0.96)$, suggesting that more emphasis should be placed on the previous frame's variance estimate. Hence, for the most part, low-energy segments use α_1, while high-energy segments use α_0.

In our study we adopted the following constants: $\eta = 0.65$ (36), $\delta = 0.2$, $\zeta_0 = 14$, $\zeta_1 = 5$, $\alpha_1 = 0.96$, and $\alpha_0 = 0.92$. Different values of α were used in (33) for different estimators. For the MMSE-SP estimator it was set to $\alpha = 0.98$, for the MMSE-SPZC estimator it was set to $\alpha = 0.97$, and for the SMPR and SMPO estimators it was set to $\alpha = 0.90$. These

TABLE II
PERFORMANCE, IN TERMS OF MSE, OF THE SMPR AND SMPO ESTIMATORS AS A FUNCTION OF THRESHOLD θ

Noise	Method	THR	15dB	10dB	5dB	0dB	Noise	Method	THR	15dB	10dB	5dB	0dB
Car	SMPR	-5dB	0.218	0.631	2.156	9.187	Street	SMPR	-5dB	0.192	0.682	2.795	19.436
		0dB	**0.216**	**0.587**	**1.673**	**5.245**			0dB	**0.186**	0.593	**2.146**	10.355
		5dB	0.232	0.636	1.725	5.196			5dB	0.194	**0.581**	2.149	**5.623**
	SMPO	-5dB	**0.221**	**0.629**	2.028	7.186		SMPO	-5dB	**0.195**	0.685	2.709	16.399
		0dB	0.228	0.631	**1.853**	5.624			0dB	0.196	0.622	**2.328**	7.545
		5dB	0.239	0.664	1.911	**5.956**			5dB	0.203	**0.613**	2.368	**5.612**
Babble	SMPR	-5dB	0.304	0.947	4.184	19.959	White	SMPR	-5dB	**0.085**	0.229	0.742	2.742
		0dB	**0.289**	**0.814**	3.026	10.778			0dB	0.086	**0.223**	**0.668**	**2.033**
		5dB	0.298	0.824	**2.818**	**7.646**			5dB	0.092	0.244	0.737	2.227
	SMPO	-5dB	0.310	0.963	4.009	17.077		SMPO	-5dB	**0.086**	**0.232**	**0.730**	2.411
		0dB	**0.305**	**0.881**	3.070	8.691			0dB	0.088	0.238	0.735	**2.269**
		5dB	0.312	0.894	**2.923**	**8.485**			5dB	0.091	0.250	0.774	2.398

The numbers in boldface indicate the best performance.

values were optimized for each estimator based on their resulting PESQ [41] score.[2] This ensured best performance from each estimator.

For the soft masking methods incorporating SNR uncertainty, i.e., SMPR (29) and SMPO (32), the $E[G_k \mid H_0]$ term was set to $G_f = -20$ dB in order to retain small amounts of residual noise and make the quality of the enhanced speech more natural.

Speech was segmented into 20 ms frames and Han-windowed with 50% overlap. The short-time Fourier transform was applied to each frame to obtain the noisy magnitude spectrum Y_k. The gain functions G_k of the derived estimators (Sections III and IV) were applied to the noisy magnitude spectrum to get the enhanced signal spectrum $\hat{X}_k$ as $\hat{X}_k = G_k Y_k$. An inverse Fourier transform was taken of $\hat{X}_k$ using the noisy speech phase spectrum to reconstruct the time-domain signal. The overlap-add method was used to obtain the enhanced signal.

VI. EXPERIMENTS

A total of 30 sentences taken from the NOIZEUS [4] database was used to evaluate the performance of the proposed estimators. The sentences were corrupted by car, street, babble and white noise at 0, 5, 10, and 15 dB. Two measures were used to assess performance, the mean-square error (MSE) between the estimated (short-time) and the true magnitude-squared spectrum, and the Perceptual Evaluation of Speech Quality (PESQ) [41] measure. The MSE measure is defined as

$$\mathrm{MSE} = \frac{1}{MN} \sum_{l=0}^{M-1} \sum_{k=0}^{N-1} \left[X_k^2(l) - \hat{X}_k^2(l) \right]^2 \tag{40}$$

where X_k^2 is the short-time magnitude-squared spectrum of the clean signal, $\hat{X}_k^2$ is the estimated magnitude-squared spectrum, N is the total number of frequency bins, and M is the total number of the frames in a sentence. While small values of MSE imply a better estimate of the true magnitude-squared spectrum, they do not imply better speech quality. For that reason, we used the PESQ [41] measure which has been found to correlate highly [42] with speech quality. Unlike the MSE, higher PESQ values indicate better performance, i.e., better speech quality.

A. Influence of Threshold Value on Performance

In the first set of experiments, we wanted to examine the influence of the selected thresholds in the performance of the SMPO and SMPR estimators. The thresholds were varied from -5 dB to 5 dB, and performance (in terms of MSE and PESQ scores) was assessed. Table II shows the MSE results and Table III shows the PESQ results. In terms of PESQ scores, better performance is obtained with the SMPR estimator when $\theta = 5$ dB. This was found to be consistent for all types of noise examined. For the SMPO estimator, good performance (in terms of PESQ scores) was obtained with $\theta = 0$ dB. The MSE values were consistently low for $\theta = 0$ dB. For that reason, we fixed the threshold to $\theta = 0$ dB for the SMPO estimator and to $\theta = 5$ dB for the SMPR estimator in subsequent experiments.

B. Evaluation of Proposed Estimators

In the second set of experiments, we first compared the performance of the magnitude-squared spectrum estimators derived in the present study against that proposed by [7] [see (3)]. The latter estimator (3) derived in [7], [6] is denoted as MMSE-SP. In addition, for benchmark purposes we report the performance of the (oracle) ideal binary mask and ideal ratio masks [25], which assume access to the true instantaneous SNR of each bin. These oracle estimators are included as they provide the upper bound in performance of the MAP estimators. The ideal binary mask (noted as IdBM) adopts the rule of (24), while the ideal ratio mask (noted as IdRM) is computed using the following gain function [43]:

$$G_{\mathrm{IdRM}}(k) = \frac{X_k^2}{X_k^2 + D_k^2}. \tag{41}$$

For further evaluation of the MMSE-SPZC (17) estimator, and following [40] and [44], we incorporated the SNR uncertainty in the estimator. In Section IV, we derived the probability of the local SNR exceeding a threshold. We assume that when the local SNR is below -20 dB, speech is absent. The hypothesis is given as follows:

$$\begin{aligned} H_0' &: \xi_{I,k} < -20 \text{ dB} : \text{Speech absent} \\ H_1' &: \xi_{I,k} \geq -20 \text{ dB} : \text{Speech present.} \end{aligned} \tag{42}$$

Therefore, the probabilities of $P(H_1')$ can be computed by (30), by setting the threshold $\theta = -20$ dB.

[2] Thirty sentences in 10 dB babble noise were used to optimize the selection of α for each estimator. Consistent results were obtained in other types of noise.

TABLE III
PERFORMANCE, IN TERMS OF PESQ SCORES, OF THE SMPR AND SMPO ESTIMATORS AS A FUNCTION OF THRESHOLD θ

Noise	Method	THR	15dB	10dB	5dB	0dB	Noise	Method	THR	15dB	10dB	5dB	0dB
Car	SMPR	-5dB	2.69	2.36	2.03	1.77	Street	SMPR	-5dB	2.70	2.40	2.06	1.74
		0dB	2.91	2.58	2.24	1.93			0dB	2.91	2.62	2.26	1.92
		5dB	**3.12**	**2.77**	**2.40**	**2.05**			5dB	**3.13**	**2.82**	**2.43**	**2.07**
	SMPO	-5dB	2.80	2.49	2.17	1.87		SMPO	-5dB	2.79	2.52	2.18	1.86
		0dB	**3.17**	**2.85**	**2.50**	**2.17**			0dB	3.15	**2.89**	**2.53**	**2.18**
		5dB	3.13	2.79	2.42	2.05			5dB	**3.18**	2.87	2.47	2.12
Babble	SMPR	-5dB	2.78	2.44	2.11	1.81	White	SMPR	-5dB	2.64	2.32	1.99	1.70
		0dB	2.95	2.60	2.25	1.93			0dB	2.91	2.59	2.24	1.91
		5dB	**3.16**	**2.78**	**2.42**	**2.07**			5dB	**3.14**	**2.79**	**2.42**	**2.04**
	SMPO	-5dB	2.85	2.51	2.18	1.86		SMPO	-5dB	2.79	2.48	2.16	1.86
		0dB	3.17	2.82	2.48	**2.14**			0dB	**3.21**	**2.90**	**2.57**	**2.21**
		5dB	**3.24**	**2.84**	**2.48**	2.12			5dB	3.18	2.82	2.46	2.07

The numbers in boldface indicate the best performance.

The MMSE-SPZC estimator incorporating *a priori* SNR uncertainty is denoted as "MMSE-SPZC-U" and is implemented as follows:

$$\hat{X}_k = G_{k,\mathrm{MMSE-SPZC}} \cdot P(H_1') \cdot Y_k^2 + G_{\min} \cdot P(H_0') \cdot Y_k^2. \tag{43}$$

When speech is absent, a minimum gain $G_{\min} = -28$ dB is used.

Finally, to determine the influence of noise estimation accuracy in the performance of the proposed estimators, we run experiments using an oracle noise estimator [10], and a different set of experiments using the minimum controlled recursive average (MCRA) noise estimator [39]. The oracle estimator of the noise variance $(\sigma_d^2(k,l))$ is computed as

$$\sigma_d^2(k,l) = \beta\sigma_d^2(k,l-1) + (1-\beta)D^2(k,l) \tag{44}$$

where $\beta = 0.85$ in this study and $D^2(k,l)$ is the true noise magnitude-squared spectrum in frame l and frequency bin k. The above oracle noise estimator was used to assess the performance of the various estimators in the absence of the confounding effect of the feedback introduced by the estimate of the noise spectrum in the computation of the *a priori* SNR in (33). To assess significant differences between the scores obtained with the various estimators, we used the Fisher's LSD statistical test.

1) Results With the Oracle Noise Estimator: Tables IV and V show the performance comparisons based on the MSE and PESQ measures respectively. In terms of MSE, lower values indicate better performance. The unprocessed corrupted speech is noted as UNProc in the Tables. The MMSE-SPZC estimator yielded significantly (significance level $p < 0.05$) lower MSE values than the MMSE-SP estimator for all four types of noise tested and for all SNR levels. The SMPR estimator yielded the lowest MSE values in most noisy conditions, followed by the SMPO estimator. The MAP estimator also yielded significantly $(p < 0.05)$ lower MSE values than the MAP-BM estimator. The MMSE-SPZC-U estimator yielded slightly higher MSE than MMSE-SPZC. The IdRM yielded lower MSE values than IdBM. This outcome was consistent with that reported in [25]. In the following discussion, comparisons in performance are analyzed only between the proposed estimators and not against the oracle estimators, IdBM and IdRM.

In terms of PESQ, higher values indicate better performance, i.e., better speech quality. The IdRM and IdBM yielded, as expected, the highest scores. The MMSE-SPZC yielded significantly higher $(p < 0.05)$ PESQ scores than MMSE-SP. The MAP estimator yielded significantly better PESQ scores than MMSE-SP, MMSE-SPZC, and MAP-BM. Finally, the performance of the SMPR and SMPO estimators was significantly higher than the other estimators (except for IdRM and IdBM), and in particular the MMSE-SP and MMSE-SPZC estimators. In babble noise (0 dB SNR), for instance, the PESQ scores improved from 1.894 with the MMSE-SP estimator [7] to 2.137 with the proposed SMPO estimator. Similar improvements were also noted at all SNR levels and with the other types of noise. The MMSE-SPZC-U estimator yielded slightly higher PESQ value than MMSE-SPZC for car, street, and babble noise, but it yielded significantly higher PESQ than the MMSE-SPZC in white-noise conditions, but still lower PESQ values than SMPR and SMPO. Overall, the SMPO estimator yielded the highest PESQ scores in all conditions.

2) Results With the MCRA Noise Estimator: Tables VI and VII show the performance, in terms of MSE and PESQ values, respectively, of the proposed estimators implemented using the MCRA noise estimation algorithm.

In terms of MSE, the MMSE-SPZC estimator yielded significantly $(p < 0.05)$ lower MSE values than MMSE-SP, for most cases except at 0 dB SNR in the street and babble noise conditions. The MMSE-SPZC-U yielded slightly higher MSE values than MMSE-SPZC. The MAP estimator yielded significantly $(p < 0.05)$ lower MSE values than MAP-BM for most cases except at 0 dB SNR in the street and babble noise conditions. The SMPR estimator yielded the lowest MSE values in the low SNR (0 dB and 5 dB) conditions, while the SMPO estimator yielded the lowest MSE values in the high SNR (10 dB and 15 dB) conditions.

In terms of PESQ, shown in Table VII, the MMSE-SPZC yielded significantly higher $(p < 0.05)$ PESQ scores than MMSE-SP. The MMSE-SPZC-U yielded slightly higher PESQ scores than MMSE-SPZC for car, street and babble noise conditions, but yielded higher (by 0.1) PESQ scores than MMSE-SPZC in white-noise conditions. The MAP estimator yielded significantly better PESQ scores than MAP-BM in

TABLE IV
PERFORMANCE COMPARISON, IN TERMS OF MSE, BETWEEN THE VARIOUS ESTIMATORS TESTED USING THE ORACLE NOISE ESTIMATOR

Noise	Method	15dB	10dB	5dB	0dB	Noise	Method	15dB	10dB	5dB	0dB
Car	UNProc	0.201	0.679	2.841	16.328	Street	UNProc	0.179	0.773	3.557	30.334
	IdBM	0.219	0.600	1.710	5.032		IdBM	0.183	0.537	2.060	4.377
	IdRM	0.202	0.511	1.239	3.126		IdRM	0.166	0.440	1.488	2.833
	MAP-BM	0.285	0.867	2.813	8.698		MAP-BM	0.235	0.806	3.319	8.713
	MMSE-SP	0.276	0.804	2.255	6.654		MMSE-SP	0.232	0.747	2.671	9.626
	MMSE-SPZC	0.232	0.657	1.885	5.696		MMSE-SPZC	0.198	0.630	2.322	7.719
	MMSE-SPZC-U	0.238	0.680	1.986	6.085		MMSE-SPZC-U	0.201	0.641	2.404	7.309
	MAP	0.246	0.726	2.211	6.751		MAP	0.208	0.681	2.653	6.404
	SMPR	0.232	0.636	**1.725**	**5.196**		SMPR	**0.194**	**0.581**	**2.149**	**5.623**
	SMPO	**0.228**	**0.631**	1.853	5.624		SMPO	0.196	0.622	2.328	7.545
Babble	UNProc	0.296	1.033	5.333	30.624	White	UNProc	0.069	0.222	0.849	4.130
	IdBM	0.289	0.786	2.534	7.121		IdBM	0.085	0.226	0.686	2.054
	IdRM	0.247	0.618	1.705	3.784		IdRM	0.083	0.208	0.579	1.566
	MAP-BM	0.371	1.216	4.428	12.355		MAP-BM	0.100	0.306	1.047	3.480
	MMSE-SP	0.365	1.087	3.724	10.372		MMSE-SP	0.104	0.292	0.939	2.878
	MMSE-SPZC	0.306	0.881	3.120	9.351		MMSE-SPZC	**0.088**	0.245	0.763	2.382
	MMSE-SPZC-U	0.310	0.903	3.208	9.333		MMSE-SPZC-U	0.090	0.251	0.796	2.522
	MAP	0.324	0.978	3.418	9.776		MAP	0.092	0.260	0.853	2.787
	SMPR	**0.298**	**0.824**	**2.818**	**7.646**		SMPR	0.092	0.244	0.737	**2.227**
	SMPO	0.305	0.881	3.070	8.691		SMPO	**0.088**	**0.238**	**0.735**	2.269

The numbers in boldface indicate the best performance.

TABLE V
PERFORMANCE COMPARISON, IN TERMS OF PESQ SCORES, BETWEEN THE VARIOUS ESTIMATORS TESTED USING THE ORACLE NOISE ESTIMATOR

Noise	Method	15dB	10dB	5dB	0dB	Noise	Method	15dB	10dB	5dB	0dB
Car	UNProc	2.53	2.20	1.89	1.63	Street	UNProc	2.54	2.25	1.90	1.56
	IdBM	3.57	3.26	2.87	2.56		IdBM	3.58	3.32	2.90	2.64
	IdRM	3.87	3.66	3.37	3.09		IdRM	3.90	3.70	3.40	3.18
	MAP-BM	2.97	2.64	2.26	1.92		MAP-BM	2.98	2.71	2.31	2.00
	MMSE-SP	2.77	2.46	2.15	1.87		MMSE-SP	2.76	2.50	2.15	1.85
	MMSE-SPZC	2.93	2.61	2.26	1.93		MMSE-SPZC	2.94	2.64	2.27	1.94
	MMSE-SPZC-U	3.03	2.72	2.39	2.06		MMSE-SPZC-U	3.02	2.74	2.39	2.04
	MAP	3.08	2.76	2.40	2.07		MAP	3.11	2.83	2.46	2.14
	SMPR	3.12	2.77	2.40	2.05		SMPR	3.13	2.82	2.43	2.07
	SMPO	**3.17**	**2.85**	**2.50**	**2.17**		SMPO	**3.15**	**2.89**	**2.53**	**2.18**
Babble	UNProc	2.65	2.32	2.01	1.71	White	UNProc	2.47	2.14	1.82	1.56
	IdBM	3.60	3.28	2.93	2.58		IdBM	3.66	3.33	2.95	2.62
	IdRM	3.87	3.64	3.37	3.11		IdRM	3.99	3.78	3.53	3.24
	MAP-BM	3.03	2.67	2.28	1.94		MAP-BM	2.99	2.68	2.32	1.98
	MMSE-SP	2.84	2.52	2.20	1.89		MMSE-SP	2.74	2.43	2.10	1.79
	MMSE-SPZC	2.98	2.64	2.30	1.97		MMSE-SPZC	2.93	2.60	2.24	1.88
	MMSE-SPZC-U	3.04	2.69	2.36	2.01		MMSE-SPZC-U	3.05	2.75	2.43	2.06
	MAP	3.16	2.82	2.47	2.09		MAP	3.09	2.79	2.47	2.12
	SMPR	3.16	2.78	2.42	2.07		SMPR	3.14	2.79	2.42	2.04
	SMPO	**3.17**	**2.82**	**2.48**	**2.14**		SMPO	**3.21**	**2.90**	**2.57**	**2.21**

The numbers in boldface indicate the best performance.

the car and white noise conditions, but no statistically significant difference $(p > 0.05)$ was noted between the MAP and MAP-BM estimators in the street and babble noise conditions. The SMPO estimator yielded significantly $(p < 0.05)$ higher PESQ scores than the other estimators in the car and white noise conditions. Finally, the performance of the SMPR estimator was significantly better than the other estimators in the street and babble noise conditions.

C. Spectrograms

Figs. 8 and 9 show sample spectrograms of speech processed by the various estimators. The sample sentence was corrupted by babble at 10 dB SNR. The IdRM output clearly resembles the clean signal. Residual noise is evident in the spectrogram showing the MMSE-SP output (Fig. 8). This residual noise is reduced considerably in the MMSE-SPZC output speech (Fig. 9). The MAP estimators greatly reduced the residual noise even further. A smaller amount of distortion was introduced with the MAP-processed speech. The SMPR speech contained more residual noise than the MAP estimator. Finally, the SMPO output speech had less speech distortion and low noise distortion. Informal listening tests confirmed that SMPO yielded the highest quality, consistent with the PESQ data shown in Table V.

VII. CONCLUSION

Statistical estimators of the magnitude-squared spectrum were derived based on the assumption that the magnitude-squared spectrum of the noisy speech signal can be computed as the sum of the clean signal and noise magnitude-squared spectrum. Aside from the two traditional estimators, based on MAP and MMSE principles, two additional soft masking methods were derived incorporating SNR uncertainty. Overall, when compared to the conventional MMSE spectral power estimators [6], [7], the proposed MAP

TABLE VI
PERFORMANCE COMPARISON, IN TERMS OF MSE, BETWEEN THE VARIOUS ESTIMATORS TESTED USING THE MCRA NOISE ESTIMATOR

Noise	Method	15dB	10dB	5dB	0dB	Noise	Method	15dB	10dB	5dB	0dB
Car	UNProc	0.201	0.679	2.841	16.328	Street	UNProc	0.179	0.773	3.557	30.334
	IdBM	0.219	0.600	1.710	5.032		IdBM	0.183	0.537	2.060	4.377
	IdRM	0.202	0.511	1.239	3.126		IdRM	0.166	0.440	1.488	2.833
	MAP-BM	0.469	1.383	3.803	12.078		MAP-BM	0.403	1.269	4.541	23.151
	MMSE-SP	0.484	1.219	2.893	8.331		MMSE-SP	0.423	1.119	3.568	18.883
	MMSE-SPZC	0.294	0.826	2.398	7.507		MMSE-SPZC	0.267	0.875	3.163	20.015
	MMSE-SPZC-U	0.308	0.875	2.582	8.249		MMSE-SPZC-U	0.279	0.910	3.319	20.285
	MAP	0.342	1.004	3.047	9.035		MAP	0.307	1.021	3.895	23.110
	SMPR	0.310	0.840	2.213	**6.991**		SMPR	0.278	0.838	**2.910**	**17.286**
	SMPO	**0.252**	**0.706**	**2.162**	6.993		SMPO	**0.228**	**0.789**	3.039	22.243
Babble	UNProc	0.296	1.033	5.333	30.624	White	UNProc	0.069	0.222	0.849	4.130
	IdBM	0.289	0.786	2.534	7.121		IdBM	0.085	0.226	0.686	2.054
	IdRM	0.247	0.618	1.705	3.784		IdRM	0.083	0.208	0.579	1.566
	MAP-BM	0.590	1.775	6.039	19.207		MAP-BM	0.180	0.505	1.587	4.777
	MMSE-SP	0.568	1.495	4.785	14.965		MMSE-SP	0.226	0.519	1.366	3.657
	MMSE-SPZC	0.389	1.196	4.440	15.483		MMSE-SPZC	0.120	0.311	0.945	2.841
	MMSE-SPZC-U	0.401	1.243	4.550	15.514		MMSE-SPZC-U	0.125	0.327	1.009	3.103
	MAP	0.459	1.464	5.311	19.476		MAP	0.138	0.370	1.160	3.532
	SMPR	0.395	1.108	**4.026**	**12.625**		SMPR	0.136	0.327	0.945	2.760
	SMPO	**0.341**	**1.050**	4.438	18.278		SMPO	**0.100**	**0.262**	**0.790**	**2.427**

The numbers in boldface indicate the best performance.

TABLE VII
PERFORMANCE COMPARISON, IN TERMS OF PESQ SCORES, BETWEEN THE VARIOUS ESTIMATORS TESTED USING THE MCRA NOISE ESTIMATOR

Noise	Method	15dB	10dB	5dB	0dB	Noise	Method	15dB	10dB	5dB	0dB
Car	UNProc	2.53	2.20	1.89	1.63	Street	UNProc	2.54	2.25	1.90	1.56
	IdBM	3.57	3.26	2.87	2.56		IdBM	3.58	3.32	2.90	2.64
	IdRM	3.87	3.66	3.37	3.09		IdRM	3.90	3.70	3.40	3.18
	MAP-BM	2.81	2.46	2.08	1.76		MAP-BM	2.70	2.38	2.01	1.65
	MMSE-SP	2.73	2.42	2.09	1.80		MMSE-SP	2.68	2.40	2.07	1.78
	MMSE-SPZC	2.89	2.56	2.18	1.88		MMSE-SPZC	2.82	2.51	2.17	1.82
	MMSE-SPZC-U	2.95	2.63	2.26	1.93		MMSE-SPZC-U	2.85	2.54	2.19	1.85
	MAP	2.88	2.55	2.14	1.85		MAP	2.73	2.41	2.04	1.67
	SMPR	**3.02**	2.66	2.27	1.94		SMPR	**2.92**	**2.60**	**2.24**	**1.87**
	SMPO	3.01	**2.67**	**2.28**	**1.97**		SMPO	2.86	2.53	2.18	1.83
Babble	UNProc	2.65	2.32	2.01	1.71	White	UNProc	2.47	2.14	1.82	1.56
	IdBM	3.60	3.28	2.93	2.58		IdBM	3.66	3.33	2.95	2.62
	IdRM	3.87	3.64	3.37	3.11		IdRM	3.99	3.78	3.53	3.24
	MAP-BM	2.79	2.38	1.99	1.62		MAP-BM	2.85	2.55	2.18	1.87
	MMSE-SP	2.79	2.46	2.15	1.80		MMSE-SP	2.70	2.40	2.05	1.77
	MMSE-SPZC	2.90	2.53	2.19	1.84		MMSE-SPZC	2.89	2.57	2.18	1.83
	MMSE-SPZC-U	2.92	2.54	2.18	1.82		MMSE-SPZC-U	2.99	2.69	2.33	1.95
	MAP	2.79	2.41	2.00	1.61		MAP	2.94	2.67	2.31	1.98
	SMPR	**2.98**	**2.58**	**2.22**	**1.87**		SMPR	3.05	2.70	2.31	1.94
	SMPO	2.91	2.535	2.13	1.76		SMPO	**3.08**	**2.78**	**2.44**	**2.07**

The numbers in boldface indicate the best performance.

estimators that incorporated SNR uncertainty yielded significantly better speech quality. The main contribution of this paper is the finding that the gain function of the MAP estimator of the magnitude-squared spectrum is identical to that of the ideal binary mask. This finding is important as it suggests that the MAP estimator of the magnitude-squared spectrum has the potential of improving speech intelligibility, given the past success of the ideal binary mask in improving, and in most cases, restoring speech intelligibility at extremely low SNR levels [23], [24], [36]. The challenge remaining is to find techniques that can estimate the local SNR reliably from the noisy observations.

APPENDIX A

In this Appendix, we derive the convergence of the MMSE gain function, given in (19), in the case that $\sigma_x^2(k) = \sigma_d^2(k)$ or equivalently when $\xi = 1$. When $\xi \neq 1$, we have

$$G_{\text{MMSE}}^2 = \frac{1}{\nu} - \frac{1}{e^\nu - 1} = \frac{e^\nu - 1 - \nu}{\nu(e^\nu - 1)}. \tag{45}$$

When $\xi \to 1$, $\nu \to 0$, and $(1/\nu) \to +\infty$. To avoid the singularity, we use the Taylor series expansion of the exponential term

$$e^\nu = 1 + \nu + \frac{\nu^2}{2} + \cdots. \tag{46}$$

In doing so, we get

$$\begin{aligned}\lim_{\nu \to 0} G_{\text{MMSE}}^2 &= \lim_{\nu \to 0} \frac{e^\nu - 1 - \nu}{\nu\left(e^\nu - 1\right)} \\ &= \lim_{\nu \to 0} \frac{\frac{\nu^2}{2} + \cdots}{\nu\left(\nu + \frac{\nu^2}{2} + \cdots\right)} = \frac{1}{2}.\end{aligned} \tag{47}$$

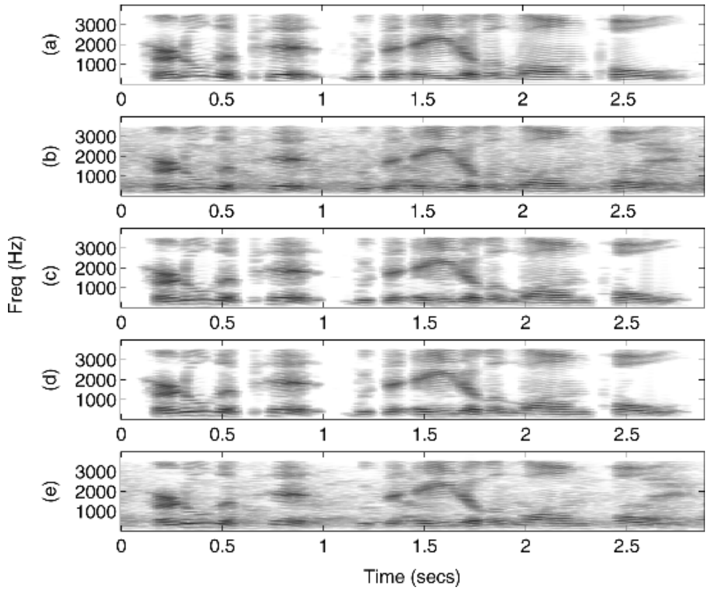

Fig. 8. Wideband spectrograms of (a) the clean sentence, b) the sentence corrupted by babble noise at 10 dB SNR, (c) the sentence processed by IdBM [25], (d) the sentence processed by IdRM [43], and (e) the sentence processed by the MMSE-SP estimator [7]. The sentence ("Hurdle the pit with the aid of a long pole") was taken from the NOIZEUS database.

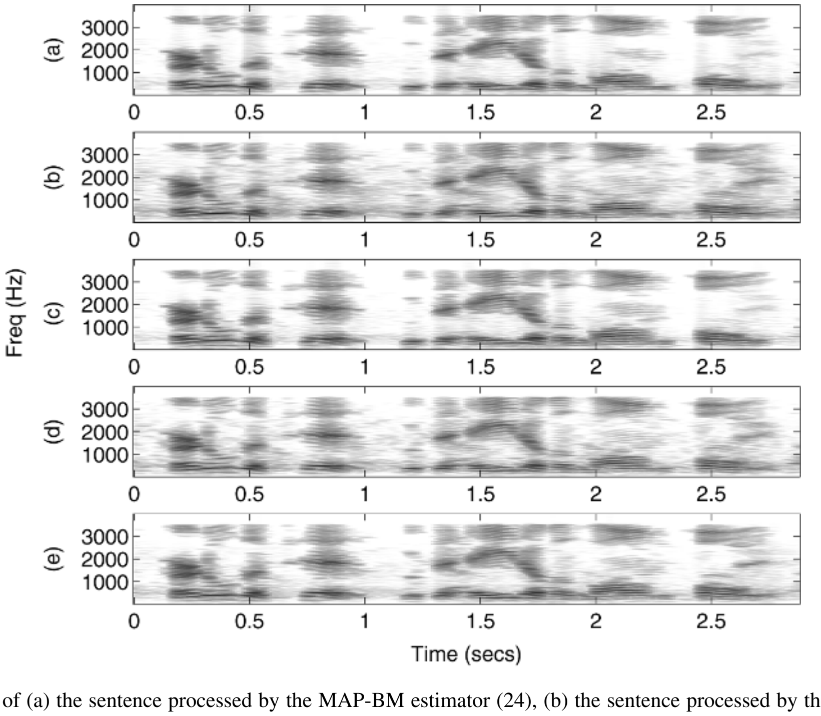

Fig. 9. Wideband spectrograms of (a) the sentence processed by the MAP-BM estimator (24), (b) the sentence processed by the MMSE-SPZC estimator (19), (c) the sentence processed by the MAP estimator (21), (d) the sentence processed by the SMPR estimator ($\theta = 5$ dB) (29), and (e) the sentence processed by the SMPO estimator ($\theta = 0$ dB) (32). The sentence was the same as in Fig. 8 and was corrupted by babble noise at 10 dB SNR.

Appendix B

In this Appendix, we derive the *a priori* distribution of the instantaneous SNR, ξ_I.

Let $\{p_i\}$ and $\{q_j\}$ be independently and identically distributed Gaussian random variables, with $p_i \sim N(0,\sigma_1^2)$ and $q_i \sim N(0,\sigma_2^2)$. Let p and q denote the sum of their squares

$$p=\sum_{i=1}^{m}p_i^2, \quad q=\sum_{j=1}^{n}q_j^2 \tag{48}$$

If $\sigma_1^2=\sigma_2^2=1$, then $F=(p/m)/(q/n)$ is known to be F-distributed [31, p. 208]

$$f_F(z)=\frac{\Gamma\left(\frac{m+n}{2}\right)m^{\frac{m}{2}}n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\frac{z^{\frac{m}{2}-1}}{(n+mz)^{\frac{m+n}{2}}}u(z) \tag{49}$$

where $\Gamma(\cdot)$ denotes the Gamma function. In our case, $p=X_k^2$, $q=D_k^2$, $m=n=2$, and $\sigma_1^2=\sigma_x^2/2$ and $\sigma_2^2=\sigma_d^2/2$. We can then express the instantaneous SNR, ξ_I, as

$$\xi_I=\frac{p}{q}=\frac{m\sigma_1^2}{n\sigma_2^2}F. \tag{50}$$

From that, we can obtain the probability density of ξ_I as

$$\begin{aligned} f_{\xi_I}(z) &= \frac{n\sigma_2^2}{m\sigma_1^2} f_F\left(z\frac{n\sigma_2^2}{m\sigma_1^2}\right) \\ &= \frac{1}{\xi} f_F\left(\frac{z}{\xi}\right) \\ &= \frac{\xi}{(\xi+z)^2} u(z) \end{aligned} \tag{51}$$

where $u(z)$ is the step function and ξ is the *a priori* SNR.

References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Taylor & Francis, 2007.

[2] 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.

[3] 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.

[4] Y. Hu and P. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[5] G. H. Ding, T. Huang, and B. Xu, "Suppression of additive noise using a power spectral density MMSE estimator," *IEEE Signal Process. Lett.*, vol. 11, no. 6, pp. 585–588, Jun. 2004.

[6] A. Accardi and R. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'99)*, Phoenix, AZ, May 1999, pp. 201–204.

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

[8] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[9] 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, no. 7–8, pp. 530–541, 2007.

[10] 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.

[11] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. 11th IEEE Signal Process. Workshop Statist. Signal Process.*, Aug. 2001, pp. 496–499.

[12] T. Lotter and P. Vary, "Noise reduction by maximum a posteriori spectral amplitude estimation with super Gaussian speech modeling," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC'03)*, Kyoto, Japan, Sep. 2003, pp. 83–86.

[13] T. Lotter and P. Vary, "Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super Gaussian speech modeling," in *Proc. EUSIPCO*, Vienna, Austria, Sep. 2004, pp. 1457–1460.

[14] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 1, pp. 1110–1126, 2005.

[15] 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.

[16] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, pp. 208–211.

[17] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[18] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 328–337, Jul. 1998.

[19] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Norwell, MA: Kluwer, 2000, pp. 155–178.

[20] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1048–1062, Sep. 2005.

[21] Y. Lu and P. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, Jun. 2008.

[22] *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds. Piscataway, NJ: Wiley/IEEE Press, 2006.

[23] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "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, 2006.

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

[25] Y. Li and D. Wang, "On the optimality of ideal binary time-frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, Mar. 2009.

[26] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[27] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[28] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, 1995.

[29] M. Jansen, *Noise Reduction by Wavelet Thresholding*, ser. Lecture notes in Statistics. Berlin, Germany: Springer-Verlag, 2001, vol. 161.

[30] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens, "A study of the distribution of time-domain speech samples and discrete Fourier coefficients," *Proc. SPS-DARTS*, vol. 1, pp. 155–158, 2005.

[31] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

[32] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[33] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1999.

[34] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, May 2008.

[35] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–299, Apr. 1994.

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

[37] G. Kim and P. C. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2080–2090, Sep. 2010.

[38] 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.

[39] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

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

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

[42] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement.," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[43] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time–frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, pp. 1486–1501, Nov. 2006.

[44] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

Yang Lu received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, and the Institute of Acoustics, Chinese Academy of Sciences, Beijing, in 2002 and 2005, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas, Richardson, in 2010.

He worked as a Research Intern with Dolby Labs, San Francisco, CA, in the summer of 2008. He is now with Cirrus Logic, Austin, TX, as a DSP Engineer. His research interests include speech enhancement, microphone array, and general audio signal processing.

Philipos C. Loizou (S'90–M'91–SM'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, in 1989, 1991, and 1995, respectively.

From 1995 to 1996, he was a Postdoctoral Fellow in the Department of Speech and Hearing Science, Arizona State University, working on research related to cochlear implants. He was an Assistant Professor at the University of Arkansas, Little Rock, from 1996 to 1999. He is now a Professor and holder of the Cecil and Ida Green Chair in the Department of Electrical Engineering, University of Texas at Dallas. His research interests are in the areas of signal processing, speech processing, and cochlear implants. He is the author of the textbook *Speech Enhancement: Theory and Practice* (CRC Press, 2007) and coauthor of the textbooks *An Interactive Approach to Signals and Systems Laboratory* (National Instruments, 2008) and *Advances in Modern Blind Signal Separation Algorithms: Theory and Applications* (Morgan & Claypool, 2010).

Dr. Loizou is a Fellow of the Acoustical Society of America. He is currently an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and *International Journal of Audiology*. He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1999–2002), IEEE SIGNAL PROCESSING LETTERS (2006–2009), and a member of the Speech Technical Committee (2008–2010) of the IEEE Signal Processing Society.