Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum

Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—The traditional minimum mean-square error (MMSE) estimator of the short-time spectral amplitude is based on the minimization of the Bayesian squared-error cost function. The squared-error cost function, however, is not subjectively meaningful in that it does not necessarily produce estimators that emphasize spectral peak (formants) information or estimators which take into account auditory masking effects. To overcome the shortcomings of the MMSE estimator, we propose in this paper Bayesian estimators of the short-time spectral magnitude of speech based on perceptually motivated cost functions. In particular, we use variants of speech distortion measures, such as the Itakura–Saito and weighted likelihood-ratio distortion measures, which have been used successfully in speech recognition. Three classes of Bayesian estimators of the speech magnitude spectrum are derived. The first class of estimators emphasizes spectral peak information, the second class uses a weighted-Euclidean cost function that implicitly takes into account auditory masking effects, and the third class of estimators is designed to penalize spectral attenuation. Of the three classes of Bayesian estimators, the estimators that implicitly take into account auditory masking effect performed the best in terms of having less residual noise and better speech quality.**

***Index Terms*—Minimum mean-square error (MMSE) estimators, perceptually-motivated speech enhancement, speech distortion measures, speech enhancement.**

I. Introduction

SINGLE-CHANNEL speech enhancement algorithms based on minimum mean-square error (MMSE) estimation of the short-time spectral magnitude have received a lot of attention in the past two decades [1]–[6], and are often compared against new algorithms. The MMSE estimators have been very popular, partly because they have been shown to be successful in eliminating musical noise [7].

It is known from estimation theory, that the MMSE estimator minimizes the Bayes risk based on a squared-error cost function [8]. The squared-error cost function is most commonly used because it is mathematically tractable and easy to evaluate. It might not be subjectively meaningful, however, in that small and large squared estimation errors might not necessarily correspond to good and poor speech quality respectively. Also, the squared error criterion might not necessarily produce estimators that preserve spectral peak (formant) information or estimators that take into account auditory masking effects. Lastly, the squared error cost function treats positive and negative estimation errors the same way. But the perceptual effect of positive error (i.e., the estimated magnitude is smaller than the true magnitude) and negative error (i.e., the estimated magnitude is larger than the true magnitude) is not the same in speech enhancement applications. Hence, the positive and negative errors need not be weighted equally.

To overcome the above problems and shortcomings of the squared-error cost function, we propose in this paper Bayesian estimators of the short-time spectral magnitude of speech based on perceptually motivated distortion measures. In particular, we use variants of speech distortion measures, such as the Itakura–Saito and weighted likelihood-ratio distortion measures, which have been applied successfully in speech recognition applications [9]–[11]. These distortion measures have been shown to be subjectively more meaningful than the squared error measure and have been applied to speech recognition tasks [12].

Three classes of Bayesian estimators are derived in this paper based on these distortion measures. In the first class, Bayesian estimators are derived that place more emphasis on spectral peaks (formants) than on spectral valleys. In the second class, Bayesian estimators are derived that take into account auditory masking effects. Lastly, in the third class, a Bayesian estimator is derived which preserves weak (low-energy) segments of speech, such as fricatives and stop consonants. This was done by using a distortion measure which penalizes positive estimation errors more than negative errors. MMSE estimators do not typically do well with such low-energy speech segments because of the low segmental SNR associated with such segments.

This paper is organized as follows. Section II provides an overview of general Bayesian estimators, Section III derives the perceptually motivated Bayesian estimators, Section IV presents the experimental results, and Section V presents the conclusions.

II. General Bayesian Estimators: Background

Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean signal $x(n)$ and the noise signal $d(n)$. Taking the short-time Fourier transform of $y(n)$, we get

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

Manuscript received June 24, 2004; revised July 17, 2004. This work was supported by NIDCD/NIH. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

The author is with the Department of Electrical Engineering, University of Texas-Dallas, Richardson, TX 75083-0688 USA (e-mail: loizou@utdallas.edu).

Digital Object Identifier 10.1109/TSA.2005.851929

for $\omega_k = 2\pi k/N$ and $k = 0, 1, 2, \ldots, N-1$, where N is the frame length in samples. 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.

In this paper, we are interested in estimating the magnitude spectrum, X_k from the noisy complex speech spectrum, $Y(\omega_k)$. Let $\varepsilon = X_k - \hat{X}_k$ denote the error in estimating the magnitude X_k at frequency bin k, and let $d(\varepsilon) \triangleq d(X_k, \hat{X}_k)$ denote a non-negative function of ε. The average cost, i.e., $E\left[d(X_k, \hat{X}_k)\right]$, is known as the *Bayes risk* $\Re_B$, and is given by

$$\begin{aligned}\Re_B &= E\left[d(X_k, \hat{X}_k)\right] \\ &= \int\int d(X_k, \hat{X}_k) p\left(X_k, Y(\omega_k)\right) dX_k dY(\omega_k) \\ &= \int\left[\int d(X_k, \hat{X}_k) p(X_k|Y(\omega_k)) dX_k\right] p(Y(\omega_k)) dY(\omega_k). \end{aligned} \tag{3}$$

Minimizing the *Bayes risk* $\Re_B$ with respect to $\hat{X}_k$ for a given cost function results in a variety of estimators. If we use the squared-error cost function $d(X_k, \hat{X}_k) = (X_k - \hat{X}_k)^2$ in (3), and we minimize the inner integral with respect to $\hat{X}_k$, while holding $Y(\omega_k)$ fixed, then we get the traditional MMSE estimator $E\left[X_k|Y(\omega_k)\right]$ [1]. If we use the "hit-or-miss" function for $d(X_k, \hat{X}_k)$, then we get the MAP estimator [8]. If we use the following cost function:

$$d_{\mathrm{LOG}}(X_k, \hat{X}_k) = \left(\log X_k - \log \hat{X}_k\right)^2 \tag{4}$$

then we get the log-MMSE estimator [2]. Non-linear cost functions that incorporated psychoacoustic constraints were proposed in [4], [5]. However, due to the nonlinearity of the constraints, no closed form solution was derived for the Bayesian estimators [4].

In summary, different Bayesian estimators of X_k can be derived depending on the choice of the cost function. Aside from the cost functions used in [4], [5], and the log square-error cost function used in [2] (since loudness is often modeled by a log function), the squared-error type cost functions used in [1]–[3], [6] were not necessarily subjectively meaningful. Next, we derive Bayesian estimators of X_k based on perceptually motivated cost functions in place of the squared-error cost function. We refer to these cost functions as "distortion measures," as they do not necessarily satisfy the metric requirements of symmetry and triangle inequality [10].

III. Perceptually Motivated Bayesian Estimators of the Speech Magnitude Spectrum

A. Psychoacoustically Motivated Distortion Measure

The proposed distortion measure is motivated by the perceptual weighting technique used in low-rate analysis-by-synthesis speech coders [13]. In most low-rate speech coders (e.g., CELP), the excitation used for LPC synthesis is selected in a closed-loop fashion using a perceptually weighted error criterion [14], [15]. This error criterion exploits the masking properties of the auditory system. More specifically, it is based on the fact that the auditory system has a limited ability in detecting quantization noise near the high-energy regions of the spectrum (e.g., near the formant peaks). Quantization noise near the formant peaks is masked by the formant peaks, and is therefore not audible. Auditory masking can be exploited by shaping the frequency spectrum of the error (estimation error in our case) so that less emphasis is placed near the formant peaks and more emphasis is placed on the spectral valleys, where any amount of noise present will be audible. We are referring here to simultaneous masking and not temporal masking (forward or backward masking) which extends in time outside the period the masker is present. Non-stationary masking effects such as forward and backward masking are not modeled in this work.

In speech coding, the perceptually-weighted error criterion is implemented by weighting the error spectrum with a filter which has the shape of the inverse spectrum of the original signal. That way, spectral peaks are not emphasized as much as spectral valleys. As a crude approximation to this perceptual weighting filter, we considered weighting the estimation error by $1/X_k$. We therefore considered the following cost function:

$$d(X_k, \hat{X}_k) = \frac{\left(X_k - \hat{X}_k\right)^2}{X_k}. \tag{5}$$

It is clear that the above distortion measure penalizes the estimation error more heavily when X_k is small (spectral valley) than when X_k is large (spectral peak). The following Bayesian risk (corresponding to the inner integral in (3), and denoted henceforth as $\Re$) was then minimized

$$\Re = \int_0^\infty \left[\frac{(X_k - \hat{X}_k)^2}{X_k}\right] p\left(X_k|Y(\omega_k)\right) dX_k. \tag{6}$$

Taking the derivative of $\Re$ with respect to $\hat{X}_k$ and setting it equal to zero, we get

$$\frac{\partial \Re}{\partial \hat{X}_k} = \int_0^\infty -2\frac{X_k - \hat{X}_k}{X_k} p\left(X_k|Y(\omega_k)\right) dX_k = 0. \tag{7}$$

Solving for $\hat{X}_k$ we get

$$\hat{X}_k = \frac{1}{\int_0^\infty \frac{1}{X_k} p\left(X_k|Y(\omega_k)\right) dX_k}. \tag{8}$$

Using the Gaussian statistical model, it can be shown (see Appendix A) that $\hat{X}_k$ evaluates to

$$\hat{X}_k = \frac{\sqrt{\lambda_k}}{\Gamma\left(\frac{1}{2}\right)} \frac{1}{\Phi\left(\frac{1}{2}, 1; -v_k\right)} \tag{9}$$

where $\Phi(a, b; x)$ denotes the confluent hypergeometric function [16, eq. 9.210.1], $\Gamma(\cdot)$ denotes the gamma function and $1/\lambda_k = 1/\lambda_d(k) + 1/\lambda_x(k)$. It is easy to show that λ_k can also be written as

$$\lambda_k = \frac{\sqrt{v_k}}{\gamma_k} Y_k \tag{10}$$

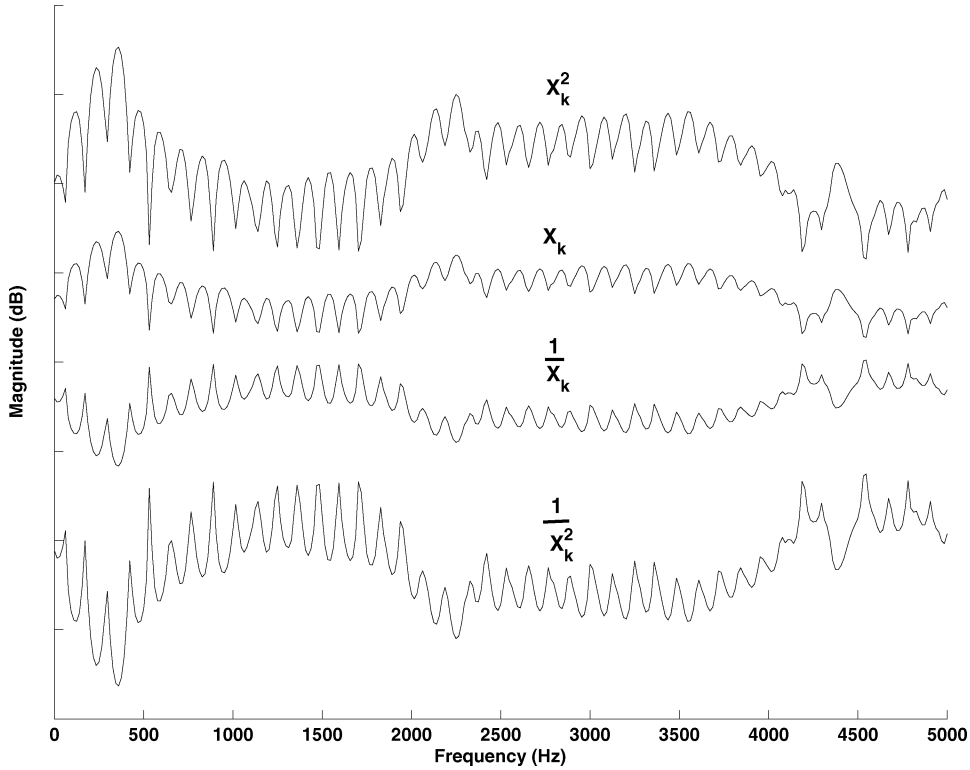


Fig. 1. Plot of the magnitude spectrum, X_k, of a 30-ms segment of the vowel /iy/ taken from the word "heed" ($\mathrm{F1} = 344$ Hz, $\mathrm{F2} = 2450$ Hz). Plots of the spectra X_k^2, $1/X_k$ and $1/X_k^2$ are superimposed for comparison. The latter spectra are shifted relative to X_k for better visual clarity.

where $v_k = \xi_k\gamma_k/(1+\xi_k)$, $\gamma_k = Y_k^2/\lambda_d(k)$, $\xi_k = \lambda_x(k)/\lambda_d(k)$, $\lambda_x(k) \triangleq E\left[X_k^2\right]$ and $\lambda_d(k) \triangleq E\left[D_k^2\right]$. Using (10), we can also express (9) as

$$\hat{X}_k = \frac{\sqrt{v_k}}{\Gamma\left(\frac{1}{2}\right)\gamma_k}\frac{1}{\Phi\left(\frac{1}{2},1;-v_k\right)}Y_k \tag{11}$$

where $\Gamma(1/2) = \sqrt{\pi}$. The above confluent hypergeometric function can also be written in terms of a Bessel function [17, eq. A1.31b], thereby simplifying the above estimator to

$$\hat{X}_k = \frac{\sqrt{v_k}}{\sqrt{\pi}\,\gamma_k}\frac{\exp\left(\frac{v_k}{2}\right)}{I_0\left(\frac{v_k}{2}\right)}Y_k \tag{12}$$

where $I_0(\cdot)$ denotes the modified Bessel function of order zero. It is worthwhile noting that the above estimator becomes the Wiener estimator when $v_k \gg 1$. To prove that, after substituting in (12) the approximation of the Bessel function, $I_0(v_k/2) \approx (1/\sqrt{\pi v_k})\exp(v_k/2)$ (for $v_k \gg 1$), we get

$$\hat{X}_k \approx \frac{\sqrt{v_k}}{\gamma_k}\sqrt{v_k}\,Y_k = \frac{\xi_k}{\xi_k+1}Y_k \qquad v_k \gg 1 \tag{13}$$

which is the Wiener estimator.

Next, we considered generalizing the cost function given in (5) to weigh the estimation error by X_k^p, i.e.,

$$d_{WE}(X_k,\hat{X}_k) = X_k^p\left(X_k-\hat{X}_k\right)^2. \tag{14}$$

Note that the above distortion measure emphasizes spectral peaks when $p > 0$, but emphasizes spectral valleys when $p < 0$. This is illustrated in Fig. 1. For $p = -2$, the above distortion measure is similar to the model distortion measure proposed by Itakura [11] for comparing two autoregressive speech models. The cost function used in (5) is obtained by setting $p = -1$. We refer to the above distortion measure as the weighted Euclidean distortion measure, since it can be written as $d_{WE}(X,\hat{X}) = \left(X-\hat{X}\right)^T W\left(X-\hat{X}\right)$, where W is a diagonal matrix, having as the kth diagonal element, $[W]_{kk} = X_k^p$. Using (14), the following risk is then minimized:

$$\Re = \int_0^\infty X_k^p\left(X_k-\hat{X}_k\right)^2 p\left(X_k|Y(\omega_k)\right)dX_k. \tag{15}$$

Taking the derivative of $\Re$ with respect to $\hat{X}_k$ and setting it equal to zero, we get

$$\frac{\partial\Re}{\partial\hat{X}_k} = \int_0^\infty -2X_k^p(X_k-\hat{X}_k)p\left(X_k|Y(\omega_k)\right)dX_k = 0. \tag{16}$$

Solving for $\hat{X}_k$ we get

$$\hat{X}_k = \frac{\int_0^\infty X_k^{p+1}p\left(X_k|Y(\omega_k)\right)dX_k}{\int_0^\infty X_k^p p\left(X_k|Y(\omega_k)\right)dX_k}. \tag{17}$$

Note that the above Bayesian estimator is the ratio of the $(p+1)$ moment of the posterior pdf $p\left(X_k|Y(\omega_k)\right)$ and the pth moment of $p\left(X_k|Y(\omega_k)\right)$, i.e., it can be written as: $\hat{X}_k = E\left[X_k^{p+1}|Y(\omega_k)\right]/E\left[X_k^p|Y(\omega_k)\right]$. In our case, p is not restricted to be an integer, however. Note also that when $p = 0$, we get the traditional MMSE estimator derived in [1].

Using the Gaussian statistical model [1], we can show (see Appendix A) that $\hat{X}_k$ evaluates to

$$\hat{X}_k = \frac{\sqrt{v_k}}{\gamma_k} \frac{\Gamma\left(\frac{p+1}{2}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)} \frac{\Phi\left(-\frac{p+1}{2}, 1; -v_k\right)}{\Phi\left(-\frac{p}{2}, 1; -v_k\right)} Y_k, \qquad p > -2. \tag{18}$$

The above equation allows us to express $\hat{X}_k$ in terms of a nonlinear gain function $G_p(\xi_k, \gamma_k) = \hat{X}_k / Y_k$ which is a function of both the *a priori* SNR ξ_k and *posteriori* SNR γ_k, much like the gain function of the MMSE estimator [1]. Fig. 2 plots the gain function $G_p(\xi_k, \gamma_k)$ as a function of the instantaneous SNR $(\gamma_k - 1)$ for a fixed value of ξ_k ($\xi_k = -5$ dB in top panel and $\xi_k = 5$ dB in bottom panel) for several values of the power exponent p. For comparative purposes, the gain functions of the MMSE [1] and log-MMSE [2] estimators are superimposed. As can be seen, the shape of the gain function $G_p(\xi_k, \gamma_k)$ is similar to that of the MMSE and log-MMSE gain functions. The amount of attenuation seems to be dependent on the value of the power exponent p. Large and positive values of p provide small attenuation, while large and negative values of p provide heavier attenuation.

Note that for large values of γ_k the gain function $G_p(\xi_k, \gamma_k)$ converges to the MMSE gain function. In fact, $G_p(\xi_k, \gamma_k)$ converges to the Wiener gain function for $\gamma_k \gg 1$ and consequently for $v_k \gg 1$. This can be proven by substituting in (18) the following asymptotic approximation of the confluent hypergeometric function [17, eq. A1.16b]:

$$\Phi(\alpha, \beta; -v_k) \approx \frac{\Gamma(\beta)}{\Gamma(\beta - \alpha)} (v_k)^{-\alpha} \qquad v_k \gg 1. \tag{19}$$

In doing so, we get

$$\hat{X}_k \approx \frac{\sqrt{v_k}}{\gamma_k} \frac{(v_k)^{(p+1)/2}}{(v_k)^{p/2}} Y_k = \frac{\xi_k}{\xi_k + 1} Y_k \qquad v_k \gg 1 \tag{20}$$

which is the Wiener estimator.

B. Itakura–Saito Measure

The Itakura–Saito measure [18] has been used successfully in speech recognition for comparing a reference power spectrum $S(\omega)$ against a test spectrum $X(\omega)$ according to

$$d_{IS}\left(X(\omega), S(\omega)\right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\frac{S(\omega)}{X(\omega)} - \log\left(\frac{S(\omega)}{X(\omega)}\right) - 1 \right] d\omega. \tag{21}$$

Due to its asymmetric nature, the IS measure is known to provide more emphasis on spectral peaks than spectral valleys.

In this paper, we consider the IS distortion measure between the estimated and true short-time power spectra at the kth frequency bin (rather than over the whole spectrum)

$$d_{IS}(X_k^2, \hat{X}_k^2) = \frac{X_k^2}{\hat{X}_k^2} - \log\left(\frac{X_k^2}{\hat{X}_k^2}\right) - 1. \tag{22}$$

Note that $d_{IS}(X_k^2, \hat{X}_k^2) \geq 0$, since $x - \log(x) - 1 \geq 0$. It is easy to show that minimization of the following Bayesian risk:

$$\Re = \int_0^{\infty} \left[\frac{X_k^2}{\hat{X}_k^2} - \log\left(\frac{X_k^2}{\hat{X}_k^2}\right) - 1 \right] p\left(X_k | Y(\omega_k)\right) dX_k \tag{23}$$

yields the following magnitude-squared estimator

$$\hat{X}_k^2 = \int_0^{\infty} X_k^2 \, p\left(X_k | Y(\omega_k)\right) dX_k \tag{24}$$

which is also the MMSE estimator of the short-time power spectrum. So, the Bayesian estimator resulting from minimization of the IS distortion measure is the same as the MMSE estimator resulting from minimization of the following distortion measure: $d(X_k, \hat{X}_k) = \left(X_k^2 - \hat{X}_k^2\right)^2$.

It is worthwhile noting that minimization of the IS measure based on the magnitude spectra (i.e., $d_{IS}(X_k, \hat{X}_k)$) of the signal, i.e., minimization of the following Bayesian risk:

$$\Re = \int_0^{\infty} \left[\frac{X_k}{\hat{X}_k} - \log\left(\frac{X_k}{\hat{X}_k}\right) - 1 \right] p\left(X_k | Y(\omega_k)\right) dX_k \tag{25}$$

results in the MMSE estimator: $\hat{X}_k = E\left[X_k | Y(\omega_k)\right]$. To verify this, after taking the derivative of $\Re$ given in (25) with respect to $\hat{X}_k$, and setting it equal to zero, we get

$$\frac{\partial \Re}{\partial \hat{X}_k} = \int_0^{\infty} \left[-\frac{X_k}{\hat{X}_k^2} + \frac{1}{\hat{X}_k} \right] p\left(X_k | Y(\omega_k)\right) dX_k = 0. \tag{26}$$

After solving for $\hat{X}_k$, we get $\hat{X}_k = E\left[X_k | Y(\omega_k)\right]$, which is the MMSE estimator of X_k.

C. COSH Measure

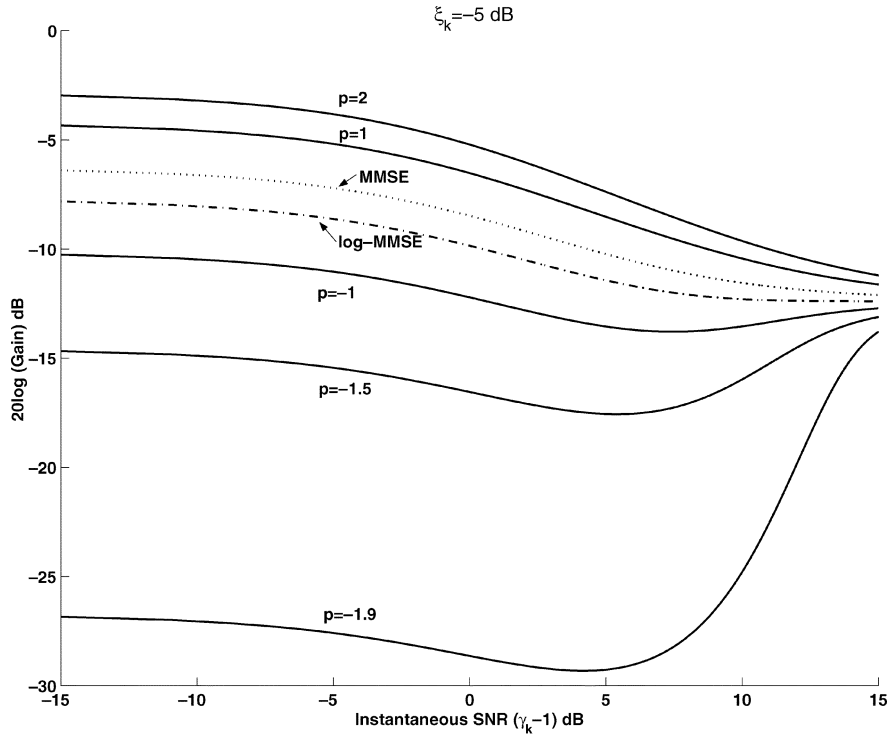
As mentioned earlier, the IS measure is asymmetric since $d_{IS}(X_k, \hat{X}_k) \neq d_{IS}(\hat{X}_k, X_k)$. A symmetric distortion measure was derived in [9] by combining the two forms of the IS measure to get a new distortion measure, termed $\cosh$ measure. The $\cosh$ measure considered here is given by

$$d_{\mathrm{COSH}}(X_k, \hat{X}_k) = \cosh\left(\log \frac{X_k}{\hat{X}_k}\right) - 1 = \frac{1}{2}\left[\frac{X_k}{\hat{X}_k} + \frac{\hat{X}_k}{X_k}\right] - 1. \tag{27}$$

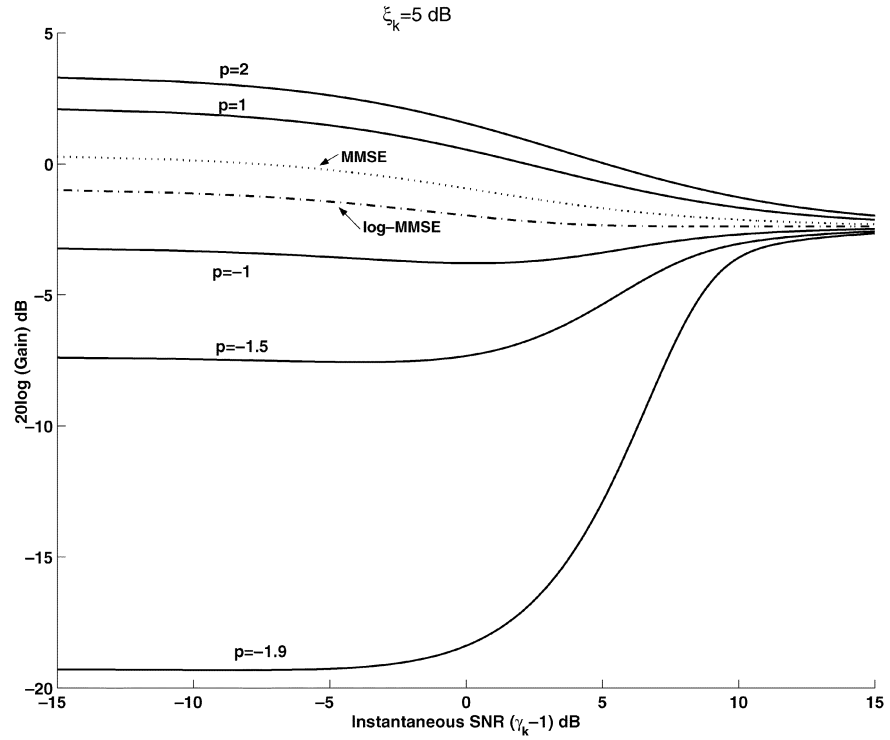
The $\cosh$ measure was shown in [9] to be nearly identical to the log spectral distortion ((4)) for small estimation errors but to differ markedly for large errors. This is illustrated in Fig. 3 which plots the $\cosh$ measure against the log spectral distortion measure, $d_{\mathrm{LOG}}(X_k, \hat{X}_k)$ given in (4). We can therefore conclude that compared to the log spectral difference measure [(4)], the $\cosh$ measure penalizes large estimation errors more heavily, but penalizes small estimation errors equally.

After minimizing the $\cosh$ risk

$$\Re = \int_0^{\infty} \left[\frac{X_k}{\hat{X}_k} + \frac{\hat{X}_k}{X_k} - 1 \right] p\left(X_k | Y(\omega_k)\right) dX_k \tag{28}$$



(a)



(b)

Fig. 2. Gain functions of the weighted-Euclidean distance estimator [(18)] as a function of the instantaneous SNR ($\gamma_k - 1$) and for several values of the power exponent p. Top panel plots the gain functions for $\xi_k = -5$ dB and bottom panel plots the gain function for $\xi_k = 5$ dB. The gain functions of the MMSE and log-MMSE estimators are also plotted for comparison.

with respect to $\hat{X}_k$ we get the following magnitude-squared estimator:

$$\hat{X}_k^2 = \frac{\int_0^\infty X_k p\left(X_k|Y(\omega_k)\right) dX_k}{\int_0^\infty \frac{1}{X_k} p\left(X_k|Y(\omega_k)\right) dX_k}. \tag{29}$$

Note that the numerator is the traditional MMSE estimator [1], and the denominator is the estimator derived in (8). Substituting (9) for the denominator and the MMSE estimator [1] for the numerator, we get

$$\hat{X}_k = \frac{1}{\gamma_k} \sqrt{\frac{v_k}{2} \frac{\Phi(-0.5, 1; -v_k)}{\Phi(0.5, 1; -v_k)}}\, Y_k. \tag{30}$$

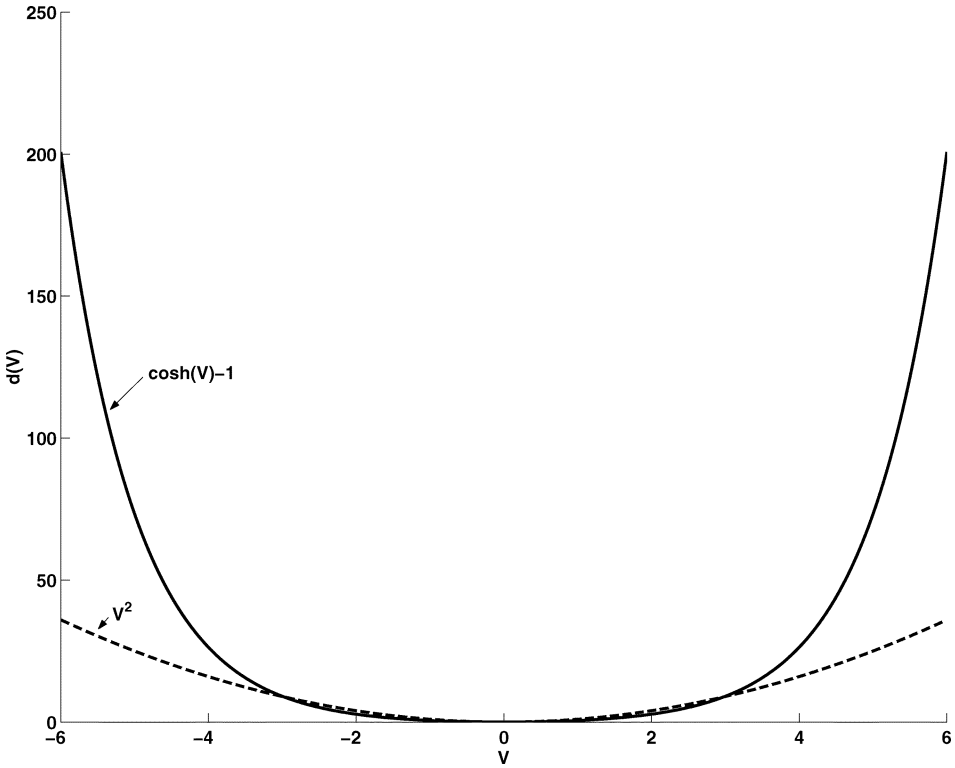


Fig. 3. Plot of the cosh distortion measure $d_{\cosh}(V) = \cosh(V) - 1$, where $V = \log X_k - \log \hat{X}_k$. The log spectral distortion measure [(4)] is also plotted (dashed line) for comparison.

The above estimator can also be expressed in terms of Bessel functions using [17, eq. A1.31a, A1.31c] as

$$\hat{X}_k = \frac{1}{\gamma_k}\sqrt{\frac{v_k + v_k^2}{2} + \frac{v_k^2\, I_1\left(\frac{v_k}{2}\right)}{2\, I_0\left(\frac{v_k}{2}\right)}}\, Y_k \tag{31}$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of order ν.

Wanting to exploit auditory masking effects, as we did with the weighted Euclidean distortion measure, we also considered the following weighted cosh distortion measure:

$$d_{W\mathrm{COSH}}(X_k, \hat{X}_k) = \left[\frac{X_k}{\hat{X}_k} + \frac{\hat{X}_k}{X_k} - 1\right] X_k^p. \tag{32}$$

Minimization of the above weighted-cosh based Bayesian risk, leads to the following magnitude-squared estimator:

$$\hat{X}_k^2 = \frac{\int_0^\infty x_k^{p+1} p\left(x_k|Y(\omega_k)\right) dx_k}{\int_0^\infty x_k^{p-1} p\left(x_k|Y(\omega_k)\right) dx_k}. \tag{33}$$

It is easy to show (see derivation in Appendix A) that the above estimator evaluates to

$$\hat{X}_k = \frac{1}{\gamma_k}\sqrt{\frac{v_k\,\Gamma\left(\frac{p+3}{2}\right)}{\Gamma\left(\frac{p+1}{2}\right)}\frac{\Phi\left(-\frac{p+1}{2}, 1; -v_k\right)}{\Phi\left(-\frac{p-1}{2}, 1; -v_k\right)}}\, Y_k, \qquad p > -1. \tag{34}$$

Fig. 4 plots the gain function $G_{W\mathrm{COSH}}(\xi_k, \gamma_k) = \hat{X}_k / Y_k$ as a function of $(\gamma_k - 1)$ for several values of p and for $\xi_k = -5\,\mathrm{dB}$. For comparative purposes we also superimpose the gain function of the log-MMSE estimator. The power exponent clearly influences attenuation with negative values providing more attenuation than positive values. When $p = 0$, we get the "unweighted" cosh estimator given in (30). Note that the cosh estimator given in (30) provides slightly more attenuation than the log-MMSE estimator. Only the parametric gain curves for $\xi_k = -5\,\mathrm{dB}$ were plotted in Fig. 4. The shape of the gain functions obtained for other values of ξ_k is similar.

D. Weighted Likelihood Ratio

As mentioned earlier, the IS measure places more emphasis on spectral peaks than spectral valleys. To further increase the sensitivity of distortion measure to the spectral peaks, Shikano and Sugiyama [19] proposed the weighted likelihood ratio (WLR) distortion measure which has the following form:

$$d_{WLR}(X_k, \hat{X}_k) = \left(\log X_k - \log \hat{X}_k\right)\left(X_k - \hat{X}_k\right). \tag{35}$$

The WLR measure can be considered to be a variant of the log spectral difference measure given in (4). The weighting function used in $d_{WLR}(X_k, \hat{X}_k)$ is the linear spectral difference $(X_k - \hat{X}_k)$ which weights log spectral peaks more than spectral valleys. In contrast, the $d_{\mathrm{LOG}}(X_k, \hat{X}_k)$ measure implicitly uses the log spectral difference, $(\log X_k - \log \hat{X}_k)$, as the weighting function, thereby weighting spectral peaks and valleys equally.

After differentiating the Bayesian risk

$$\Re = \int_0^\infty d_{WLR}(X_k, \hat{X}_k)\, p\left(X_k|Y(\omega_k)\right) dX_k \tag{36}$$

with respect to $\hat{X}_k$, we get the following nonlinear equation in $\hat{X}_k$

$$\log \hat{X}_k + a_k - \frac{b_k}{\hat{X}_k} = 0 \tag{37}$$

where

$$\begin{aligned} a_k &= 1 - E\left[\log X_k|Y(\omega_k)\right] \\ &= 1 - \frac{1}{2}\left[\log \lambda_k + \log v_k + \int_{v_k}^\infty \frac{e^{-t}}{t}\, dt\right] \end{aligned} \tag{38}$$

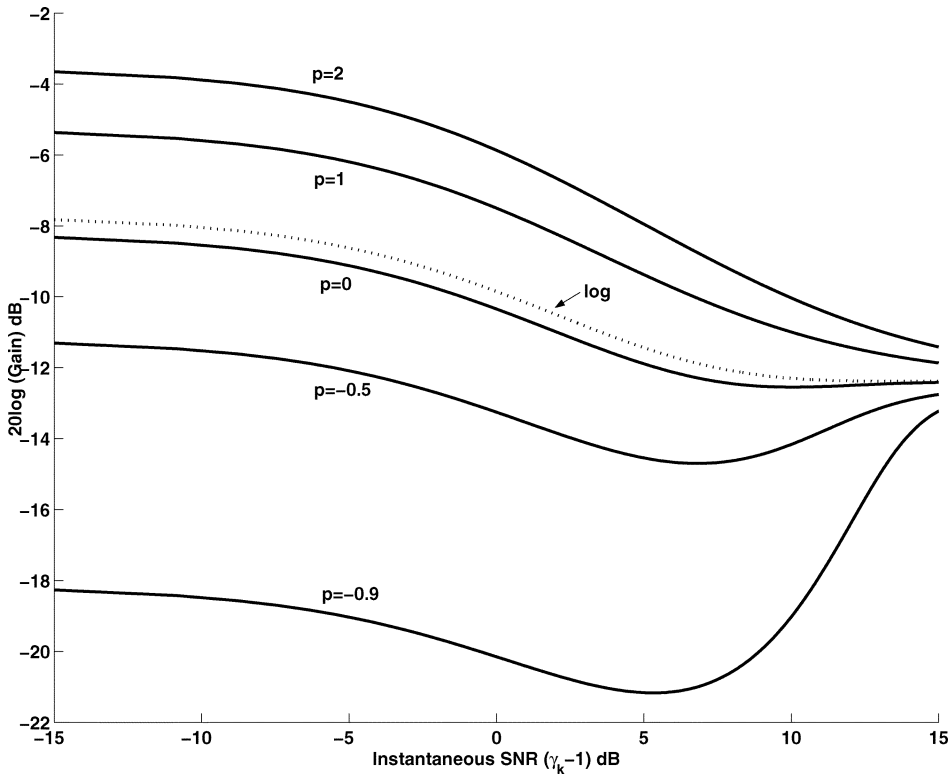


Fig. 4. Gain function of the weighted-$\cosh$ estimator [(34)] as a function of the instantaneous SNR ($\gamma_k - 1$) and for several values of the power exponent p. The *a priori* SNR is fixed at $\xi_k = -5$ dB. The gain function of the log-MMSE estimator is also plotted for comparison.

and $b_k = E[X_k|Y(\omega_k)]$ is the MMSE estimator [1]. The $E[\log X_k|Y(\omega_k)]$ term above was derived in [2]. It is easy to show that the function $g(x) = \log x + a - b/x$ in (37) is monotonically increasing in (0, ∞) with $\lim_{x\to 0+} g(x) = -\infty$ (given that $b \geq 0$) and $\lim_{x\to\infty} g(x) = \infty$, and therefore has a single zero. That is, the solution of the nonlinear equation in (37) yields a unique estimator. Numerical techniques [20] can be used to find the single zero of $g(x)$.

E. Modified Itakura–Saito Distortion Measure

With the exception of the asymmetric IS measure, the other distortion measures discussed so far were symmetric. The symmetry property is certainly desirable in pattern recognition applications, where we would like the distortion measure to yield the same value regardless of whether we compare the reference spectrum (or parametric model) against the test spectrum or the test spectrum against the reference spectrum. In speech enhancement applications, however, the distortion measure need not be symmetric, as we may want to penalize positive errors more than negative errors or vice versa. A positive estimation error ($X_k - \hat{X}_k > 0$) would suggest that the estimated spectral amplitude is attenuated since $\hat{X}_k < X_k$, while a negative error ($X_k - \hat{X}_k < 0$) would suggest that the estimated amplitude is amplified, since $\hat{X}_k > X_k$. The perceptual effects of these two types of error, however, are not equivalent and therefore the positive and negative errors need not be weighted equally. Wanting to prevent attenuation of the weak speech segments (e.g., stops, fricatives), we chose a distortion measure that penalizes the positive errors more heavily than the negative errors.

The following distortion measure was therefore considered:

$$d_{MIS}(X_k, \hat{X}_k) = \exp(X_k - \hat{X}_k) - (X_k - \hat{X}_k) - 1 \quad (39)$$

which is referred to as the modified IS (MIS) measure. Note that the original IS measure had the form $d_{IS}(x, \hat{x}) = \exp(V) - V - 1$ where $V = \log x - \log \hat{x}$, whereas in our case, $V = x - \hat{x}$. Fig. 5 plots the above measure as a function of $V_k = X_k - \hat{X}_k$. As can be seen, the above distortion measure is indeed nonsymmetric in that it penalizes the positive errors ($V_k > 0$ or equivalently, $\hat{X}_k < X_k$) more than the negative errors. After minimizing the Bayesian risk

$$\Re = \int_0^\infty \left[e^{X_k - \hat{X}_k} - X_k + \hat{X}_k - 1\right] p(X_k|Y(\omega_k))\, dX_k \quad (40)$$

with respect to $\hat{X}_k$, we get the following estimator:

$$\hat{X}_k = \log\left[\int_0^\infty e^{X_k} p(X_k|Y(\omega_k))\, dX_k\right]. \quad (41)$$

The integral in the above equation evaluates to (see derivation in Appendix B)

$$\begin{aligned}\int_0^\infty & e^{x_k} p(x_k|Y(\omega_k))\, dx_k \\ &= \exp(-v_k) \sum_{m=0}^{\infty} \frac{1}{m!} (v_k)^m F\left(-m, -m, \frac{1}{2}; \frac{Y_k^2}{4\gamma_k^2}\right) \\ &\quad + \exp(-v_k) \frac{\sqrt{v_k}}{\gamma_k} Y_k \\ &\quad \times \sum_{m=0}^{\infty} \frac{\Gamma(m+1.5)}{m!\Gamma(m+1)} (v_k)^m F\left(-m, -m, \frac{3}{2}; \frac{Y_k^2}{4\gamma_k^2}\right) \end{aligned} \quad (42)$$

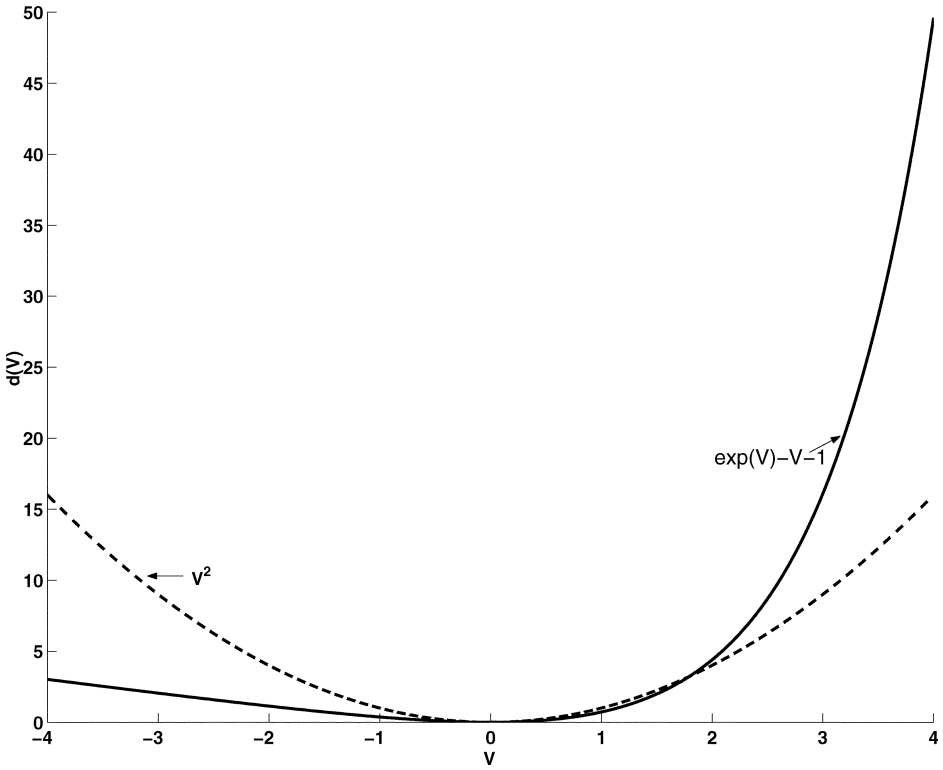


Fig. 5. Plot of the modified Itakura–Saito distortion measure $d_{MIS}(V) = \exp(V) - V - 1$, where $V = X_k - \hat{X}_k$. The squared error measure (V^2) used in the MMSE estimator is also plotted (dashed line) for comparison.

where $F(a, b, c; x)$ denotes the Gaussian hypergeometric function [16, eq. 9.100]. In our implementation, we truncated the above infinite series to the first Q terms as follows:

$$\begin{aligned}\int_0^\infty & e^{x_k} p\left(x_k|Y(\omega_k)\right) dx_k \\ &\approx \exp(-v_k) \sum_{m=0}^{Q-1} \frac{1}{m!} (v_k)^m F\left(-m, -m, \frac{1}{2}; \frac{Y_k^2}{4\gamma_k^2}\right) \\ &\quad + \exp(-v_k) \frac{\sqrt{v_k}}{\gamma_k} Y_k \\ &\quad \times \sum_{m=0}^{Q-1} \frac{\Gamma(m+1.5)}{m!\Gamma(m+1)} (v_k)^m F\left(-m, -m, \frac{3}{2}; \frac{Y_k^2}{4\gamma_k^2}\right). \quad (43)\end{aligned}$$

Good performance was obtained using Q in the range of 30 to 40. Due to highly nonlinear nature of the resulting estimator, we are unable to plot its gain function. We can easily prove, however, that the above estimator always provides less attenuation than the MMSE estimator. Acknowledging the fact that the integral in (41) is $E\left[e^{X_k}|Y(\omega_k)\right]$, and after using Jensen's inequality, we have

$$\log E\left[e^{X_k}|Y(\omega_k)\right] \geq E\left[\log(e^{X_k})|Y(\omega_k)\right] = E\left[X_k|Y(\omega_k)\right]. \quad (44)$$

IV. Results

The proposed estimators were evaluated using both objective measures and subjective listening tests. Twenty sentences from the TIMIT database were used for the objective evaluation of the proposed estimators, ten produced by female speakers and ten produced by male speakers. The TIMIT sentences were downsampled to 8 kHz. Speech-shaped noise constructed from the long-term spectrum of the TIMIT sentences was added to the clean speech files at 0, 5, and 10 dB SNR. An estimate of the noise spectrum was obtained from the initial 100-ms segment of each sentence. The noise spectrum estimate was not updated in subsequent frames.

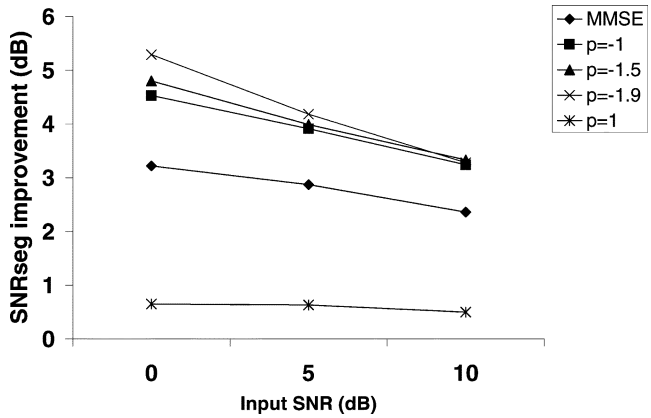


Fig. 6. Performance, in terms of segmental SNR improvement (dB), of the weighted Euclidean estimator [(18)] for different values of p and for different input SNR levels. The performance of the MMSE estimator is also shown for comparison.

The proposed estimators were applied to 20-ms duration frames of speech using a Hamming window, with 50% overlap between frames. The "decision-directed" approach [1] was used in all proposed Bayesian estimators to compute the *a priori* SNR ξ_k, with $a = 0.98$. The enhanced signal was combined

TABLE I
COMPARISON BETWEEN THE LOG-MMSE [2] AND WLR BAYESIAN ESTIMATORS [(37)] IN TERMS OF SEGMENTAL SNR IMPROVEMENT (DB) AND IN TERMS OF THE IS MEASURE

	0 dB		5 dB		10 dB	
Estimator	**SNRseg improvement**	**IS**	**SNRseg improvement**	**IS**	**SNRseg improvement**	**IS**
Log-MMSE	4.11	3.28	3.67	2.54	3.09	1.92
WLR	3.56	2.89	2.93	2.34	1.99	2.00

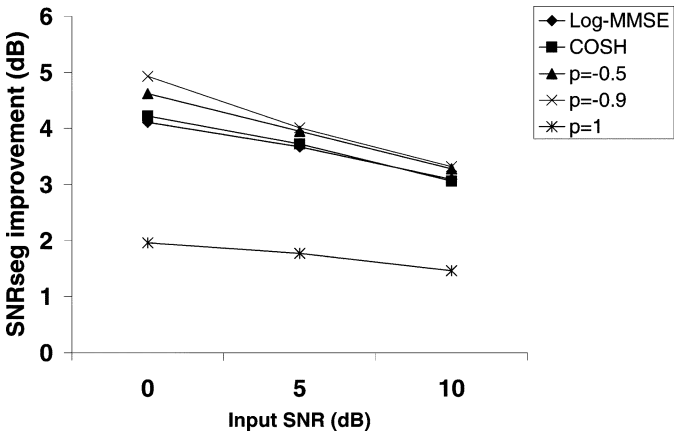


Fig. 7. Performance, in terms of segmental SNR improvement (dB), of the weighted $\cosh$ estimator [(34)] for different values of p and for different input SNR levels. The performance of the log-MMSE estimator is also shown for comparison.

using the overlap and add approach. For comparative purposes, we evaluated the performance of the MMSE and log-MMSE estimators. MATLAB implementations of the proposed estimators are available upon request from the author.

A. Objective Evaluations

Objective measures, in terms of the segmental SNR, were used to evaluate the performance of the proposed Bayesian estimators of the speech magnitude spectrum. We first compared the performance obtained with the weighted Euclidean Bayesian estimator [(18)] against the performance obtained with the MMSE estimator [1]. For the implementation of the $\Phi(a,b,c;x)$ function in (18), we used the first 100 terms of the confluent hypergeometric series. The results are shown in Fig. 6 in terms of segmental SNR improvement (over the noisy speech) for different values of p and for three input SNRs (0, 5, 10 dB). Clearly, better performance is obtained with negative values of p. Listening tests indicated that the residual noise is reduced significantly when $p < 0$. Speech distortion is introduced however when p takes on large negative values, particularly when p gets close to -2. Hence, the value of p controls the tradeoff between speech distortion and residual noise. A good compromise was found with $p = -1$ (see next section).

Fig. 7 compares the performance obtained with the $\cosh$ and weighted-$\cosh$ estimators, against the performance obtained with the log-MMSE estimator [2]. The estimator based on the weighted-$\cosh$ measure performs a little better than the log-MMSE estimator for $\mathrm{SNR} = 0$ dB, and performs equally well for higher SNR levels. Listening tests (see next section) indicated that the residual noise is reduced significantly by the weighted-$\cosh$ Bayesian estimators when $p < 0$. Speech distortion is introduced however when p gets close to -1. A good compromise between residual noise and speech distortion was found with $p = -0.5$.

TABLE II
COMPARISON BETWEEN THE MMSE [1] AND MIS BAYESIAN ESTIMATORS [(43) WITH $Q = 40$] IN TERMS OF SEGMENTAL SNR IMPROVEMENT (DB)

Estimator	0 dB	5 dB	10 dB
MMSE	3.22	2.87	2.36
MIS	2.00	-0.7	-5.14

Table I compares the performance obtained with the weighted-likelihood Bayesian estimator against the performance obtained with the log-MMSE estimator. Brent's algorithm [20] was used to solve for the WLR estimator satisfying (37). In addition to the segmental SNR measure, we also evaluated the performance of the two estimators using the Itakura–Saito distortion measure. This was done to assess the spectral-peak matching ability of the WLR estimator. Large improvement in IS values was indeed observed for $\mathrm{SNR} = 0$ dB with the WLR estimator.

Table II compares the performance obtained with the MMSE estimator against the performance obtained with the MIS estimator [(43)] using $Q = 40$. For the implementation of the $F(a,b,c;x)$ function in (43), we used the first 40 terms of the Gaussian hypergeometric series. Overall, the MMSE estimator performs better, however, closer examination of some of the enhanced signals revealed that the MIS Bayesian estimator does a better job in preserving weak (low-energy) speech segments such as stops and fricatives. This is illustrated in Fig. 8 which compares the enhanced signals obtained with the two estimators. Note that the fricative /s/ at $t \approx 0.4$, 1.4 and 2.4 secs is hardly present in the signal enhanced by the MMSE estimator, but is quite evident in the signal enhanced by the MIS estimator. The MMSE estimator, however, does a better job in enhancing the voiced segments of speech and preserving the consonant-to-vowel amplitude ratio. Informal listening tests indicated that the quality of speech produced by the MIS estimator was sensitive to the number of terms, Q, used to truncate the infinite series in (43). We found that Q in the range of 30 to 40 gives modest performance, but the MIS estimator becomes very aggressive with "musical"-type of noise if a smaller number of terms is used.

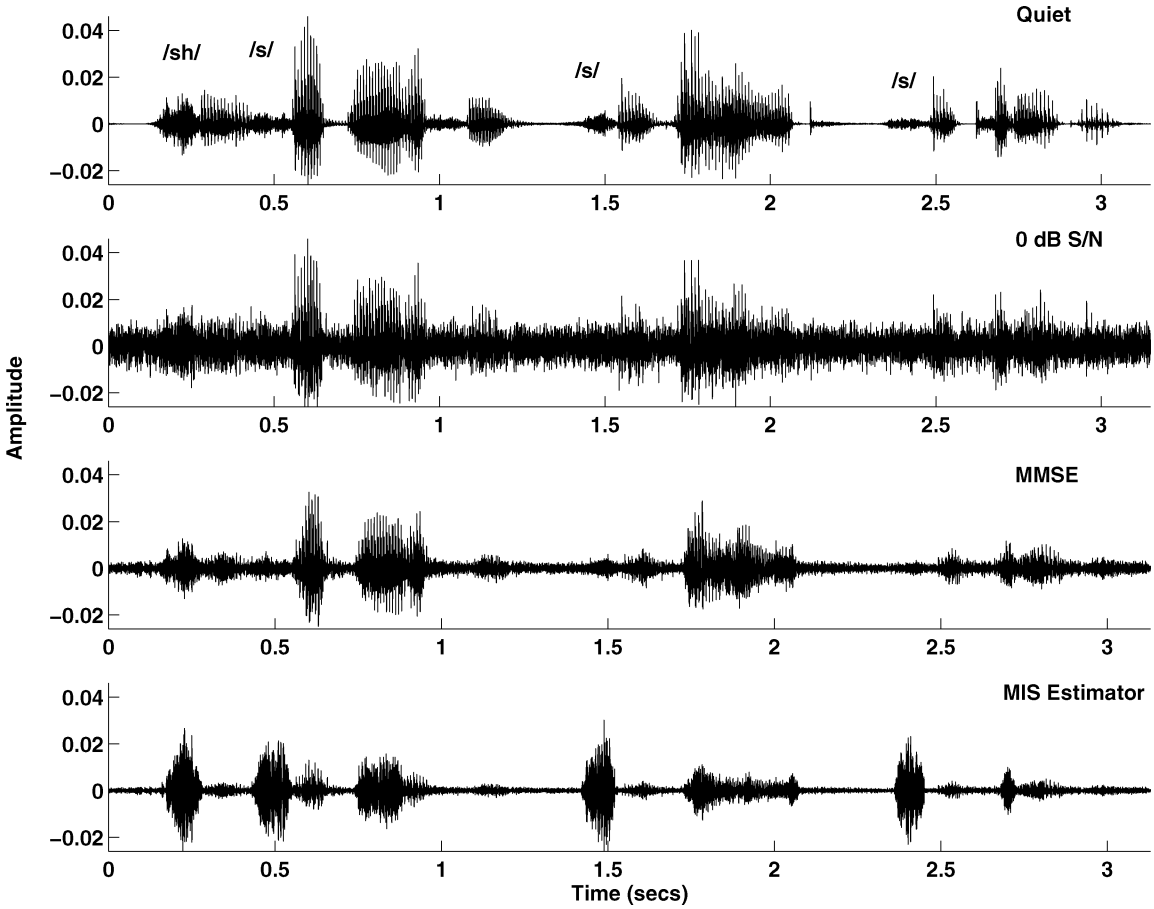


Fig. 8. Top panel shows the waveform of the TIMIT sentence "She was so beautiful, so valiant, so pitiable." produced by a male speaker. Second panel from the top shows the noisy waveform at 0 dB SNR. The bottom two panels show the enhanced signals by the MMSE and MIS Bayesian estimators. Note that the fricative /s/ at $t \approx 0.4$, 1.4 and 2.4 secs is hardly present in the signal enhanced by the MMSE estimator, but is quite evident in the signal enhanced by the MIS estimator.

B. Subjective Evaluations

Ten TIMIT sentences (subset of the sentences used for the objective evaluations) produced by five male and five male speakers were used in the listening tests. Five normal-hearing listeners, age 20–25 yrs, participated in the listening tests [all listeners were paid for their participation]. Listeners were presented randomly with 30 pairs ($= 10\,\text{sentences} \times 3\,\text{repetitions}$) of sentences processed by the MMSE estimator and the weighted-Euclidean based Bayesian estimators using $p = -1$ and $p = -1.9$. In a different listening session, listeners were presented with 30 pairs of sentences processed by the log-MMSE estimator and the weighted-cosh based Bayesian estimators using $p = -0.5$ and $p = -0.9$. Subjects were asked to (1) choose the sentence they prefer in terms of being more natural and having less distortion, and (2) indicate which sentence had more residual noise.

The results, scored in terms of preference percentage, are given in Table III. Results indicated that the value of p clearly influenced the amount of distortion perceived, with large negative value of p producing more distortion than small negative values of p. Small values of p produce speech with little distortion, but with more residual noise. Subjects preferred the quality of speech produced by the MMSE estimator over speech enhanced by the weighted-Euclidean estimator for large negative values of p, i.e., when $p = -1.9$. Subjects also preferred the quality of speech produced by the log-MMSE estimator over speech enhanced by the weighted-cosh estimator when $p = -0.9$. For moderately smaller values of p, however, subjects preferred the quality of speech produced by the proposed weighted-Euclidean estimator ($p = -1$) over speech enhanced by the MMSE estimator. Quality of speech by the log-MMSE estimator and the weighted-cosh estimator ($p = -0.5$) was found to be comparable ($p = -1.9$), but with substantially lower residual noise reported for the weighted-cosh estimator. Listeners overwhelmingly reported that speech produced by the proposed estimators had less residual noise than either the MMSE or log-MMSE estimators.

TABLE III
TOP TWO ROWS GIVE THE MEAN PERCENT PREFERENCE SCORES FOR SPEECH ENHANCED BY THE PROPOSED WEIGHTED EUCLIDEAN ESTIMATORS OVER THE MMSE ESTIMATOR. BOTTOM TWO ROWS GIVE THE MEAN PERCENT-PREFERENCE SCORES FOR SPEECH ENHANCED BY THE PROPOSED WEIGHTED-cosh ESTIMATORS OVER THE LOG-MMSE ESTIMATOR

Proposed Estimator	Percent preference
Weighted-Euclidean ($p = -1$)	70%
Weighted-Euclidean ($p = -1.9$)	36 %
Weighted cosh ($p = -0.5$)	40 %
Weighted cosh ($p = -0.9$)	17 %

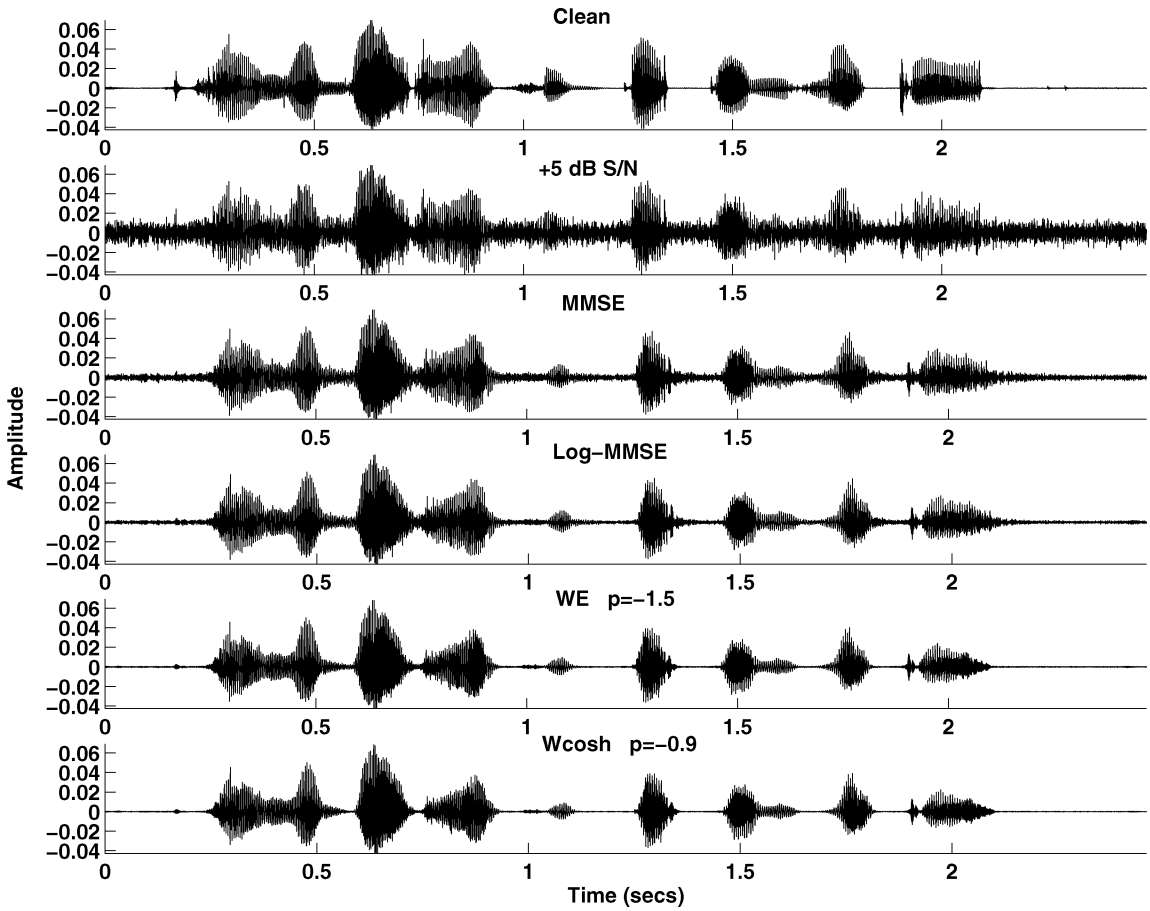


Fig. 9. Top panel shows the waveform of the TIMIT sentence "The angry boy answered, but didn't look up." produced by a female speaker. Second panel from the top shows the noisy waveform at 5 dB SNR. The remaining panels show the enhanced signals produced by the MMSE, log-MMSE, weighted-Euclidean estimator ($\mathrm{p} = -1.5$) and weighted-$\cosh$ estimator ($\mathrm{p} = -0.9$). Note that the residual noise is significantly reduced by the proposed estimators.

V. Summary and Conclusions

The present study focused on the derivation of perceptually-motivated Bayesian estimators of the magnitude spectrum. Several other perceptually-motivated methods were proposed in [21]–[23] but used a different approach for incorporating psychoacoustic constraints.

Six different Bayesian estimators of the spectral magnitude were derived in this paper. Unlike the previous MMSE estimators derived in [1], [2], the proposed Bayesian estimators are based on perceptually motivated distortion measures. Based on the evaluation of the proposed estimators, we can draw the following conclusions.

1) Bayesian estimators which over-emphasize spectral peak information performed the worst. These include the traditional MMSE estimator [1], the WLR estimator [(37)] and the estimators given in (18) and (34) with $p > 0$. The enhanced speech signal produced by these estimators (including the traditional MMSE estimator) had a significant amount of residual noise which was audible (see Fig. 9). This was confirmed by listening tests. We believe that this is due to the fact that the estimation error produced by these estimators is small near the spectral peaks (where it is masked anyway) and large in the spectral valleys, where the residual noise is audible [15].
2) Bayesian estimators that emphasize spectral valleys more than the spectral peaks performed the best in terms of having less residual noise and better speech quality (see Fig. 9). These include the estimator given in (18) with $p = -1$ and the estimator given in (34) with $p = -0.5$. Listening tests confirmed that the weighted-Euclidean estimator ($p = -1$) performed significantly better than the MMSE estimator. The weighted-$\cosh$ estimator ($p = -0.5$) performed comparably with the log-MMSE estimator, but with substantially reduced residual noise. This class of estimators exploits implicitly auditory masking effects by taking into account the fact that estimation errors near the spectral peaks are masked.
3) The derived Bayesian estimators based on the Itakura–Saito measure of the magnitude and power spectrum were identical to the MMSE estimator of the magnitude and power-spectrum, respectively.
4) The Bayesian estimator based on the asymmetric MIS measure seems to perform well in preserving weak speech segments (e.g., fricatives) but not in enhancing voiced segments. This was based on visual inspection of spectrograms and waveforms of the enhanced signals. This estimator was designed to penalize positive errors more than negative errors, thereby avoiding spectral

attenuation. Although the performance of the MIS Bayesian estimator was not consistently or equally well for all voiced and unvoiced speech segments, conceivably, a hybrid estimator can be implemented which uses the MIS estimator for unvoiced segments and a different estimator (e.g., (18) with $p < 0$) for voiced segments.

APPENDIX A

In this Appendix, we derive the estimators given in (8), (18), (29) and (33). Assuming the Gaussian statistical model [1], we know that [2]

$$E\left[X_k^p|Y(\omega_k)\right] = \frac{\int_0^\infty x_k^{p+1} \exp\left(\frac{-x_k^2}{\lambda_k}\right) I_0\left(2x_k\sqrt{\frac{v_k}{\lambda_k}}\right) dx_k}{\int_0^\infty x_k \exp\left(\frac{-x_k^2}{\lambda_k}\right) I_0\left(2x_k\sqrt{\frac{v_k}{\lambda_k}}\right) dx_k}. \tag{45}$$

Using [16, eq. 6.631.1, 8.406.3, 9.212.1] it is easy to show that [2]

$$E\left[X_k^p|Y(\omega_k)\right] = \lambda_k^{p/2}\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2},1;-v_k\right). \tag{46}$$

By setting $p = -1$ in the above equation, we can evaluate the integral in (8) to get (9).

The above equation is similarly used in (17) to evaluate the estimator given in (18). The restriction on the value of p to be larger than -2 comes from the evaluation of the integral in the numerator of (45). From [16, eq. 6.631.1], the power exponent of the x_k^{p+1} term has to be larger than -1, which leads to the condition that $p > -2$.

The cosh estimator in (29) and the weighted-cosh estimator in (33) are derived in a similar way using (46).

APPENDIX B

In this Appendix, we evaluate the MIS Bayesian estimator given in (41). Using Bayes' rule, we can write

$$\int_0^\infty e^{x_k} p\left(x_k|Y(\omega_k)\right) dx_k = \frac{1}{p\left(Y(\omega_k)\right)} \int_0^\infty e^{x_k} p\left(Y(\omega_k)|x_k\right) p(x_k)\, dx_k. \tag{47}$$

After using the Gaussian statistical model, and after integrating over θ_x, we get

$$\int_0^\infty e^{x_k} p\left(Y(\omega_k)|x_k\right) p(x_k)\, dx_k = \int_0^\infty e^{x_k} x_k \exp\left(\frac{-x_k^2}{\lambda_k}\right) I_0\left(2x_k\sqrt{\frac{v_k}{\lambda_k}}\right) dx_k \tag{48}$$

where $I_0(\cdot)$ indicates the modified Bessel function of order zero. After using the following identity for the exponential term [17, eq. A.1.47c]:

$$e^x = \sqrt{\frac{\pi x}{2}}\left[I_{-0.5}(x) + I_{0.5}(x)\right] \tag{49}$$

in (48), we get

$$\int_0^\infty e^{x_k} p\left(Y(\omega_k)|x_k\right) p(x_k)\, dx_k = \sqrt{\frac{\pi}{2}}\int_0^\infty x_k^{3/2} \exp\left(\frac{-x_k^2}{\lambda_k}\right) I_{-0.5}(x_k) I_0\left(2x_k\sqrt{\frac{v_k}{\lambda_k}}\right) dx_k + \sqrt{\frac{\pi}{2}}\int_0^\infty x_k^{3/2} \exp\left(\frac{-x_k^2}{\lambda_k}\right) I_{0.5}(x_k) I_0\left(2x_k\sqrt{\frac{v_k}{\lambda_k}}\right) dx_k. \tag{50}$$

The above integrals can be evaluated using [16, eq. 6.633.1, 8.406.3] to

$$\int_0^\infty e^{x_k} p\left(Y(\omega_k)|x_k\right) p(x_k) dx_k = \frac{\lambda_k}{2}\sum_{m=0}^\infty \frac{1}{m!}(v_k)^m F\left(-m,-m,\frac{1}{2};\frac{\lambda_k}{4v_k}\right) + \frac{(\lambda_k)^{3/2}}{2}\sum_{m=0}^\infty \frac{\Gamma(m+1.5)}{m!\Gamma(m+1)}(v_k)^m F\left(-m,-m,\frac{3}{2};\frac{\lambda_k}{4v_k}\right) \tag{51}$$

where $F(a,b,c;x)$ denotes the Gaussian hypergeometric function [16, eq. 9.100]. Finally, substituting $p\left(Y(\omega_k)\right) = (\lambda_k/2)\exp(v_k)$ in (47), we get

$$\int_0^\infty e^{x_k} p\left(x_k|Y(\omega_k)\right) dx_k = \exp(-v_k)\sum_{m=0}^\infty \frac{1}{m!}(v_k)^m F\left(-m,-m,\frac{1}{2};\frac{\lambda_k}{4v_k}\right) + \exp(-v_k)\sqrt{\lambda_k}\sum_{m=0}^\infty \frac{\Gamma(m+1.5)}{m!\Gamma(m+1)}(v_k)^m F\left(-m,-m,\frac{3}{2};\frac{\lambda_k}{4v_k}\right). \tag{52}$$

Using $\sqrt{\lambda_k} = \sqrt{v_k}\, Y_k/\gamma_k$, we can express the above equation as

$$\int_0^\infty e^{x_k} p\left(x_k|Y(\omega_k)\right) dx_k = \exp(-v_k)\sum_{m=0}^\infty \frac{1}{m!}(v_k)^m F\left(-m,-m,\frac{1}{2};\frac{Y_k^2}{4\gamma_k^2}\right) + \exp(-v_k)\frac{\sqrt{v_k}}{\gamma_k}Y_k \times \sum_{m=0}^\infty \frac{\Gamma(m+1.5)}{m!\Gamma(m+1)}(v_k)^m F\left(-m,-m,\frac{3}{2};\frac{Y_k^2}{4\gamma_k^2}\right). \tag{53}$$

The Gaussian hypergeometric infinite series $F(a,b,c;x)$ is known to converge if $Y_k^2/(4\gamma_k^2) < 1$ or equivalently if $Y_k^2 > \lambda_d^2(k)/4$. Simulation results indicated that this condition was rarely violated even at extremely low SNR conditions.

ACKNOWLEDGMENT

Many thanks go to S. Rangachari for all his help with the listening tests. The author would also like to thank the reviewers

for their useful comments that helped improved the present manuscript.

REFERENCES

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

[2] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. , Speech, Signal Processing*, vol. ASSP-23, no. 2, pp. 443–445, 1985.

[3] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proc. Int. Conf. Speech, Acoustics, Signal Processing*, vol. I, 2002, pp. 253–256.

[4] P. Wolfe and S. Godsill, "Toward a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2000, pp. 821–824.

[5] ——, "A perceptually balanced loss function for short-time spectral amplitude estimation," in *Proc. IEEE Int. Conf. Acoust. , Speech, Signal Processing*, vol. V, 2003, pp. 425–428.

[6] C. You, S. Koh, and S. Rahardja, "Adaptive b-order MMSE estimation for speech enhancement," in *Proc. IEEE Int. Conf. Acoust. , Speech, Signal Processing*, vol. I, 2003, pp. 900–903.

[7] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 346–349, 1994.

[8] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[9] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380–391, 1976.

[10] R. Gray, A. Buzo, A. Gray, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 367–376, 1980.

[11] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, 1975.

[12] N. Nocerino, F. K. Soong, L. Rabiner, and D. Klatt, "Comparative study of several distortion measures for speech recognition," *Speech Commun.*, vol. 4, no. 4, pp. 317–331, 1985.

[13] P. Kroon and B. Atal, "Predictive coding of speech using analysis-by-synthesis techniques," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi, Eds. New York: Marcel-Dekker, 1992, pp. 141–164.

[14] B. Atal and M. Schroeder, "Predictive coding of speech and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, 1979.

[15] M. Schroeder, B. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, 1979.

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

[17] D. Middleton, *An Introduction to Statistical Communication Theory*. New York: IEEE Press, 1996.

[18] F. Itakura and S. Saito, "An analysis-synthesis telephony based on maximum likelihood method," in *Proc. 6th Int. Conf. Acoustics*, 1968, pp. 17–20.

[19] K. Shikano and M. Sugiyama, "Evaluation of LPC spectral matching measures for spoken word recognition," *Trans. IECE*, vol. 565-D, no. 5, pp. 535–541, 1982.

[20] R. Brent, *Algorithms for Minimization without Derivatives*. Upper Saddle River, NJ: Prentice-Hall, 1973.

[21] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.

[22] Y. Hu and P. Loizou, "Incorporating a psychoacoustical model in frequency domain speech enhancement," *IEEE Signal Processing Lett.*, vol. 11, pp. 270–273, 2004.

[23] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 126–137, 1999.

Philipos C. Loizou (S'90–M'91–SM'04) received the B.S., M.S., and Ph.D. degrees, all 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 at Little Rock from 1996 to 1999. He is now a Professor in the Department of Electrical Engineering at the University of Texas at Dallas. His research interests are in the areas of signal processing, speech processing, and cochlear implants.

Dr. Loizou was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1999–2002) and is currently a member of the Industrial Technology Track Technical Committee of the IEEE Signal Processing Society.