A PERCEPTUALLY BALANCED LOSS FUNCTION FOR SHORT-TIME SPECTRAL AMPLITUDE ESTIMATION

Patrick J. Wolfe and Simon J. Godsill

Signal Processing Group, University of Cambridge
Department of Engineering, Trumpington Street
CB2 1PZ, Cambridge, UK
{pjw47, sjg}@eng.cam.ac.uk
http://www-sigproc.eng.cam.ac.uk

ABSTRACT

Here we present a novel approach to audio signal enhancement based on psychoacoustic principles. Specifically, we describe a short-time spectral amplitude estimator whose form comprises a weighted sum of the minimum mean-square error solution and the observed spectral value, where the weighting factor is given by the ratio of the masked threshold and this observed value. We then explore the connection between our approach and the idea of so-called balanced loss functions in statistics, showing the former to be an instance of the latter with a very special choice of weighting factor. Lastly, we present results indicating the relative merits of our approach in both objective and subjective terms, as compared to standard minimum mean-square error estimation under the assumed model. Software and sound examples are available at http://www-sigproc.eng.cam.ac.uk/∼pjw47.

1. INTRODUCTION

1.1. Short-Time Spectral Attenuation

Here we consider the task of noise reduction within the standard engineering framework of *short-time spectral attenuation*. In this method a time-varying filter is applied to the frequency-domain transform of a noisy signal, using the overlap-add method of short-time Fourier analysis and synthesis. The observed signal is first divided into overlapping segments via multiplication by a smooth, 'sliding' window function (which is non-zero only for a duration on the order of tens of milliseconds); the Fourier transform is then taken on each interval. Plotted side by side, the resultant spectra comprise a time-frequency representation known as the Gabor transform, or subsampled short-time Fourier transform—the modulus of which is the well-known spectrogram. The coefficients of this transform are attenuated to some degree in order to reduce the noise; individual short-time intervals are then inverse-transformed, multiplied by a smoothing window, and added together in an appropriate manner to reconstruct an estimate of the original signal.

1.2. Signal Model

We assume that the observed audio time series $\mathbf{y}$ may be modelled as the sum of an underlying signal $\mathbf{x}$ and a white, Gaussian noise process of zero mean $\mathbf{d}$. The goal is thus to reduce the noise level of the observed data $\mathbf{y} = \mathbf{x} + \mathbf{d}$; i.e., to provide an estimate of $\mathbf{x}$.

Within the framework of short-time spectral attenuation, the perceptual importance of spectral amplitude relative to phase [1] has led researchers to re-cast the resultant spectral estimation problem in terms of the former quantity (see [2] and references therein). In particular, Ephraim and Malah [2] derive an MMSE short-time spectral amplitude estimator under the assumption that the Gabor coefficients of the original signal as well as the noise may be modelled as statistically independent, zero-mean, Gaussian random variables. Thus the k-th observed spectral component, $Y_k \triangleq R_k \exp(j\vartheta_k)$, is equal to the sum of the spectral components of the signal, $X_k \triangleq A_k \exp(j\alpha_k)$, and the noise, D_k.

This model leads to a Rayleigh distribution of spectral amplitudes a_k and a uniform distribution of phases α_k over $[0, 2\pi)$. Keeping the notation of [2] for convenience, $\lambda_x(k) \triangleq E\left[\,|X_k|^2\,\right]$ and $\lambda_d(k) \triangleq E\left[\,|D_k|^2\,\right]$ denote the respective variances of the k-th component of the signal and noise, and a_k is a realisation of the random variable A_k to be estimated. Additionally, define

$$\frac{1}{\lambda(k)} \triangleq \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)}$$

and

$$\upsilon_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k; \quad \xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \quad \gamma_k \triangleq \frac{R_k^2}{\lambda_d(k)}, \tag{1}$$

where ξ_k and γ_k are interpreted after [2] as the a priori and a posteriori signal-to-noise ratios (SNR), respectively.

1.3. Optimal Estimation

Under the assumed model, the posterior density $p(a_k \,|\, Y_k)$ (following integration w.r.t. the phase term α_k) is Rician [3] with parameters $\sigma_k^2 \triangleq \lambda(k)/2$ and $s_k^2 \triangleq \upsilon_k \lambda(k)$:

$$p(a_k \,|\, Y_k) = \frac{a_k}{\sigma_k^2} \exp\left(-\frac{a_k^2 + s_k^2}{2\sigma_k^2}\right) I_0\left(\frac{a_k s_k}{\sigma_k^2}\right), \tag{2}$$

where $I_n(\cdot)$ denotes the modified Bessel function of order n.

The MMSE solution of [2] is simply the first moment of (2); when combined with the optimal phase estimator (the observed phase ϑ_k [2]), it takes the form of a suppression rule, or real-valued gain H_k applied to the observed spectral amplitude R_k:

$$\widehat{A}_k = \lambda(k)^{\frac{1}{2}}\, \Gamma(1.5)\, \Phi(-0.5, 1; -\upsilon_k) \tag{3}$$

$$\Rightarrow H_k = \frac{\sqrt{\pi \upsilon_k}}{2\gamma_k}\left[(1+\upsilon_k) I_0\left(\frac{\upsilon_k}{2}\right) + \upsilon_k I_1\left(\frac{\upsilon_k}{2}\right)\right] e^{\frac{-\upsilon_k}{2}}, \tag{4}$$

where $\Gamma(\cdot)$ is the gamma function [4, eq. 8.310.1] and $\Phi(\cdot)$ is the confluent hypergeometric function [4, eq. 9.210.1].

Of course, this and other optimal estimators may also be obtained directly in terms of the expected loss $E[L(a_k,\hat{a}_k)\,|\,Y_k]$ with respect to some posterior density $p(a_k\,|\,Y_k)$, as we now proceed to detail.

2. A PERCEPTUALLY MOTIVATED ESTIMATOR

It is well-known that in audio signal processing applications, careless processing via short-time spectral attenuation may lead to unacceptable levels of signal distortion (see, e.g., [5, Chapter 6]). In an attempt to remedy this shortcoming, we pursue here an approach based on auditory perception. That is to say, we consider a nonnegative masked threshold m_k for each point on the time-frequency lattice defined implicitly by the parameters of the overlap-add procedure (window type and support, amount of overlap, and so on), below which additive noise is taken to be inaudible.

A natural way of incorporating this masked threshold into the restoration procedure is via the Bayesian approach hinted at earlier, in which a loss function incorporating perceptual criteria may be employed, along with a prior signal model such as that given in Section 1.2, to determine the optimal (i.e., minimum-loss) Bayes estimator for the quantity of interest [6].

2.1. An Approach Based on the 'Principle of Least Processing'

An important point to note is that any processing we choose to apply is likely to lead to a dulling, or loss of 'naturalness', to the sound. This is especially true in high-fidelity applications such as the restoration of degraded audio recordings [5]. We therefore wish to proceed according to what one might term the 'principle of least processing'; i.e., that it is preferable to subject the signal to the least amount of processing necessary.

In the context of spectral estimation, we thus seek a loss function having the characteristic that it penalises *any* processing to some extent, thereby minimising the amount of attenuation applied (while, of course, still effecting a reduction in the noise level!). One way to accomplish this is through the consideration of perceptual criteria; as noise reduction inevitably occurs at the expense of signal resolution, why not take advantage of human auditory perception in order to optimise this trade-off?

In keeping with this principle, we may—for starters—choose not to attenuate the coefficient at any lattice point whose observed spectral energy is deemed to be masked a priori. However, this constitutes merely a *qualitative* incorporation of auditory perception into the restoration process. Additionally, we may also consider the *quantitative* integration of perceptual criteria into the estimation process, by way of the (a priori unknown) masked threshold m_k. Accordingly, we may formulate a perceptually motivated loss function which is also consistent with the principle of least processing as follows:

$$L(a_k,\hat{a}_k) = \left(\hat{a}_k - \left(1-\frac{m_k}{R_k}\right)a_k - m_k\right)^2 \qquad (5)$$

Intuitively, (5) may be understood in the following manner: when the *relative* masking level m_k/R_k is small, (5) is similar to the standard squared-error loss function, but with an extra term corresponding to the masked threshold. On the other hand, as $m_k/R_k \to 1$, the resultant estimator tends toward the masked threshold m_k, reflecting our intuition that any further attenuation violates the least processing principle. (Note that by processing only those points whose observed spectral energy is not masked, we are effectively imposing the constraint $m_k/R_k \leq 1$.)

We may now obtain the optimal Bayes estimator under the given loss function of (5) and the assumed model of Section 1.2, as detailed in Appendix A, in order to yield a solution of the form

$$\widehat{A}_k = \left(1-\frac{m_k}{R_k}\right)\hat{a}_k^* + m_k \Rightarrow H_k = \left(1-\frac{m_k}{R_k}\right)H_k^* + \frac{m_k}{R_k},$$

where $\hat{a}_k^*$ is given by (3) and H_k^* by (4).

2.2. Balanced Loss Functions: An Alternative Interpretation

An interesting connection can be made with so-called *balanced loss functions* [7], where the objective is to consider both loss due to estimation (i.e., shrinkage) and prediction (i.e., goodness-of-fit). In the case at hand, the form of such a function is as follows:

$$L(a_k,\hat{a}_k) = (1-w)(\hat{a}_k - a_k)^2 + w(\hat{a}_k - R_k)^2, \qquad (6)$$

where $0 \leq w \leq 1$ is a weighting factor determined by the user.

Specifically, consider the weighting factor as m_k/R_k; i.e., as the relative masking level. In this case *the estimator resulting from the balanced loss function of* (6) *is identical to that given by the least-processing loss function of* (5). This may be verified according to Appendix A, as differentiation identifies functions differing by a constant.

In a similar vein, perceptually motivated subspace methods for speech enhancement such as [8–10] seek a compromise between noise reduction and resultant signal distortion. Our solution is also related to the multiple-hypothesis method of [11], where a given Gabor coefficient is assumed either to be masked or not, and the resultant estimator is accordingly a weighted sum of the observed spectral magnitude and some other spectral estimator which takes into account the uncertainty of speech presence.

3. RESULTS AND CONCLUSIONS

3.1. Resultant Suppression Rules

It is instructive to compare the suppression rules resulting from (5) with those of [6], in which a zero-loss region is determined according to the masked threshold for a given spectral observation. To this end, Figs. 1 and 3 compare these suppression rules as a function of instantaneous SNR $\gamma_k - 1$ and a priori SNR ξ_k, and Figs. 2 and 4 compare them under the constraint $\gamma_k - 1 = \xi_k$, in a manner similar to simpler methods such as spectral subtraction and Wiener filtering. (Note that in all cases the $m_k = 0$ solution corresponds to the MMSE suppression rule.)

An inspection indicates that the perceptually balanced suppression rules shown in Figs. 1 and 2 do indeed induce less attenuation than those of Fig. 3 and 4 for a given relative masking level, in a manner consistent with the principle of least processing. This formulation thus provides suppression rules with the same qualitative trends as those in [6], but without the necessity of storage of tabulated gain values, since the solution is obtainable in closed form. Moreover, a similar derivation in terms of the *second* posterior moment—i.e., optimal spectral power estimation—yields a solution consisting entirely of simple functions and mathematical operations [12], thereby providing an ideal solution for on-line applications where speed is of the essence.

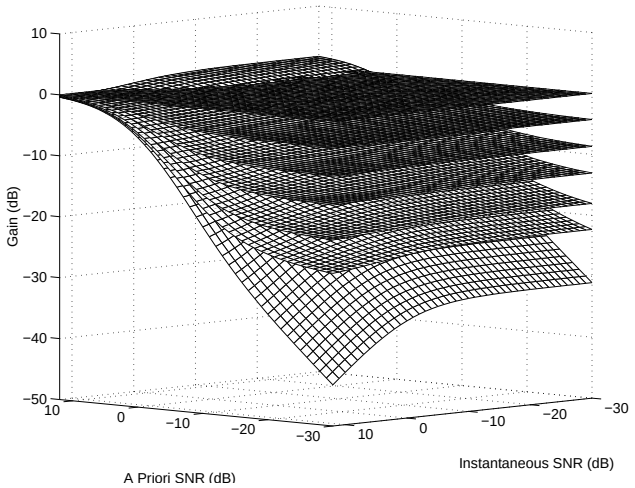

Fig. 1. Parametric suppression rules resulting from the minimisation of expected loss using (5) for relative masking levels $m_k/R_k \in \{0, 0.05, 0.1, 0.2, 0.35, 0.6, 1\}$

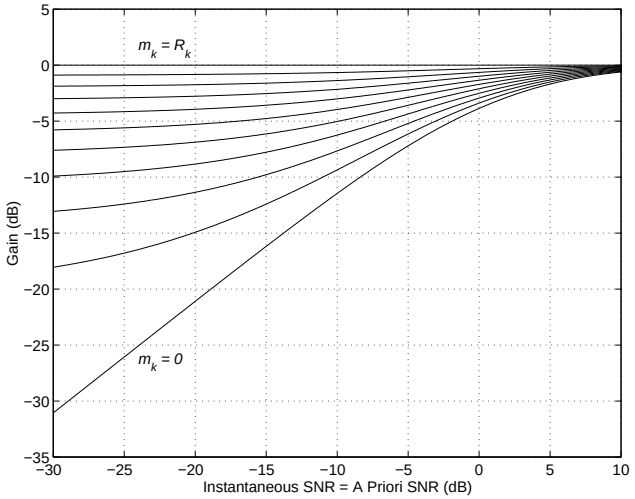

Fig. 2. Parametric suppression rules resulting from the minimisation of expected loss using (5) for relative masking levels $m_k/R_k \in \{0, 0.1, \ldots, 1\}$, with $\gamma_k - 1 = \xi_k$

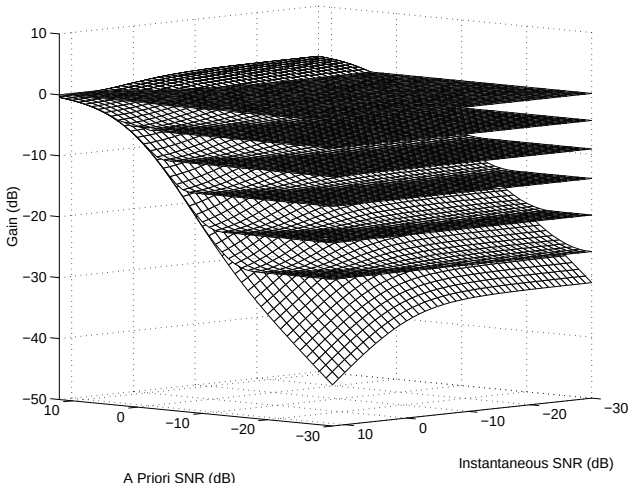

Fig. 3. Parametric suppression rules resulting from the minimisation of expected loss according to [6] for relative masking levels $m_k/R_k \in \{0, 0.05, 0.1, 0.2, 0.35, 0.6, 1\}$

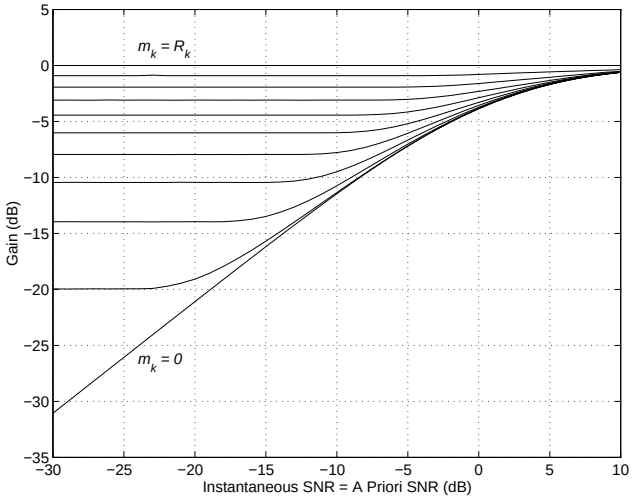

Fig. 4. Parametric suppression rules resulting from the minimisation of expected loss according to [6] for relative masking levels $m_k/R_k \in \{0, 0.1, \ldots, 1\}$, with $\gamma_k - 1 = \xi_k$

3.2. Enhancement Performance

In the interest of brevity we limit our quantitative performance comparison here to the MMSE suppression rule of Ephraim and Malah. In evaluating the performance of our perceptually balanced suppression rule, we considered a variety of speech and music signals corrupted by broadband noise, the variance of which was assumed to be known. Here we report results for broadband male speech and a solo piano recording—these being typical of the results obtained across a range of examples—artificially degraded with Gaussian noise to yield an SNR between -10 and 30 dB.

The signals were analysed using window lengths of 512 and 2048 samples, respectively, corresponding to durations of approximately 12 and 45 ms, and a redundancy factor of two, corresponding to a 50% window overlap. In keeping with the standard approach in the literature, we first obtained an estimate of the relative masking level via the model of [13], applied to an estimate of the original signal obtained according to the MMSE suppression rule.

Figure 5 shows a comparison of the objective restoration quality measured in terms of SNR gain, from which it can be seen the the perceptually balanced loss function approach often outperforms the MMSE suppression rule. In terms of subjective enhancement quality, we conclude that the restorations obtained using the perceptually balanced suppression rule are more natural and less dull-sounding than those resulting from the MMSE suppression rule. Lastly, we note that software (for the repro-

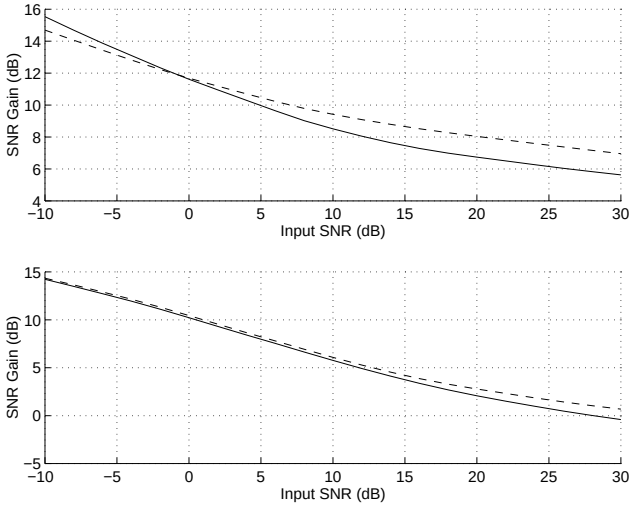

Fig. 5. Resultant SNR gain for a music (top) and speech (bottom) example (note difference in ordinate scales). The MMSE restoration performance is indicated by the solid line, and that according to the perceptually balanced loss function of (5) by the dashed line.

duction of these results as well as further experimentation) and sound examples (including real-world recordings) are available at `http://www-sigproc.eng.cam.ac.uk/~pjw47`.

A. ESTIMATOR DERIVATION

The key relation to note is that given by [4, p. 737, eqn. 6.631.1], rewritten here according to the relation $I_0(z) = J_0(jz)$:

$$\int_0^\infty x^\mu e^{-\alpha x^2} I_0(\beta x)\, dx = \frac{\Gamma(\kappa)}{2\alpha^\kappa}\Phi\left(\kappa, 1; \frac{\beta^2}{4\alpha}\right), \qquad (7)$$

where

$$\kappa = \frac{\mu+1}{2}; \quad \mathrm{Re}\,\alpha > 0, \quad \mathrm{Re}[\mu+\nu] > -1.$$

Note that the form of the integrand in (7) is that of a Rician density multiplied by some additional power. This is, of course, one particular example of the integral we seek to minimise in order to obtain the optimal Bayes estimator with respect to a given loss function and density. Thus, with regard to the case at hand, consider situations in which the loss function of interest is quadratic—say $Q(a)$—yielding an expression of the form

$$\int Q(a)\, a \exp\left(-\frac{a^2}{\lambda}\right) I_0\left(2a\sqrt{\frac{\nu}{\lambda}}\right) da. \qquad (8)$$

Specifically, if $L(a,\hat{a}) = Aa^2 + Ba + C$, then the solution of (8), according to (7), is

$$A\frac{\lambda^2}{2}\Phi(2,1;\nu) + B\frac{\Gamma\left(\frac{3}{2}\right)\lambda^{\frac{3}{2}}}{2}\Phi\left(\frac{3}{2},1;\nu\right) + C\frac{\lambda}{2}\Phi(1,1;\nu).$$

However, since $\Phi(\alpha,\alpha;z) = e^z$ and $\Phi(\alpha,\gamma;z) = e^z\Phi(\gamma-\alpha,\gamma;-z)$ by [4, p. 1086, eq. 9.212.1], this form may be reduced to the following:

$$\frac{A\lambda^2}{2}\Phi(2,1;\nu) + \frac{\lambda e^\nu}{2}\left[B\Gamma\left(\frac{3}{2}\right)\lambda^{\frac{1}{2}}\Phi\left(-\frac{1}{2},1;-\nu\right) + C\right].$$

If A is not a function of $\hat{a}$, then differentiating the above w.r.t. $\hat{a}$ implies that its critical points satisfy

$$\frac{d}{d\hat{a}}C(\hat{a}) = -\frac{d}{d\hat{a}}B(\hat{a})\,\Gamma\left(\frac{3}{2}\right)\lambda^{\frac{1}{2}}\Phi\left(-\frac{1}{2},1;-\nu\right), \qquad (9)$$

where we note that the latter terms on the right-hand side of (9) comprise the MMSE spectral amplitude estimator of (3) under the assumed model. Thus for the quadratic loss functions considered herein, we obtain the weighted solution described in Section 2.

B. REFERENCES

[1] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 4, pp. 679–681, 1982.

[2] 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.

[3] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell Syst. Tech. J.*, vol. 27, pp. 109–157, 1948.

[4] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, fifth edition, 1994.

[5] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*, Springer-Verlag, Berlin, 1998.

[6] P. J. Wolfe and S. J. Godsill, "Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement," in *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2000, vol. 2, pp. 821–824.

[7] A. Zellner, "Bayesian and non-Bayesian estimation using balanced loss functions," in *Statistical Decision Theory and Related Topics V*, S. S. Gupta and J. O. Berger, Eds., New York, 1994, pp. 377–390, Springer-Verlag.

[8] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[9] F. Jabloun and B. Champagne, "A perceptual signal subspace approach for speech enhancement in colored noise," in *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2002, vol. 1, pp. 569–572.

[10] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2002, vol. 1, pp. 537–540.

[11] A. Akbari Azirani, R. le Bouquin Jeannès, and G. Faucon, "Optimizing speech enhancement by exploiting masking properties of the human ear," in *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Processing (ICASSP)*, 1995, vol. 1, pp. 800–803.

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

[13] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Area Commun.*, vol. 6, no. 2, pp. 314–323, 1988.