Journal of New Music Research
2001, Vol. 30, No. 1, pp. 83–92

0929-8215/01/3001-083$16.00
© Swets & Zeitlinger

Perceptually Motivated Approaches to Music Restoration

Patrick J. Wolfe and Simon J. Godsill

University of Cambridge, Signal Processing Group, Department of Engineering, Cambridge, UK

Abstract

Spurred by the success of perceptual models in audio coding applications, researchers have recently begun to address audio signal enhancement in a similar manner. Here we consider the case of musical recordings degraded by additive broadband noise such as tape hiss, in which the prevention of signal distortion is tantamount to noise removal. We review perceptually motivated approaches to music restoration and describe a statistical model based framework we have recently proposed. By integrating psychoacoustics into the restoration process through the use of perceptual optimality criteria, our method aims to take advantage of human auditory perception to yield improvements in both noise reduction and perceived signal fidelity. Audio examples and related software may be found at http://www.sigproc.eng.cam.ac.uk/~pjw47.

1 Introduction

The advent of digital signal processing has promoted speech and audio enhancement methods beyond the simple, fixed, frequency-domain filters of the past towards more powerful and flexible techniques. Although research in audio signal enhancement has traditionally been driven by speech enhancement needs for military and civilian communications under low- or medium-fidelity conditions, the area of high-fidelity audio restoration is at least as challenging. Moreover, its range of application spans the entire history of sound recording. Consider, for example, a musicologist studying early recordings that may be degraded almost to the point of uselessness, or the amateur musician of today with a home recording studio, who may be without the knowledge or equipment necessary to produce a noise-free recording.

It is with such practical and artistic motivations that researchers have sought perceptually motivated means to restore musical recordings. Although most techniques developed for speech enhancement are in principle applicable to audio restoration, differences in the criteria governing these tasks dictate a significant change of approach (Godsill & Rayner, 1998, Ch. 6). Whereas the ultimate goal of speech enhancement is to increase intelligibility through noise removal, audio restoration must also address the inevitable signal distortion incurred at its expense; in fact, this distortion is often the limiting factor in musical applications (Cappé & Laroche, 1995). In addition to the stringent fidelity requirements of audio restoration, dissimilarities in the time and frequency characteristics of speech and music signals magnify the distinction.

As perceptually motivated methods to date have been largely limited to the removal of additive broadband noise such as tape hiss, we restrict our discussion to this aspect of audio restoration. We first address a popular class of audio signal enhancement methods known collectively as short-time spectral attenuation. We follow with a review of recent methods that incorporate perceptual criteria, usually auditory masking, in an attempt to improve upon standard mathematical approaches. We emphasise techniques applicable to the restoration of musical recordings; however, as many noise reduction techniques were originally developed for speech enhancement, we discuss their relevance to music restoration as well. We then introduce our own method, which employs perceptual optimality criteria in a statistical model based framework, and conclude with some general observations on the performance of these techniques.

Complex numbers, such as those of the discrete Fourier transform of a signal, appear in bold throughout. The $\hat{\cdot}$ symbol denotes an estimate; e.g., $\hat{x}$ is an estimate of x.

1.1 Standard broadband noise reduction techniques

Background noise is common to all analogue systems, and as it is localised in neither time nor frequency its removal is one of the most ubiquitous and difficult signal enhancement

Accepted for publication: January 18, 2001

Correspondence: Patrick J. Wolfe, University of Cambridge, Department of Engineering, Signal Processing Group, Trumpington Street, CB2 1PZ, Cambridge, UK. Tel.: +44 1223 332 767; Fax: +44 1223 332 662; E-mail: pjw47@eng.cam.ac.uk

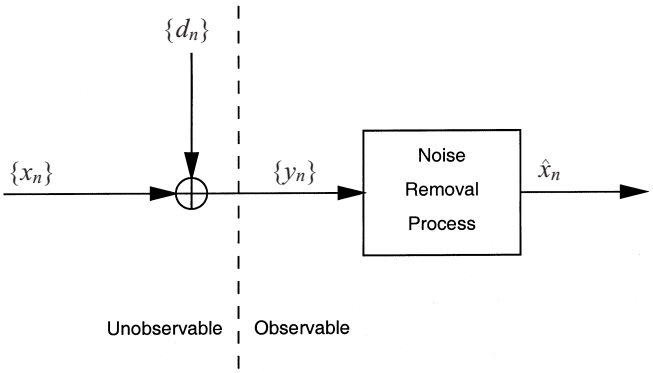

Fig. 1. Noise removal for music restoration.

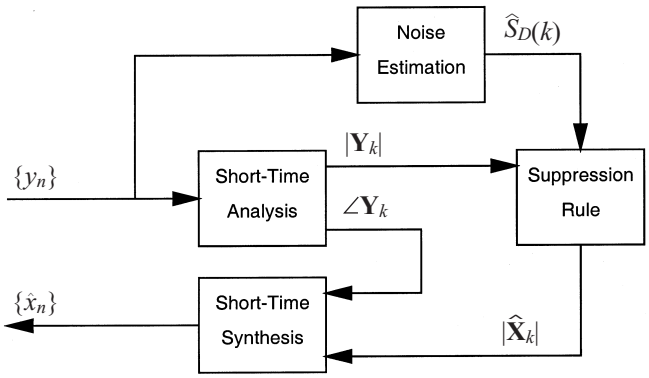

Fig. 2. Short-time spectral attenuation.

tasks. Here we assume that such noise is additive in the time domain and random; one example is noise introduced into an audio recording by the electrical circuits employed in the recording process. Furthermore, we assume a sampled and quantised analogue signal which we process by means of a digital filter. From a statistical point of view, we are given a sequence of sampled observations of an underlying process (in our case the original musical performance) that has been corrupted by additive random noise. Let $\{x_n\} \triangleq \{x(nT)\}$ in general represent a set of values from a finite-duration analogue signal sampled at a regular interval of T; a noisy data sequence observed at time n may then be represented by the additive model

$$y_n = x_n + d_n,$$

where y_n is the observed signal, x_n is the original signal, and d_n is statistically independent random noise. Our goal is then to make an estimate $\hat{x}_n$ of the original signal x_n at time n, based on some subset of the observations $\{y_n\}$, as shown in Figure 1.

The most popular methods of broadband noise reduction to date involve the application of a time-varying filter to the frequency-domain transform of a noisy signal. Because many audio signals of interest (e.g., music and voiced speech) are composed of spectral components corresponding to fundamental pitches and harmonics, the Fourier transform provides a useful tool for analysis. However, while the squared magnitude of the Fourier transform shows the distribution of signal energy with respect to frequency, it does not tell us anything about the relative timing of different frequency components (as that information is hidden in the phase of the complex Fourier spectrum). One solution to this problem is to assume the frequency content to be relatively constant over some (usually short) time interval, and then to Fourier transform each interval. Plotting the resultant spectra side by side, we arrive at a time-frequency representation of a signal not unlike a musical score.

Often it may be necessary to process the signal of interest in real time, in which case the overlap-add method of short-time Fourier analysis and synthesis is often used. Here the signal is analysed using the discrete Fourier transform as described above, after being divided into overlapping intervals by application of a window function. Following modification of the short-time spectrum, individual blocks are inverse-transformed, windowed, added together, and then scaled to account for the effects of pre- and post-windowing.[1] Modification of the noisy signal in this manner is known as short-time spectral attenuation (STSA), and is equivalent to the application of a nonnegative real-valued gain H_k to each frequency bin k of the observed short-time spectrum $\mathbf{Y}_k$, in order to form an estimate $\hat{\mathbf{X}}_k$ of the original spectrum $\mathbf{X}_k$:

$$\hat{\mathbf{X}}_k = H_k \cdot \mathbf{Y}_k.$$

The formula governing H_k is known in the literature as a noise suppression rule, and it depends in general on the power spectra of the signal and the noise, $S_X(\omega)$ and $S_D(\omega)$, respectively.[2] Figure 2 shows the sequence of STSA events in the form of a block diagram.

Intuitively, one straightforward approach to noise reduction is simply to subtract the estimated noise magnitude or power spectrum from the observed spectrum, leaving the phase unchanged. Indeed, the two most basic suppression rules are based on this concept – magnitude spectral subtraction (Boll, 1979):

$$|\hat{\mathbf{X}}_k| = \begin{cases} |\mathbf{Y}_k| - |\hat{\mathbf{D}}_k| & \text{if } |\mathbf{Y}_k| > |\hat{\mathbf{D}}_k|, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

and power spectral subtraction (Berouti et al., 1979):

$$|\hat{\mathbf{X}}_k|^2 = \begin{cases} \hat{S}_Y(k) - \alpha\hat{S}_D(k) & \\ & \text{if } \hat{S}_Y(k) - \alpha\hat{S}_D(k) > \beta\hat{S}_D(k), \\ \beta\hat{S}_D(k) & \text{otherwise.} \end{cases} \tag{2}$$

[1] For a Gabor analysis interpretation of the overlap-add and other time-frequency methods applied to musical signals, see the discussion elsewhere in this issue by Dörfler (2001).

[2] We proceed under the standard assumption that an estimate $\hat{S}_D(k)$ of the noise power spectrum is available; for instance, such an estimate may be taken from sections of $\{y_n\}$ consisting only of recorded silence.

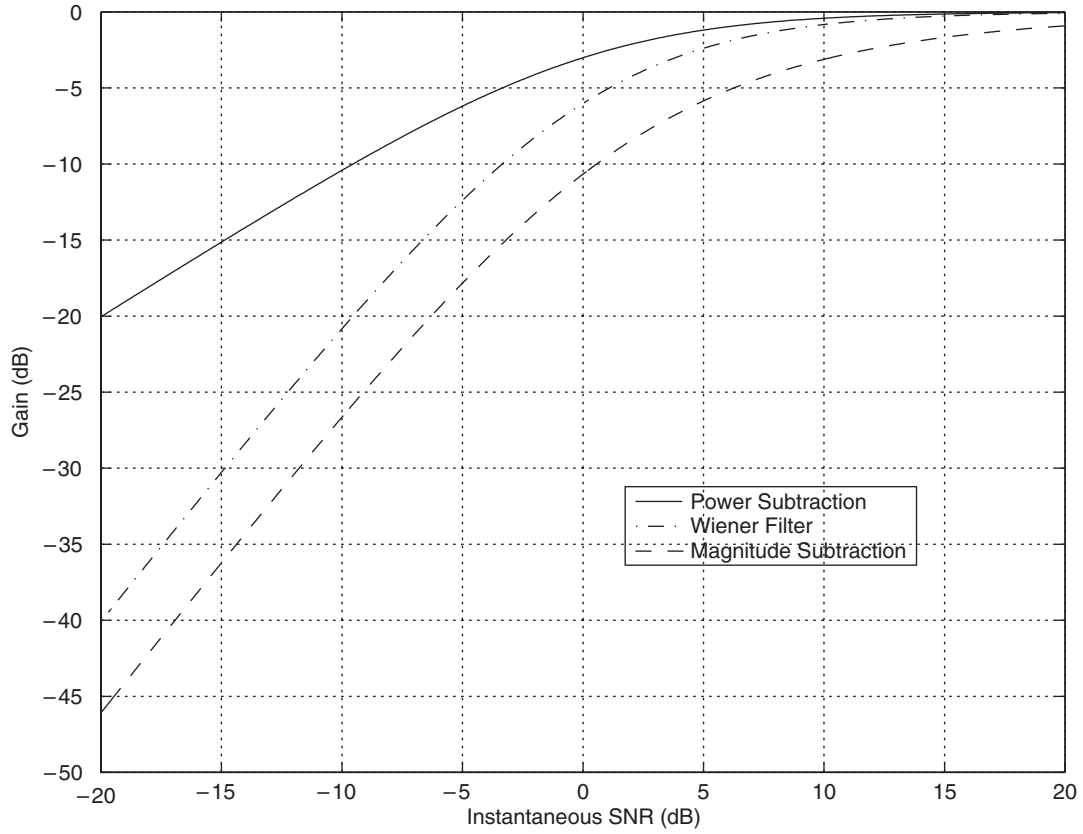

Fig. 3. Common short-time spectral attenuation suppression rules.

Parameters α and β in (2) may be adjusted to control the amount of noise reduction and the level of the remaining noise floor, respectively. Note that power subtraction requires an estimate $\hat{S}_Y(k)$ of the power spectrum of the observed signal. Following Berouti et al. (1979), a simple estimate thereof may be defined as

$$\hat{S}_Y(k) \triangleq \left|\hat{\mathbf{Y}}_k\right|^2. \tag{3}$$

Another standard technique is to approximate the Wiener filter (Wiener, 1949), a filter that minimises the mean-square error of the estimate's time domain reconstruction for the case of zero-mean, additive noise (see, e.g., Van Trees, 1968, pp. 198–219):

$$H_k = \frac{S_X(k)}{S_X(k)+S_D(k)}. \tag{4}$$

Although in general neither $S_X(k)$ nor $S_D(k)$ is known, an estimate $\hat{S}_X(k)$ of the signal power spectrum may be obtained in a manner analogous to (3) by applying the power subtraction method of (2) with $\alpha = 1$ and $\beta = 0$:

$$\hat{S}_X(k) \triangleq \begin{cases} \hat{S}_Y(k) - \hat{S}_D(k) & \text{if } \hat{S}_Y(k) - \hat{S}_D(k) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Figure 3 shows a comparison of the standard suppression rules given by (1), (2), and (4). Intuitively, a low input signal-to-noise ratio (SNR) – corresponding to a greater degree of signal degradation – effects a greater degree of spectral attenuation than a high SNR.

We pause now to consider one additional model, as it provides the basis for a noise reduction method we have recently proposed (Wolfe & Godsill, 2000). Ephraim and Malah (1984) derive a minimum mean-square error (MMSE) spectral amplitude estimator under the assumption that the Fourier expansion coefficients of the original signal may be modelled as statistically independent, zero-mean, Gaussian random variables. For the case of additive, zero-mean, white Gaussian noise, the resultant suppression rule is

$$H_k \frac{\sqrt{\pi \upsilon_k}}{2\gamma_k}\left[(1+\upsilon_k) I_0\left(\frac{\upsilon_k}{2}\right) + \upsilon_k I_1\left(\frac{\upsilon_k}{2}\right)\right] \exp\left(\frac{-\upsilon_k}{2}\right), \tag{6}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of order zero and one, respectively (Gradshteyn & Ryzhik, 1994, eq. 8.431.3), and

$$\upsilon_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k.$$

The ratio of the variances of the Fourier expansion coefficients of the underlying signal and the noise is given by ξ_k. Since both are assumed to be Gaussian, ξ_k is interpreted as the *a priori* (i.e., the true but unobservable) SNR. McAulay

and Malpass (1980), and Ephraim and Malah (1984), define the *a posteriori* SNR as

$$\gamma_k \triangleq \frac{\hat{S}_Y(k)}{\hat{S}_D(k)}.$$

The term *a posteriori* reflects the fact that this SNR is related to the noisy observation $\mathbf{Y}_k$ rather than the underlying signal $\mathbf{X}_k$. Although not immediately obvious upon inspection of (6), ξ_k is in fact the dominant parameter, with γ_k functioning as a corrective factor when ξ_k is low (Cappé, 1994); an approximation to this suppression rule yields a more intuitive interpretation (Wolfe & Godsill, 2001).

Note that the Wiener noise suppression rule may be expressed as a function of $(\gamma_k - 1)$, defined by Ephraim and Malah (1984) as instantaneous SNR:[3]

$$H_k = \begin{cases} \dfrac{\gamma_k - 1}{\gamma_k} & \text{if } \gamma_k - 1 > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Similarly, the basic power subtraction scheme of (2), with $\alpha = 1$ and $\beta = 0$, is equivalent to applying the square root of the Wiener gain of (7) to $\mathbf{Y}_k$.

1.2 Related aspects of auditory perception

We now briefly introduce those aspects of human auditory perception that play a central role in perceptually motivated approaches to music restoration. Physiologically, the peripheral auditory system comprises three parts: the outer, middle, and inner ear. The outer ear consists of the pinna (the outermost, visible part) and the auditory canal, and acts as a fixed acoustic filter to modify incoming sound. The middle ear provides a mechanism whereby an acoustic wave incident on the eardrum is transmitted to the inner ear (or cochlea), a fluid-filled tube wound into a spiral shape. The cochlea houses the basilar membrane, along which a pressure wave propagates following its transmission from the middle ear.

When excited by an acoustic stimulus, the basilar membrane reaches mechanical resonance at different places along its length as a function of the frequencies of the input wave. Thus, the cochlea behaves as a frequency analyser – a group of bandpass filters effecting frequency-to-place conversion. Auditory nerve fibres grouped along the basilar membrane exhibit a bandpass response threshold similar to its mechanical resonance, with each fibre showing a characteristic frequency at which it is most sensitive. The distribution of this neural auditory activity, taken as a function of characteristic frequency, is called an excitation pattern. Measured in decibels and plotted on a logarithmic frequency scale, excitation patterns may be considered an effective internal representation of an input spectrum (Moore, 1997, p. 37); their importance will be seen shortly.

Beginning with Fletcher (1940), researchers have sought to characterise auditory frequency selectivity by measuring the ability of subjects to detect a pure tone presented in a narrow frequency band of noise. Experimental results led Fletcher to hypothesise that the peripheral auditory system acts as a filter bank in which a listener makes use of a filter centred near the frequency of the stimulus; he denoted the width of such a filter as a critical band (Fletcher, 1940). Beginning with this simple rectangular approximation of the ideal auditory filter, researchers have worked towards more accurate models. One widely accepted concept of a rounded exponential auditory filter is due to Patterson and Nimmo-Smith (1980), and Patterson et al. (1982); however, it and other models are regarded simply as a weighting function operating on the power spectrum of the input signal. More recently, Irino and Patterson (1997) describe the gammachirp auditory filter model, which is a true filter in the sense of having a well-defined impulse response.

The idea of critical bands and the auditory filter as a means of signal detection is closely related to the concept of loudness. Here we take loudness to be the perceptual correlate of acoustic intensity, beginning at the threshold of hearing. Zwicker and Fastl (1999, pp. 220–233; see also references therein), Moore and Glasberg (1986), Stuart (1994), Moore and Glasberg (1996), and Moore et al. (1997) all propose predictive loudness models. Such models generally consist of four steps:

1. Time-invariant filtering according to a model of the outer and middle ear
2. Calculation of an excitation pattern
3. Transformation to loudness density per unit bandwidth
4. Integration of loudness density to produce a loudness value

Note that auditory loudness levels are not deterministic quantities. Hearing thresholds are defined by detection rate; for example, as being detectable 75% of the time in a forced-choice yes/no experiment (Moore et al., 1997). Individual responses may also differ above 1 kHz, with variation being greatest above 6 kHz (Moore, 1997, p. 51). Additionally, thresholds increase as a result of age (presbyacusis) as well as damage suffered through exposure to high sound levels. Thus, the most one may ever hope for is a general prediction of loudness in a statistical sense.

The special case of zero loudness in the presence of nonzero acoustic intensity is termed auditory masking; it corresponds to the phenomenon whereby certain sounds (signals) may be rendered partially or wholly inaudible in the presence of others (maskers). Insofar as this is the case, masking reflects the limits of auditory frequency selectivity (Moore, 1997, p. 89). It is thought to be due to a combination of swamping (Moore, 1997, pp. 117–118), in which the masker produces a level of neural activity sufficient to prevent the detection of

[3] Some authors, e.g., Cappé (1994), prefer a slightly different interpretation in which the instantaneous SNR is called the *a posteriori* SNR; that is, the *a posteriori* SNR is given by $\hat{S}_X(k)/\hat{S}_D(k)$, with $\hat{S}_X(k)$ as defined in (5).

the signal, and suppression (Moore & Glasberg, 1983a; Fastl & Bechly, 1983; Delgutte, 1990), in which neural activity at a given frequency suppresses that at nearby frequencies.

The most repeatable masking effects are those involving steady-state sounds, usually a band-limited noise masker and a tone signal. However, masking is also seen in sounds that change over time, both before and after the occurrence of a masker (see, e.g., Zwicker, 1976; Fastl, 1976, 1977, 1979). Backward masking, commonly known as premasking or prestimulatory masking, refers to the masking of a signal prior to masker onset, and is relatively poorly understood. On the other hand, forward masking, also known as postmasking or poststimulatory masking, refers to the continuation of masking effects following cessation of the masker. It is a much more repeatable phenomenon, and as such has been the subject of greater study (see, e.g., Kidd & Feth, 1982; Jesteadt et al., 1982; Moore & Glasberg, 1983a; Zwicker, 1984). Forward masking effects last on the order of 100–200 ms regardless of masker level; however, the amount of forward masking is known to increase with increasing masker duration, at least for durations up to 20 ms (Moore, 1997, p. 130).

The excitation patterns of tones and narrow bands of noise are known to exhibit a nonlinear growth towards higher frequencies, a phenomenon known as the upward spread of masking. Moore and Glasberg (1983b) explain this effect by considering the excitation level at a given frequency to be the output of the auditory filter centred at that frequency. An estimate of the excitation pattern of a masker may thus be obtained by convolving the auditory filter (whose width increases with frequency) with the masker spectrum (Patterson & Moore, 1986, p. 174). In fact, most models of masking approximate this convolution (see, e.g., Johnston, 1988).[4] In the case of simultaneous masking, a signal is detected when it evokes an excitation pattern that exceeds some constant proportion of that of the masker, generally about −4 dB (Moore, 1997, p. 110). For complex sounds (such as a tone in noise) loudness may be viewed as a measurement of partial masking, a case in which one spectral component does not completely mask another but rather causes a reduction in its perceived loudness.

2 Perceptually motivated noise reduction methods

Spurred by the success of perceptual models in audio coding applications, researchers have recently begun to address audio restoration in a similar manner. Although for the sake of clarity we divide perceptually motivated methods into the two distinct categories of audible noise reduction and perceptual transforms, this division is often blurred, with methods ranging from attempts to classify noise components as either masked or unmasked, to noise reduction schemes implemented entirely in a perceptual domain. As there have so far been few applications of these methods to music restoration rather than speech enhancement, we review the relevant algorithms from both areas.

2.1 Audible noise reduction

The most prevalent means of incorporating perceptual criteria into the noise removal process is via suppression rules that take into account auditory masking effects. Since the masked thresholds due to the underlying signal are not known, they must be estimated from the noisy signal. Usually this is done by first applying a basic STSA technique to yield an estimate of the original signal spectrum, for which masked thresholds are calculated according to the model of Johnston (1988). Most perceptually motivated suppression rules then use these estimated masked thresholds to adjust the parameters of the filtering operation. Perhaps the most basic use of auditory masking is in a qualitative attempt to mask time-varying residual noise by specifying a noise floor (see, e.g., Kim et al., 2000).

Czyzewski and Krolikowski (1999) describe a direct extension of the perceptual audio coding approach. By increasing the amplitude of those components classified as signal and decreasing the amplitude of those classified as noise, masked thresholds are manipulated in such as way as to suppress noise in the underlying signal. Tsoukalas et al. (1993) describe a method aimed at minimising the audible noise spectrum, defined as the difference between the audible power spectrum of the noisy and clean signals. They formulate an enhancement criterion by attempting to constrain the audible noise spectrum to be less than or equal to zero at all frequencies. Tsoukalas et al. (1997a) extend this approach to include the estimation of speech parameters per critical band directly from the noisy signal.

Ephraim and Malah (1985) propose a filter that minimises the mean-square error of the log-spectrum, on the basis of its relation to speech perception. Ephraim and Van Trees (1995a,b) describe a signal subspace approach in which signal distortion is minimised for a given residual noise spectrum criterion. Similarly Gustafsson et al. (1998) detail a method aimed at preserving a pre-defined amount of noise in the processed signal, and give a parameter which may be adjusted to balance this with the amount of resultant signal distortion. Virag (1999) also proposes a variation on spectral subtraction in which masked threshold estimates control the filter parameters in a search for the best perceptual compromise between noise reduction and speech distortion.

Azirani et al. (1995) describe a two-state probabilistic method: if the noise is considered to be in a masked state, the observed spectral value is left alone. Otherwise, it is subject to filtering as a function of the *a priori* SNR ξ_k. Furthermore, if this filter gain results in an estimator below the estimated masked threshold, the estimator is set to that threshold. This can be viewed as an application of what might be termed the principle of least processing: in order to mini-

[4] Such models take the same general form as those of loudness; indeed, loudness models include masking as a special case.

mise processing of the distorted signal (both for reasons of fidelity and efficiency), noisy spectral amplitudes below the estimated masked threshold are not processed, as they are assumed to be masked by the underlying signal.

2.2 Perceptual transforms

Another class of methods aims to transform the signal into an internal auditory representation (such as an excitation pattern) and then determine suppression rule filter coefficients in that domain. In an algorithm developed specifically for music restoration, Tsoukalas et al. (1997b) suggest a linear filter in the form of a Wiener filter, but with its power spectra replaced by their respective psychoacoustic representations based on a perceptual model. Canagarajah (1993, Ch. 2) details a similar scheme applied to speech enhancement, perceptual spectral subtraction, in which the filter gain H_k is given by $1 - E_D(k)/E_X(k)$, where $E_D(k)$ and $E_X(k)$ represent estimates of the internal auditory excitation patterns of the noise and the signal, respectively.

Canazza et al. (1999) propose another method in which a suppression rule is calculated after transforming the noisy signal to a psychoacoustic representation, the idea being to remove, within model fidelity, only the audible noise components from the signal. In this manner the spreading of masking in time and frequency is captured, resulting in a smoothed estimate of the underlying signal spectrum. Similarly, Lorber and Hoeldrich (1997) combine spectral attenuation with a psychoacoustic filter simulating the increase of masking bandwidth with increasing frequency. This filter is first applied to the noisy signal prior to estimating the *a posteriori* SNR γ_k, which in turn is used to evaluate a suppression rule.

In general, transforms for signal analysis and synthesis designed to mimic the human auditory system have thus far been applied to perceptual audio coding rather than noise reduction. However, Petersen and Boll (1981) detail spectral subtraction in a perceptual domain, using a transform approximating the critical bands of the ear. Similarly Gülzow et al. (1998) compare two types of perceptually motivated transforms for STSA-based speech enhancement. Most recently, Irino (1999) employs a gammachirp auditory filter for noise reduction in a perceptual domain.

2.3 Noise reduction based on perceptual optimality criteria

As we have seen, many researchers have recently focussed attention on the enhancement of degraded speech or audio signals through the use of perceptual criteria. However, rather than apply psychoacoustic principles heuristically, we prefer to integrate these criteria into the restoration process quantitatively. One way to do this, as we now detail, is through the use of cost functions that explicitly consider auditory perception.

Often, as in the case at hand, it is desirable to estimate the value of a random variable X (in our case, the underlying spectral amplitude) as a function of the observations of some related random variable Y (here, the observed spectral amplitude). By defining a model in terms of the probability distributions of these quantities, we may estimate X as a function of Y. Here we use the model of Ephraim and Malah (1984) underlying (6), in which the Fourier expansion coefficients of the original signal and the noise are modelled as statistically independent, zero-mean, Gaussian random variables.

Cost functions represent a standard technique for signal estimation in which one defines a nonnegative cost function $C(x, \hat{x})$ of x and its estimate $\hat{x}$, and minimises the risk R, defined as the average or expected cost. The Bayes estimate is that which minimises the risk with respect to $f_{X,Y}(x, y)$, the joint probability density function of X and Y:

$$\begin{aligned} R &\triangleq E[C(x,\hat{x})] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(x,\hat{x}) f_{X,Y}(x,y)\,dxdy \\ &= \int_{-\infty}^{\infty} f_Y(y)\int_{-\infty}^{\infty} C(x,\hat{x}) f_{X|Y}(x|y)\,dxdy \end{aligned} \tag{8}$$

Since $C(x, \hat{x})$ is nonnegative it is sufficient to minimise the inner integral of (8).

Although choice of a cost function is arbitrary, generally one is chosen to be a compromise between analytical tractability and an accurate reflection of the (possibly subjective) cost of an estimation error. As a first step towards a perceptual cost function (i.e., one that measures the subjective perceptual cost of an error) we incorporate auditory masking into an STSA noise reduction procedure by considering some masked threshold m_k for each bin k in a given short-time magnitude spectrum. We then formulate perceptual cost functions incorporating m_k, with which we derive suppression rules for noise reduction. Thus, we view the restoration process as a search for a best estimate of the underlying signal in which the estimation error is considered to be noise. Through or choice of cost function we attempt to mask this noise with the underlying signal, as is done in perceptual audio coding.[5]

The Bayes estimate corresponding to a quadratic cost function (i.e., when the cost is equal to the square of the error $\hat{x} - x$) is the MMSE estimate, which is also the mean of the posterior density $f_{X|Y}(x|y)$. We propose as a cost function a generalisation of the MMSE criterion to include a masked threshold, below which the cost of an estimation error is always zero (Wolfe & Godsill, 2000):

$$C(\mathbf{X}_k, \hat{\mathbf{X}}_k)\begin{cases} \left(|\hat{\mathbf{X}}_k| - |\mathbf{X}_k|\right)^2 - m_k^2 & \\ & \left|\,|\hat{\mathbf{X}}_k| - |\mathbf{X}_k|\,\right| > m_k, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

[5] Here we consider only spectral amplitude estimation. Although we do not necessarily discount the importance of phase in the restoration of music signals, in the absence of a quantative perceptual motivation we retain the original phase estimator for the model under consideration; i.e., the noisy spectral phase (Ephraim & Malah, 1984).

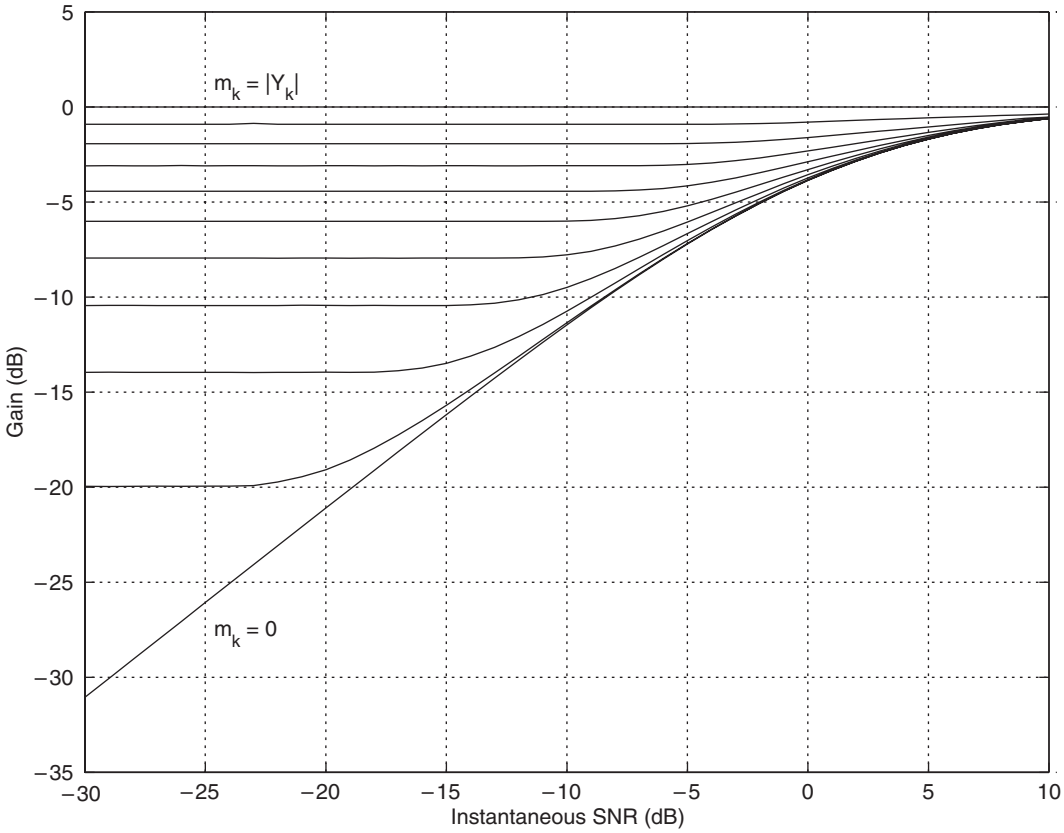

Fig. 4. Parametric suppression rules resulting from the incorporation of a masked threshold into the quadratic cost function, for relative masking levels $m_k/|\mathbf{Y}_k| \in \{0, 0.1, \ldots, 1\}$ (Wolfe & Godsill, 2000).

We then minimise the expected cost R given an observed spectral value $\mathbf{Y}_k$, using numerical methods. We formulate this result as a family of suppression rules, plotted in Figure 4 as a function of instantaneous SNR (with $\xi_k = \gamma_k - 1$) for different masking levels. For the special case of zero masking, (9) reduces to a quadratic cost function and hence the resultant estimator reduces to that of Ephraim and Malah (1984). For the general case, the resultant estimator may be seen in Figure 4 to approach the masked threshold value m_k asymptotically at low SNR, and the MMSE estimator of (6) at high SNR.

Although here we only briefly outline an extension of the MMSE optimality criterion to include a masked threshold, full details concerning this and other criteria are presented elsewhere (Wolfe & Godsill, 1999, 2000).

3 Discussion

As mentioned previously, a main reason for the success of STSA techniques is the suitability of a Fourier-based analysis for most audio signals of interest. However, basic STSA methods are not without their drawbacks. We now consider the general performance of such techniques, beginning with the problem of musical noise.

3.1 Musical noise

Owing to the random nature of broadband noise, the observed short-time magnitude spectrum $|\mathbf{Y}_k|$ may exhibit severe fluctuations from one short-time block to the next. As a result, basic STSA techniques generate a residual noise consisting of random, isolated, time-varying spectral components. (Intuitively, subtracting the average noise spectrum from a short-time block containing mostly noise energy will result in a series of spectral peaks at random locations.) This effect has come to be known as musical noise; it is highly undesirable in any event but especially so for the case of audio restoration.

Techniques for reducing musical noise include subtracting an overestimate of the noise spectrum, specifying a remaining noise floor to mask the musical noise, and using temporal information from surrounding short-time blocks. The power subtraction scheme of (2) allows for both subtraction of a noise power overestimate $\alpha\hat{S}_D(k)$ and the specification of a residual noise floor $\beta\hat{S}_D(k)$. Alternatively, a lower limit may be set on the gain of H_k (Cappé, 1994; Arslan et al., 1995). The use of temporal information from nearby short-time blocks leads to a more effective reduction of musical noise. Most notably, in an effect investigated by Cappé (1994), the decision-directed approach of Ephraim & Malah (1984) is free of musical noise. This is due to a

smoothing of the *a priori* SNR ξ_k, which is given by a geometric weighting of the instantaneous SNR from the previous and current short-time blocks. Arslan et al. (1995) apply a related idea, implemented as a moving-average filter.

Other means of employing temporal information include the replacement of spectral values in adjacent short-time blocks by either the minimum (Boll, 1979) or median (Godsill & Rayner, 1998, p. 147) value thereof. Similar ideas are proposed by Sondhi et al. (1981) and Vaseghi and Frayling-Cork (1992). Goh et al. (1998) also describe a speech-specific method which attempts to classify residual spectral components as either speech or noise and applies a median filter to those classified as the latter. Many perceptually motivated methods, including our own, also incorporate information from the surrounding time-frequency plane in that the masked threshold estimate (m_k in our case) may be a function of nearby frequencies and even neighbouring times, if nonsimultaneous masking effects are taken into account.

3.2 Qualitative performance comparison

We tested several of the aforementioned standard and perceptually motivated noise reduction techniques using high-fidelity 16-bit, 44.1-kHz noise-free recordings to which white Gaussian noise had been added to yield a SNR of 0–30 dB, as well as older recordings degraded by broadband noise. We implemented these noise reduction methods via the overlap-add method of short-time Fourier transform analysis and synthesis (Allen, 1977); representative audio examples may be found at http://www-sigproc.eng.cam.ac.uk/~pjw47. In addition, an STSA Matlab toolbox is available for free distribution, allowing for the reproduction of these examples as well as further experimentation.

For the case of music signals, differences in the suppression rules of Figure 3 tend to be subtle (Godsill & Rayner, 1998, p. 141), and in fact an evaluation of STSA techniques applied to music signals by Cappé and Laroche (1995) indicates that parameters such as short-time block duration and estimated noise variance are more important. Most perceptually motivated suppression rules, and indeed all of the more advanced techniques, yield superior performance in comparison with basic suppression rules, which generate substantial levels of musical noise and fail completely at low SNR. As is to be expected from smoothing due to the decision-directed estimate of the *a priori* SNR, audio signals restored using this method exhibit a more colourless residual noise.

Differences between our estimator and advanced non-perceptually motivated suppression rules such as the MMSE estimator of (6) are subtle at high SNR, but we conclude that ours is able to prevent the attenuation of some low-level signal components – an important consideration for music restoration. At low SNR differences are more noticeable, and the quality of the restoration obtained depends increasingly on the choice of masking model. We view these results as a first step towards a perceptually optimal spectral amplitude estimator; work is presently underway to extend both the probabilistic modelling framework and the perceptual models employed.

4 Conclusion

We have presented an overview of perceptually motivated approaches to music restoration. Beginning with an explanation of standard broadband noise reduction techniques, we then introduced elements of auditory perception shown to play a central role in perceptually motivated noise reduction methods. After a review of such methods we described our own framework for the incorporation of perceptual optimality criteria into the noise reduction process. This method can easily be incorporated into existing audio restoration frameworks, and its computational complexity is equivalent to that of the other algorithms we have presented. We then described in general terms the performance of STSA techniques. In evaluating the performance of our method, we concluded that it is superior to basic suppression rules and compares favourably with more advanced ones.

Audio restoration always involves a compromise between the amount of noise reduction obtained and the amount of distortion introduced into the signal, and the perceptually motivated methods we have reviewed are significant in that they take advantage of human auditory perception in an attempt to optimise this compromise. Although fine-tuning will always be necessary in practice to obtain optimum performance from any noise reduction technique applied to audio restoration, we have attempted to minimise the need for such adjustments through development of a statistical modelling approach based on perceptual optimality criteria.

5 Acknowledgements

Material by the first author is based upon work supported under a U.S. National Science Foundation Graduate Fellowship.

References

Allen, J.B. (1977). Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-25, 235–238.

Arslan, L., McCree, A., & Viswanathan, V. (1995). New methods for adaptive noise suppression. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 812–815). New York: Institute of Electrical and Electronics Engineers, Inc.

Azirani, A.A., le Bouquin Jeannès, R., & Faucon, G. (1995). Optimizing speech enhancement by exploiting masking properties of the human ear. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal*

Processing, Vol. 1 (pp. 800–803). New York: Institute of Electrical and Electronics Engineers, Inc.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 208–211). New York: Institute of Electrical and Electronics Engineers, Inc.

Boll, S.F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-27, 113–120.

Canagarajah, C.N. (1993). *Digital Signal Processing Techniques for Speech Enhancement in Hearing Aids*. Ph.D. dissertation, University of Cambridge.

Canazza, S., De Poli, G., Maesno, S., & Mian, G.A. (1999). On the performance of a noise reduction technique based on a psychoacoustic model for restoration of old audio recordings. In: H.G. Feichtinger & M. Dörfler (Eds.), *Proceedings of the Diderot Forum on Mathematics and Music: Computational and Mathematical Methods in Music* (pp. 29–35). Vienna: Austrian Computer Society.

Cappé, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, *2*, 345–349.

Cappé, O. & Laroche, J. (1995). Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings. *IEEE Transactions on Speech and Audio Processing*, *3*, 84–93.

Czyzewski, A. & Krolikowski, R. (1999). Noise reduction in audio signals based on the perceptual coding approach. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 147–150). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Delgutte, B. (1990). Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *Journal of the Acoustical Society of America*, *87*, 791–809.

Dörfler, M. (2001). Time-frequency analysis for music signals: A mathematical approach. In: H.G. Feichtinger & M. Dörfler (Eds.), *Computational and Mathematical Methods in Music*, a special issue of the *Journal of New Music Research*, *30*, 3–12.

Ephraim, Y. & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32, 1109–1121.

Ephraim, Y. & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33, 443–445.

Ephraim, Y. & Van Trees, H.L. (1995a). A signal subspace approach for speech enhancement. *IEEE Transactions on Speech and Audio Processing*, *3*, 251–266.

Ephraim, Y. & Van Trees, H.L. (1995b). A spectrally-based signal subspace approach for speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 804–807). New York: Institute of Electrical and Electronics Engineers, Inc.

Fastl, H. (1976). Temporal masking effects: I. Broad band noise masker. *Acustica*, *35*, 287–302.

Fastl, H. (1977). Temporal masking effects: II. Critical band noise masker. *Acustica*, *36*, 317–331.

Fastl, H. (1979). Temporal masking effects: III. Pure tone masker. *Acustica*, *43*, 282–294.

Fastl, H. & Bechly, M. (1983). Suppression in simultaneous masking. *Journal of the Acoustical Society of America*, *74*, 754–757.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, *12*, 47–65.

Godsill, S.J. & Rayner, P.J.W. (1998). *Digital Audio Restoration: A Statistical Model Based Approach*. Berlin: Springer-Verlag.

Goh, Z., Tan, K.-C., & Tan, B.T.G. (1998). Postprocessing methods for suppressing musical noise generated by spectral subtraction. *IEEE Transactions on Speech and Audio Processing*, *6*, 287–292.

Gradshteyn, I.S. & Ryzhik, I.M. (1994). *Table of Integrals, Series, and Products*. San Diego: Academic Press, Inc., fifth edition.

Gülzow, T., Engelsberg, A., & Heute, U. (1998). Comparison of a discrete wavelet transformation and a nonuniform polyphase filterbank applied to spectral-subtraction speech enhancement. *Signal Processing*, *64*, 5–19.

Gustafsson, S., Jax, P., & Vary, P. (1998). A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 397–400). New York: Institute of Electrical and Electronics Engineers, Inc.

Irino, T. (1999). Noise suppression using a time-varying, analysis/synthesis gammachirp filterbank. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 97–100). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Irino, T. & Patterson, R.D. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *Journal of the Acoustical Society of America*, *101*, 412–419.

Jesteadt, W., Bacon, S.P., & Lehman, J.R. (1982). Forward masking as a function of frequency, masker level, and signal delay. *Journal of the Acoustical Society of America*, *71*, 950–962.

Johnston, J.D. (1988). Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, *6*, 314–323.

Kidd, Jr., G. & Feth, L.L. (1982). Effects of masker duration in pure-tone forward masking. *Journal of the Acoustical Society of America*, *72*, 1384–1386.

Kim, W., Kang, S., & Ko, H. (2000). Spectral subtraction based on phonetic dependency and masking effects. *IEE Proceedings-Vision, Image, and Signal Processing*, *147*, 423–427.

Lorber, M. & Hoeldrich, R. (1997). A combined approach for broadband noise reduction. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and*

Acoustics. New York: Institute of Electrical and Electronics Engineers, Inc.

McAulay, R.J. & Malpass, M.L. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28, 137–145.

Moore, B.C.J. (1997). *An Introduction to the Psychology of Hearing*. San Diego: Academic Press, fourth edition.

Moore, B.C.J. & Glasberg, B.R. (1983a). Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise. *Journal of the Acoustical Society of America*, *73*, 1249–1259.

Moore, B.C.J. & Glasberg, B.R. (1983b). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*, 750–753.

Moore, B.C.J. & Glasberg, B.R. (1986). The role of frequency selectivity in the perception of loudness, pitch, and time. In: B.C.J. Moore (Ed.), *Frequency Selectivity in Hearing* (pp. 251–308). London: Academic Press.

Moore, B.C.J. & Glasberg, B.R. (1996). A revision of Zwicker's loudness model. *Acustica*, *82*, 335–345.

Moore, B.C.J., Glasberg, B.R., & Baer, T. (1997). A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, *45*, 224–240.

Patterson, R.D. & Moore, B.C.J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In: B.C.J. Moore (Ed.), *Frequency Selectivity in Hearing* (pp. 123–177). London: Academic Press.

Patterson, R.D. & Nimmo-Smith, I. (1980). Off-frequency listening and auditory-filter asymmetry. *Journal of the Acoustical Society of America*, *67*, 229–245.

Patterson, R.D., Nimmo-Smith, I., Weber, D.L., & Milroy, R. (1982). The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America*, *72*, 1788–1803.

Petersen, T.L. & Boll, S.F. (1981). Acoustic noise suppression in the context of a perceptual model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 3 (pp. 1086–1088). New York: Institute of Electrical and Electronics Engineers, Inc.

Sondhi, M.M., Schmidt, C.E., & Rabiner, L.R. (1981). Improving the quality of a noisy speech signal. *Bell System Technical Journal*, *60*, 1847–1859.

Stuart, J.R. (1994). Noise: Methods for estimating detectability and threshold. *Journal of the Audio Engineering Society*, *42*, 124–140.

Tsoukalas, D., Paraskevas, M., & Mourjopoulos, J. (1993). Speech enhancement using psychoacoustic criteria. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2 (pp. 359–362). New York: Institute of Electrical and Electronics Engineers, Inc.

Tsoukalas, D.E., Mourjopoulos, J., & Kokkinakis, G. (1997a). Perceptual filters for audio signal enhancement. *Journal of the Audio Engineering Society*, *45*, 22–36.

Tsoukalas, D.E., Mourjopoulos, J.N., & Kokkinakis, G. (1997b). Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing*, *5*, 497–514.

Van Trees, H.L. (1968). *Detection, Estimation and Modulation Theory: Part 1, Detection, Estimation and Linear Modulation Theory*. New York: John Wiley & Sons, Inc.

Vaseghi, S.V. & Frayling-Cork, R. (1992). Restoration of old gramophone recordings. *Journal of the Audio Engineering Society*, *40*, 791–801.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on Speech and Audio Processing*, *7*, 126–137.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. Principles of Electrical Engineering Series. Cambridge, MA: MIT Press.

Wolfe, P.J. & Godsill, S.J. (1999). Formalising perceptually motivated approaches to music restoration. In: H.G. Feichtinger & M. Dörfler (Eds.), *Proceedings of the Diderot Forum on Mathematics and Music: Computational and Mathematical Methods in Music* (pp. 367–373). Vienna: Austrian Computer Society.

Wolfe, P.J. & Godsill, S.J. (2000). Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2 (pp. 821–824). Piscataway, NJ: Institute of Electrical and Electronics Engineers, Inc.

Wolfe, P.J. & Godsill, S.J. (2001). *On Bayesian Estimation of Spectral Components for Broadband Noise Reduction in Audio Signals*. Technical Report CUED/F-INFENG/TR.404, Department of Engineering, University of Cambridge.

Zwicker, E. (1976). Influence of a complex masker's time structure on masking. *Acustica*, *34*, 138–146.

Zwicker, E. (1984). Dependence of post-masking on masker duration and its relation to temporal effects in loudness. *Journal of the Acoustical Society of America*, *75*, 219–223.

Zwicker, E. & Fastl, H. (1999). *Psychoacoustics: Facts and Models*, Vol. 22 of *Springer Series in Information Sciences*. Berlin: Springer, second edition.