Digital Audio Restoration

Simon Godsill and Peter Rayner
Dept. of Engineering
University of Cambridge,
Cambridge, U.K.
sjg,pjwr@eng.cam.ac.uk

Olivier Cappé
ENST
Paris, France
cappe@sig.enst.fr

June 2, 1997

Abstract

This chapter is concerned with the application of modern signal processing techniques to the restoration of degraded audio signals. Although attention is focussed on gramophone recordings, film sound tracks and tape recordings, many of the techniques discussed have applications in other areas where degraded audio signals occur, such as speech transmission, telephony and hearing aids.

We aim to provide a wide coverage of existing methodology while giving insight into current areas of research and future trends.

1 Introduction

The introduction of high quality digital audio media such as Compact Disk (CD) and Digital Audio Tape (DAT) has dramatically raised general awareness and expectations about sound quality in all types of recordings. This, combined with an upsurge in interest in historical and nostalgic material, has led to a growing requirement for restoration of degraded sources ranging from the earliest recordings made on wax cylinders in the nineteenth century, through disc recordings (78rpm, LP, etc.) and finally magnetic tape recording technology, which has been available since the 1950's. Noise reduction may occasionally be required even in a contemporary digital recording if background noise is judged to be intrusive.

Degradation of an audio source will be considered as any undesirable modification to the audio signal which occurs as a result of (or subsequent to) the

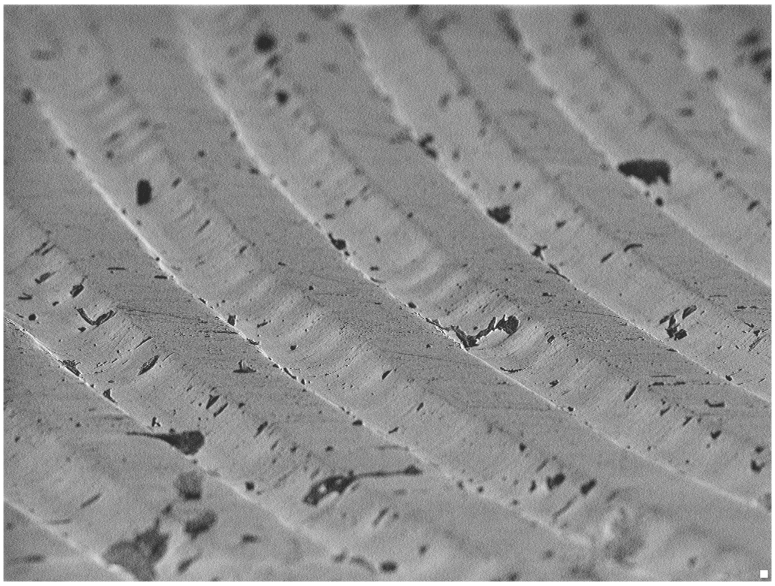

Figure 1: Electron micrograph showing dust and damage to the grooves of a 78rpm gramophone disc.

recording process. For example, in a recording made direct-to-disc from a microphone, degradations could include noise in the microphone and amplifier as well as noise in the disc cutting process. Further noise may be introduced by imperfections in the pressing material, transcription to other media or wear and tear of the medium itself. An example of such noise can be seen in the electron micrograph shown in figure 1[1]. We do not strictly consider any noise present in the recording environment such as audience noise at a musical performance to be degradation, since this is part of the 'performance'. Removal of such performance interference is a related topic which is considered in other applications, such as speaker separation for hearing aid design. An ideal restoration would then reconstruct the original sound source exactly as received by the transducing equipment (microphone, acoustic horn, etc.). Of course, this ideal can never be achieved perfectly in practice, and methods can only be devised which come close according to some suitable error criterion. This should ideally be based on the perceptual characteristics of the human listener.

Analogue restoration techniques have been available for at least as long as magnetic tape, in the form of manual cut-and-splice editing for clicks and frequency domain equalization for background noise (early mechanical disk playback equipment will also have this effect by virtue of its poor response at high frequencies). More sophisticated electronic click reducers were based upon high pass filtering for detection of clicks, and low pass filtering to mask their ef-

[1]With acknowledgment to Mr. B.C. Breton, Scientific Imaging Group, CUED

fect [28, 78].[2] None of these methods was sophisticated enough to perform a significant degree of noise reduction without interfering with the underlying signal quality. Digital methods allow for a much greater degree of flexibility in processing, and hence greater potential for noise removal, although indiscriminate application of inappropriate digital methods can be more disastrous than analogue processing!

Some of the earliest digital signal processing work for audio restoration involved deconvolution for enhancement of a solo voice (Caruso) from an acoustically recorded source (see Miller [100] and Stockham *et al.* [135]). Since then, research groups at Cambridge, Le Mans, Paris and elsewhere have worked in the area, developing sophisticated techniques for treatment of degraded audio. The results of this research are summarized and referenced later in the chapter.

There are several distinct types of degradation common in audio sources. These can be broadly classified into two groups: *localized* degradations and *global* degradations. Localized degradations are discontinuities in the waveform which affect only certain samples, including clicks, crackles, scratches , breakages and clipping. Global degradations affect all samples of the waveform and include background noise, wow and flutter and certain types of non-linear distortion. Mechanisms by which all of these defects can occur are discussed later.

The chapter is organized as follows. We firstly describe models which are suitable for audio signal restoration, in particular those which are used in later work. Subsequent sections describe individual restoration problems separately, considering the alternative methods available to the restorer. A concluding section summarizes the work and discusses future trends.

2 Modelling of audio signals

Many signal processing techniques will be model-based, either explicitly or implicitly, and this certainly applies to most of the audio restoration algorithms currently available. The quality of processing will depend largely on how well the modelling assumptions fit the data. For an audio signal, which might contain speech, music and general acoustical noises the model must be quite general and robust to deviations from the assumptions. It should also be noted that most audio signals are non-stationary in nature, although practical modelling will often assume short-term stationarity of the signal. We now discuss some models which are appropriate for audio signals.

A model which has found application in many areas of time series processing, including audio restoration (see sections 3 and 7), is the autoregressive (AR) or all-pole model (see Box and Jenkins [18], Priestley [120] and also Makhoul [93] for an introduction to linear predictive analysis) in which the current value of a signal is represented as a weighted sum of P previous signal values and a white

[2] the 'Packburn' unit achieved masking within a stereo setup by switching between channels

noise term:

$$s[n] = \sum_{i=1}^{P} s[n-i]\, a_i \;+\; e[n]. \tag{1}$$

The AR model is a reasonable representation for many stationary linear processes, allowing for noise-like signals (poles close to origin) and near-harmonic signals (poles close to unit circle). A more appropriate model for many situations might be the autoregressive moving-average (ARMA) model which allows zeros as well as poles. However, the AR model offers far greater analytic flexibility than the ARMA model, so a high order AR model will often be used in practice to approximate an ARMA signal (it is well known that an infinite order AR model can represent any finite-order ARMA model (see, e.g. [137])). Model order for the autoregressive process will reflect the complexity of the signal under consideration. For example, a highly complex musical signal can require a model order of $P > 100$ to represent the waveform adequately, while simpler signals may be modelled by an order 30 system. Strictly, any processing procedure should thus include a model order selection strategy. For many applications, however, it is sufficient to fix the model order to a value high enough for representation of the most complex signal likely to be encountered. Clearly no audio signal is truly stationary, so it will be necessary to implement the model in a block-based or adaptive fashion. Suitable block lengths and adaptation rates will depend upon the signal type, but block lengths between 500 and 2000 samples at the 44.1kHz sampling rate are generally found to be appropriate.

There are many well-known methods for estimating AR models, including maximum likelihood/least-squares [93] and methods robust to noise [73, 132]. Adaptive parameter estimation schemes are reviewed in [70]. The class of methods robust to noise, both block-based and adaptive, will be of importance to many audio restoration applications, since standard parameter estimation schemes can be heavily biased in the presence of noise, in particular impulsive noise such as is commonly encountered in click-degraded audio. A standard approach to this problem is the M-estimator [73, 132]. This method achieves robustness by iteratively re-weighting excitation values in the least-squares estimator using a non-linear function such as Huber's psi-function [72] or Tukey's bisquare function [104]. Applications of these methods to parameter estimation, detection of impulses and robust filtering include [95, 6, 40].

Another model which is a strong candidate for musical signals is the sinusoidal model, which has been used effectively for speech applications ([97] and chapter **??** of this book). A constrained form of the sinusoidal model is implicitly at the heart of short-time spectral attenuation (STSA) methods of noise reduction (see section 5.1). The model is also a fundamental assumption of the pitch variation algorithms presented in section 6. In its general form the signal can be expressed as:

$$s[n] = \sum_{i=1}^{P_n} a_i[n] \sin\left(\int_0^{nT} \omega_i(t)\, dt \;+\; \phi_i \right). \tag{2}$$

This is quite a general model, allowing for frequency and amplitude modulation (by allowing $a_i[n]$ and $\omega_i(t)$ to vary with time) as well as the 'birth' and 'death' of individual components (by allowing P_n to vary with time). However, parameter estimation for such a general model is difficult, and restrictive constraints must typically be placed upon the amplitude and frequency variations. The sinusoidal model is not suited to modelling noise-like signals, although an acceptable representation can be achieved by using a large number of sinusoids in the expansion.

Other models include adaptations to the basic AR/ARMA models to allow for speech-like periodic excitation pulses [123] and non-linearity (see section 7). Further 'non-parametric' modelling possibilities arise from other basis function expansions which might be more appropriate for audio signal analysis, including Wavelets [1] and signal dependent transforms which employ principal component-based analysis [52]. Choice of model will in general involve a compromise between prior knowledge of signal characteristics, computational power and how critical the accuracy of the model is to the application.

3 Click Removal

The term 'clicks' is used here to refer to a generic localized type of degradation which is common to many audio media. We will classify all finite duration defects which occur at random positions in the waveform as clicks. Clicks are perceived in a number of ways by the listener, ranging from tiny 'tick' noises which can occur in any recording medium, including modern digital sources, through the characteristic 'scratch' and 'crackle' noise associated with most analogue disc recording methods. For example, a poor quality 78rpm record might typically have around 2,000 clicks per second of recorded material, with durations ranging from less than 20μs up to 4ms in extreme cases. See figure 2 for a typical example of a recorded music waveform degraded by localized clicks. In most examples at least 90% of samples remain undegraded, so it is reasonable to hope that a convincing restoration can be achieved.

There are many mechanisms by which clicks can occur. Typical examples are specks of dirt and dust adhering to the grooves of a gramophone disc (see figure 1) or granularity in the material used for pressing such a disc. Further click-type degradation may be caused through damage to the disc in the form of small scratches on the surface. Similar artefacts are encountered in other analogue media, including optical film sound tracks and early wax cylinder recordings, although magnetic tape recordings are generally free of clicks. Ticks can occur in digital recordings as a result of poorly concealed digital errors and timing problems.

Peak-related distortion, occurring as a result either of overload during recording or wear and tear during playback, can give rise to a similar perceived effect to clicks, but is really a different area which should receive separate attention (see section 7), even though click removal systems can often go some way towards alleviating the worst effects.

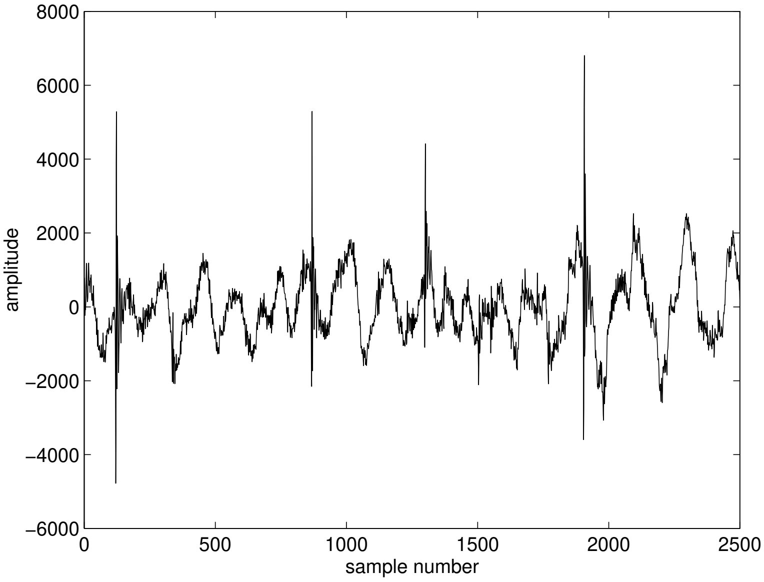

Figure 2: Click-degraded Music Waveform taken from 78rpm recording

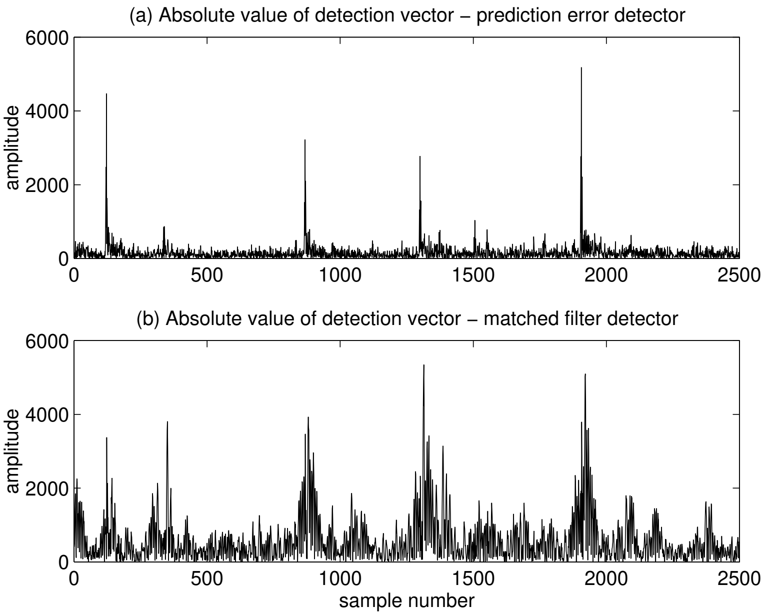

Figure 3: AR-based detection, P=50. (a) Prediction error filter (b) Matched filter.

3.1 Modelling of clicks

Localized defects may be modelled in many different ways. For example, a defect may be additive to the underlying audio signal, or it may replace the signal altogether for some short period. An additive model has been found to be acceptable for most surface defects in recording media, including small scratches, dust and dirt. A replacement model may be appropriate for very large scratches and breakages which completely obliterate any underlying signal information, although such defects usually excite long-term resonances in mechanical playback systems and must be treated differently (see section 4). Here we will consider primarily the additive model , although many of the results are at least robust to replacement noise.

An additive model for localized degradation can be expressed as:

$$x[n] \quad = \quad s[n] \, + \, i[n] \, v[n] \tag{3}$$

where $s[n]$ is the underlying audio signal, $v[n]$ is a corrupting noise process and $i[n]$ is a 0/1 'switching' process which takes the value 1 only when the localized degradation is present. Clearly the value of $v[n]$ is irrelevant to the output when the switch is in position 0. The statistics of the switching process $i[n]$ thus govern which samples are degraded, while the statistics of $v[n]$ determine the amplitude characteristics of the corrupting process.

This model is quite general and can account for a wide variety of noise characteristics encountered in audio recordings. It does, however, assume that the degradation process does not interfere with the timing content of the original signal, as observed in $x[n]$. This is reasonable for all but very severe degradations, which might temporarily upset the speed of playback, or actual breakages in the medium which have been mechanically repaired (such as a broken disc recording which has been glued back together).

Any procedure which is designed to remove localized defects in audio signals must take account of the typical characteristics of these artefacts. Some important features which are common to many click-degraded audio media include:

- Degradation tends to occur in contiguous 'bursts' of corrupted samples, starting at random positions in the waveform and of random duration (typically between 1 and 200 samples at 44.1 kHz sampling rates). Thus there is strong dependence between successive samples of the switching process $i[n]$, and the noise cannot be assumed to follow a classical impulsive noise pattern in which single impulses occur independently of each other (the Bernoulli model). It is considerably more difficult to treat clusters of impulsive disturbance than single impulses, since the effects of adjacent impulses can cancel each other in the detection space ('missed detections') or add constructively to give the impression of more impulses ('false alarms').

- The amplitude of the degradation can vary greatly within the same recorded extract, owing to a range of size of physical defects. For example, in many recordings the largest click amplitudes will be well above the largest signal amplitudes, while the smallest audible defects can be more than 40dB below the local signal level (depending on psychoacoustical masking by the signal and the amount of background noise). This leads to a number of difficulties. In particular, large amplitude defects will tend to bias any parameter estimation and threshold determination procedures, leaving smaller defects undetected. As we shall see in section 3.2.1, threshold selection for some detection schemes becomes a difficult problem in this case.

Many approaches are possible for the restoration of such defects. It is clear, however, that the ideal system will process only on those samples which are degraded, leaving the others untouched in the interests of fidelity to the original source. Two tasks can thus be identified for a successful click restoration system. The first is a detection procedure in which we estimate the process $i[n]$, that is decide which samples are degraded. The second is an estimation procedure in which we attempt to reconstruct the underlying audio data when corruption is present. A method which assumes that no useful information about the underlying signal is contained in the degraded samples will involve a pure interpolation of the audio data using the undegraded samples, while more sophisticated techniques will attempt in addition to extract extra information from samples degraded with noise using some degree of noise modelling.

3.2 Detection

Click detection for audio signals involves the identification of samples which are not drawn from the underlying audio signal; in other words they are drawn from a spurious 'outlier' distribution. We will see a close relationship between click detection and work in robust parameter estimation and treatment of outliers, from fields as diverse as medical signal processing, underwater signal processing and statistical data analysis. In the statistical field in particular there has been a vast amount of work in the treatment of outliers (see e.g. [9, 8] for extensive review material, and further references in section 3.4). Various criteria for detection are possible, including minimum probability of error, P_E, and related concepts, but strictly speaking the aim of any audio restoration is to remove only those artefacts which are audible to the listener. Any further processing is not only unnecessary but will increase the chance of distorting the perceived signal quality. Hence a truly optimal system should take into account the trade-off between the audibility of artefacts and perceived distortion as a result of processing, and will involve consideration of complex psychoacoustical effects in the human ear (see e.g. [102]). Such an approach, however, is difficult both to formulate and to realize, so we will limit discussion here only to criteria which are well understood in a mathematical sense.

The simplest click detection methods involve a high-pass filtering operation on the signal, the assumption being that most audio signals contain little information at high frequencies, while impulses have spectral content at all frequencies.

Clicks are thus enhanced relative to the signal by the high-pass filtering operation and can easily be detected by thresholding the filtered output. The method has the advantage of being simple to implement and having no unknown system parameters (except for a detection threshold). This principle is the basis of most analogue de-clicking equipment [28, 78] and some simple digital click detectors [76]. Of course, the method will fail if the audio signal has strong high frequency content or the clicks are band-limited. Along similar lines, wavelets and multiresolution methods in general [1, 31, 32] have useful localization properties for singularities in signals (see e.g. [94]), and a Wavelet filter at a fine resolution can be used for the detection of clicks. Such methods have been studied and demonstrated successfully by Montresor, Valière *et al.* [142, 101].

Other methods attempt to incorporate prior information about signal and noise into a model-based detection procedure. Techniques for detection and removal of impulses from autoregressive signals have been developed from robust filtering principles (see section 2 and [6, 40]). These methods apply non-linear functions to the autoregressive excitation sequence, and can be related to the click detection methods of Vaseghi and Rayner [147, 145, 149], which are now discussed. See also section 3.4 for recent detection methods based on statistical decision theory.

3.2.1 Autoregressive (AR) model-based Click Detection

In this method ([147, 145, 149]) the underlying audio data $s[n]$ is assumed to be drawn from a short-term stationary autoregressive (AR) process (see equation (1)). The AR model parameters $\mathbf{a}$ and the excitation variance σ_e^2 are estimated from the corrupted data $x[n]$ using some procedure robust to impulsive noise, such as the M-estimator (see section 2).

The corrupted data $x[n]$ is filtered using the prediction error filter $H(z) = (1 - \sum_{i=1}^{P} a_i\, z^{-i})$ to give a detection signal $e_d[n]$:

$$e_d[n] = x[n] \; - \sum_{i=1}^{P} x[n-i]\, a_i. \tag{4}$$

Substituting for $x[n]$ from (3) and using (1) gives:

$$e_d[n] = e[n] \; + \; i[n]\, v[n] - \sum_{i=1}^{P} i[n-i]\, v[n-i]\, a_i \tag{5}$$

which is composed of the signal excitation $e[n]$ and a weighted sum of present and past impulsive noise values. If $s[n]$ is zero mean and has variance σ_s^2 then $e[n]$ is white noise with variance $\sigma_e^2 = 2\pi \frac{\sigma_s^2}{\int_{-\pi}^{\pi} \frac{1}{|H(e^{j\theta})|^2} d\theta}$. The reduction in power here from signal to excitation can be 40dB or more for highly correlated audio signals. Consideration of (5), however, shows that a single impulse contributes the impulse response of the prediction error filter, weighted by the impulse amplitude, to the detection signal $e_d[n]$, with maximum amplitude corresponding to the maximum in the impulse response. This means that considerable amplification of the impulse relative to the signal can be achieved for all but uncorrelated, noise-like signals. It should be noted, however, that this amplification is achieved at the expense of localization in time of the impulse, whose effect is now spread over $P + 1$ samples of the detection signal $e_d[n]$. This will have adverse consequences when a number of impulses is present in the same vicinity, since their impulse responses may cancel one another out or add constructively to give false detections . More generally, threshold selection will be troublesome when impulses of widely differing amplitudes are present, since a low threshold which is appropriate for very small clicks will lead to false detections in the P detection values which follow a large impulse.

Detection can then be performed by thresholding $e_d[n]^2$ to identify likely impulses. Choice of threshold will depend upon the AR model, the variance of $e[n]$ and the size of impulses present (see [54] for optimal thresholds under Gaussian signal and noise assumptions), and will reflect trade-offs between false and missed detection rates. See figure 3(a) for a typical example of detection using this method, which shows how the impulsive interference is strongly amplified relative to the signal component.

An adaptation of this method, also devised by Vaseghi and Rayner, considers the impulse detection problem from a matched filtering perspective [143]. The 'signal' is the impulse itself, while the autoregressive audio data is regarded as coloured additive noise. The prediction error filter described above can then be viewed as a pre-whitening stage for the autoregressive noise, and the full matched filter is given by $H(z)H(z^{-1})$, a non-causal filter with $2P+1$ coefficients which can be realized with P samples of lookahead. The matched filtering approach provides additional amplification of impulses relative to the signal, but further reduces localization of impulses for a given model order. Choice between the two methods will thus depend on the range of click amplitudes present in a particular recording and the degree of separation of individual impulses in the waveform. See figure 3(b) for an example of detection using the matched filter. Notice that the matched filter has high-lighted a few additional impulse positions, but at the expense of a much more 'smeared' response which will make accurate localization very awkward. Hence the prediction-error detector is usually preferred in practice.

Both the prediction error detection algorithm and the matched filtering algorithm are efficient to implement and can be operated in real time using DSP microprocessors. Results of a very high standard can be achieved if a careful strategy is adopted for extracting the precise click locations from the detection signal. Iterative schemes are also possible which re-apply the detection algorithms to the restored data (see section 3.3) in order to achieve improved parameter estimates and to ensure that any previously undetected clicks are detected.

3.3 Replacement of corrupted samples

Once clicks have been detected, a replacement strategy must be devised to mask their effect. It is usually appropriate to assume that clicks have in no way interfered with the timing of the material, so the task is then to fill in the 'gap' with appropriate material of identical duration to the click. As discussed above, this amounts to an interpolation or generalized prediction problem, making use of the good data values surrounding the corruption and possibly taking account of signal information which is buried in the corrupted section. An effective technique will have the ability to interpolate gap lengths from one sample up to at least 100 samples at a sampling rate of 44.1kHz.

The replacement problem may be formulated as follows. Consider N samples of audio data, forming a vector $\mathbf{s}$. The corresponding click-degraded data vector is $\mathbf{x}$, and the (known) vector of detection values $i[n]$ is $\mathbf{i}$. The audio data $\mathbf{s}$ may be partitioned into two sub-vectors, one containing elements whose value is known (i.e. $i[n] = 0$), denoted by $\mathbf{s}_{\mathcal{K}}$, and the second containing unknown elements which are corrupted by noise ($i[n] = 1$), denoted by $\mathbf{s}_{\mathcal{U}}$. Vectors $\mathbf{x}$ and $\mathbf{i}$ are partitioned in a similar fashion. The replacement problem requires the estimation of the unknown data $\mathbf{s}_{\mathcal{U}}$, given the observed (corrupted) data $\mathbf{x}$. This will be a statistical estimation procedure for audio signals, which are stochastic in nature, and estimation methods might be chosen to satisfy criteria such as minimum mean-square error (MMSE), maximum likelihood (ML), maximum *a posteriori* (MAP) or perceptual features.

Numerous methods have been developed for the interpolation of corrupted or missing samples in speech and audio signals. The 'classical' approach is perhaps the median filter [141, 116] which can replace corrupted samples with a median value while retaining detail in the signal waveform. A suitable system is described in [76], while a hybrid autoregressive prediction/ median filtering method is presented in [109]. Median filters, however, are too crude to deal with gap lengths greater than a few samples. Other techniques 'splice' uncorrupted data from nearby into the gap [90, 117] in such a manner that there is no signal discontinuity at the start or end of the gap. These methods rely on the periodic nature of many speech and music signals and also require a reliable estimate of pitch period.

The most effective and flexible methods to date have been model-based, allowing for the incorporation of reasonable prior information about signal characteristics. A good coverage is given by Veldhuis [150], and a number of interpolators suited to speech and audio signals is presented. These are based on minimum variance estimation under various modelling assumptions, including sinusoidal, autoregressive, and periodic. The autoregressive interpolator, originally derived in [74], was later developed by Vaseghi and Rayner [145, 147, 149] for the restoration of gramophone recordings. This interpolator and other developments based on autoregressive modelling are discussed in the next section.

3.3.1 Autoregressive interpolation

An interpolation procedure which has proved highly successful is the Least Squares AR-based (LSAR) method [74, 150], devised originally for the concealment of uncorrectable errors in CD systems. Corrupted data is considered truly 'missing' in that no account is taken of its value in making the interpolation. We present the algorithm in a matrix/vector notation in which the locations of degraded samples can be arbitrarily specified within the data block through the detection vector $\mathbf{i}$.

Consider a block of N data samples $\mathbf{s}$ which are drawn from a short-term stationary AR process with parameters $\mathbf{a}$. Equation 1 can be re-written in matrix/vector notation as:

$$\mathbf{e} = \mathbf{A}\,\mathbf{s} \tag{6}$$

where $\mathbf{A}$ is an $((N-P)\times N)$ matrix, whose $(j-P)$th row is constructed so as to generate the prediction error, $e[j] = s[j] - \sum_{i=1}^{P} s[j-i]\,a_j$. Elements on the right hand side of this equation can be partitioned into known and unknown sections as described above, with $\mathbf{A}$ being partitioned by column. The least squares solution is then obtained by minimizing the sum of squares $E = \mathbf{e}^T\mathbf{e}$ w.r.t. the unknown data segment, to give the solution:

$$\mathbf{s}_{\mathcal{U}}^{\text{LS}} = -(\mathbf{A}_{\mathcal{U}}{}^T\mathbf{A}_{\mathcal{U}})^{-1}\mathbf{A}_{\mathcal{U}}{}^T\mathbf{A}_{\mathcal{K}}\;\mathbf{s}_{\mathcal{K}}. \tag{7}$$

This interpolator has useful properties, being the minimum-variance unbiased estimator for the missing data [150]. Viewed from a probabilistic perspective, it corresponds to maximization of $p(\mathbf{s}_{\mathcal{U}} \mid \mathbf{s}_{\mathcal{K}}, a, \sigma_e^2)$ under Gaussian assumptions,[3] and is hence also the maximum *a posteriori* (MAP) estimator [54, 150]. In cases where corruption occurs in contiguous bursts separated by at least P 'good' samples, the interpolator leads to a Toeplitz system of equations which can be efficiently solved using the Levinson-Durbin recursion [39]. See figure 4 for examples of interpolation using the LSAR method. A succession of interpolations has been performed, with increasing numbers of missing samples from left to right in the data (gap lengths increase from 25 samples up to more than 100). The autoregressive model order is 60. The shorter length interpolations are almost indistinguishable from the true signal (left-hand side of figure 4(a)), while the interpolation is much poorer as the number of missing samples becomes large (right-hand side of figure 4(b)). This is to be expected of any interpolation scheme when the data is drawn from a random process, but the situation can often be improved by use of a higher order autoregressive model. Despite poor accuracy of the interpoland for longer gap lengths, good continuity is maintained at the start and end of the missing data blocks, and the signal appears to have the right 'character'. Thus effective removal of click artefacts in typical audio sources can usually be achieved.

The basic formulation given in (7) assumes that the AR parameters are known *a priori*. In practice we may have a robust estimate of the parameters

[3] provided that no samples are missing from the first P elements of $\mathbf{s}$; otherwise a correction must be made to the data covariance matrix (see [54])

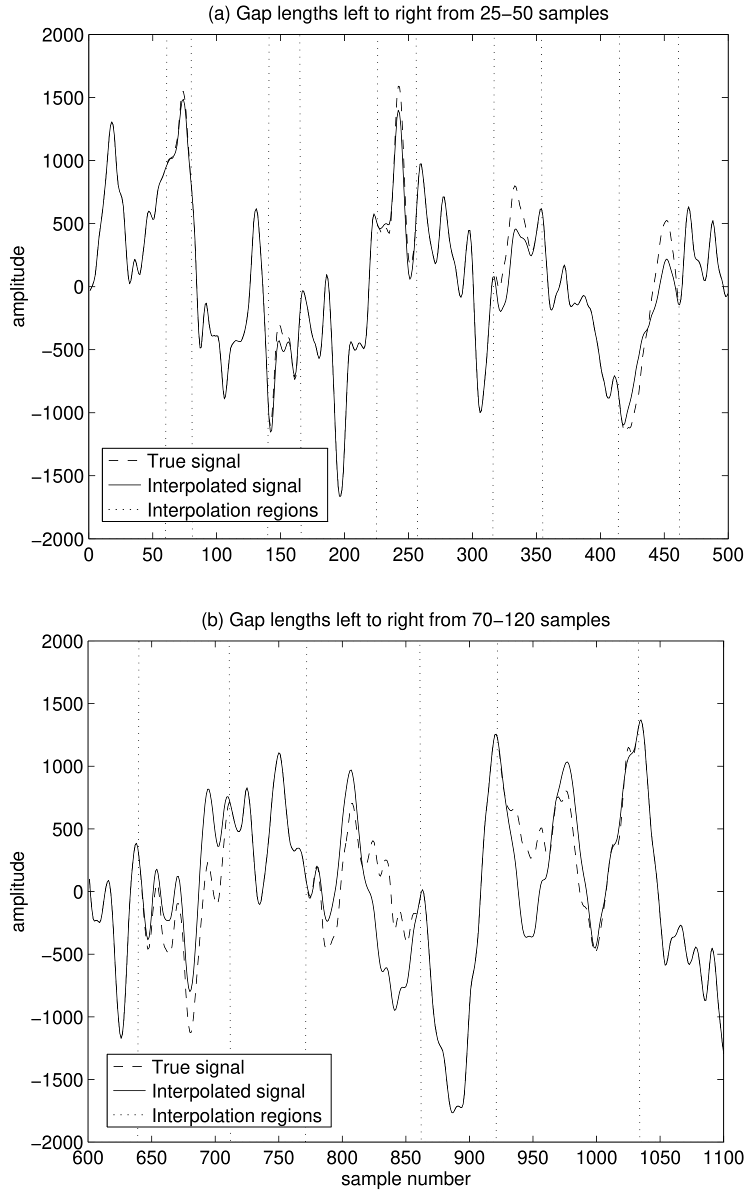

Figure 4: AR-based interpolation, P=60, classical chamber music, (a) short gaps, (b) long gaps

obtained during the detection stage (see section 3.2.1). This, however, is strictly sub-optimal and we should perhaps consider interpolation methods which treat the parameters as unknown. Minimization of the term $E = \mathbf{e}^T\mathbf{e}$ w.r.t. both $\mathbf{s}_\mathcal{U}$ and $\mathbf{a}$ corresponds to the joint least squares estimator for the parameters and the missing data, and also to the approximate joint ML estimator.[4] E, however, contains fourth-order terms in the unknowns and cannot be minimized analytically. Janssen, Veldhuis and Vries [74] propose an alternating variables iteration which performs linear maximizations w.r.t. data and parameters in turn, and is guaranteed to converge at least to a local maximum of the likelihood. The true likelihood for the missing data, $p(\mathbf{s}_\mathcal{K} \mid \mathbf{s}_\mathcal{U})$, can be maximized using the expectation-maximize (EM) algorithm [35], an approach which has been investigated by Ó Ruanaidh and Fitzgerald [127, 128]. Convergence to local maxima of the likelihood is also a potential difficulty with this method.

The LSAR approach to interpolation performs well in most cases. However, certain classes of signal which do not fit the modelling assumptions (such as periodic pulse-driven voiced speech) and very long gap lengths can lead to an audible 'dulling' of the signal or unsatisfactory masking of the original corruption. Increasing the order of the AR model will usually improve the results; however, several developments to the method are now outlined which can lead to better performance.

Vaseghi and Rayner [149] propose an extended AR model to take account of signals with long-term correlation structure, such as voiced speech, singing or near-periodic music. The model, which is similar to the long term prediction schemes used in some speech coders, introduces extra predictor parameters around the pitch period T, so that equation 1 becomes:

$$s[n] = \sum_{i=1}^{P} s[n-i]\, a_i \; + \sum_{j=-Q}^{Q} s[n-T-j]\, b_j \; + \; e[n], \tag{8}$$

where Q is typically smaller than P. Least squares/ML interpolation using this model is of a similar form to equation 7, and parameter estimation is straightforwardly derived as an extension of standard AR parameter estimation methods (see section 2). The method gives a useful extra degree of support from adjacent pitch periods which can only be obtained using very high model orders in the standard AR case. As a result, the 'under-prediction' sometimes observed when interpolating long gaps is improved. Of course, an estimate of T is required, but results are quite robust to errors in this. Veldhuis [150][chapter 4] presents a special case of this interpolation method in which the signal is modelled by one single 'prediction' element at the pitch period (i.e. $Q = 0$ and $P = 0$ in the above equation).

A second modification to the LSAR method is concerned with the characteristics of the excitation signal. We notice that the LSAR procedure (7) seeks to minimize the excitation energy of the signal, irrespective of its time domain

[4] the approximation assumes that the parameter likelihood for the first P data samples is insignificant [18]

autocorrelation. This is quite correct, and desirable mathematical properties result (see above). However, figure 6 shows that the resulting excitation signal corresponding to the corrupted region can be correlated and well below the level of surrounding excitation. As a result, the 'most probable' interpolands may under-predict the true signal levels and be over-smooth compared with the surrounding signal. In other words, ML/MAP procedures do not necessarily generate interpolands which are *typical* for the underlying model, which is an important factor in the *perceived* effect of the restoration. Rayner and Godsill [125] have devised a method which addresses this problem. Instead of minimizing the excitation energy, we consider interpolands with constant excitation energy. The excitation energy may be expressed as:

$$E = (\mathbf{s}_{\mathcal{U}} - \mathbf{s}_{\mathcal{U}}^{\text{LS}})^T \, {\mathbf{A}_{\mathcal{U}}}^T \mathbf{A}_{\mathcal{U}} \, (\mathbf{s}_{\mathcal{U}} - \mathbf{s}_{\mathcal{U}}^{\text{LS}}) + E_{\text{LS}}, \qquad E > E_{\text{LS}}, \tag{9}$$

where E_{LS} is the excitation energy corresponding to the LSAR estimate $\mathbf{s}_{\mathcal{U}}^{\text{LS}}$. The positive definite matrix ${\mathbf{A}_{\mathcal{U}}}^T \mathbf{A}_{\mathcal{U}}$ can be factorized into 'square roots' by Cholesky or any other suitable matrix decomposition [66] to give ${\mathbf{A}_{\mathcal{U}}}^T \mathbf{A}_{\mathcal{U}} = \mathbf{M}^T \mathbf{M}$, where $\mathbf{M}$ is a non-singular square matrix. A transformation of variables $\mathbf{u} = \mathbf{M}\,(\mathbf{s}_{\mathcal{U}} - \mathbf{s}_{\mathcal{U}}^{\text{LS}})$ then serves to de-correlate the missing data samples, simplifying equation (9) to:

$$E = \mathbf{u}^T \mathbf{u} + E_{\text{LS}}, \tag{10}$$

from which it can be seen that the (non-unique) solutions with constant excitation energy correspond to vectors $\mathbf{u}$ with constant L_2-norm. The resulting interpoland can be obtained by the inverse transformation $\mathbf{s}_{\mathcal{U}} = \mathbf{M}^{-1}\mathbf{u} + \mathbf{s}_{\mathcal{U}}^{\text{LS}}$. One suitable criterion for selecting $\mathbf{u}$ might be to minimize the autocorrelation at non-zero lags of the resulting excitation signal, since the excitation is assumed to be white noise. This, however, requires a non-linear minimization, and a practical alternative is to generate $\mathbf{u}$ as Gaussian white noise with variance $(E - E_{\text{LS}})/l$, where l is the number of corrupted samples. The resulting excitation will have approximately the desired energy and uncorrelated character. A suitable value for E is the expected excitation energy for the AR model, provided this is greater than E_{LS}, i.e. $E = max(E_{\text{LS}}, N\,\sigma_e^2)$. Viewed within a probabilistic framework, the case when $E = E_{\text{LS}} + l\sigma_e^2$, where l is the number of unknown sample values, is equivalent to drawing a sample from the posterior density for the missing data, $p(\mathbf{s}_{\mathcal{U}} \mid \mathbf{s}_{\mathcal{K}}, \mathbf{a}, \sigma_e^2)$. Figures 5-7 illustrate the principles involved in this sampled interpolation method. A short section taken from a modern solo vocal recording is shown in figure 5, alongside its estimated autoregressive excitation. The waveform has a fairly 'noise-like' character, and the corresponding excitation is noise-like as expected. The standard LSAR interpolation and corresponding excitation is shown in figure 6. The interpolated section (between the dotted vertical lines) is reasonable, but has lost the random noise-like quality of the original. Examination of the excitation signal shows that the LSAR interpolator has done 'too good' a job of minimizing the excitation energy, producing an interpolant which, while optimal in a mean-square error sense, cannot be regarded as typical of the autoregressive process. This might be heard as a momentary change in sound quality at the point of interpolation.

The sampling-based interpolator is shown in figure 7. Its waveform retains the random quality of the original signal, and likewise the excitation signal in the gap matches the surrounding excitation. Hence the sub-optimal interpolant is likely to sound more convincing to the listener than the LSAR reconstruction.

Ó Ruanaidh and Fitzgerald [112, 127] have successfully extended the idea of sampled interpolates to a full Gibbs' Sampling framework [51, 50] in order to generate typical interpolates from the marginal posterior density $p(\mathbf{s}_{\mathcal{U}} \mid \mathbf{s}_{\mathcal{K}})$. The method is iterative and involves sampling from the conditional posterior densities of $\mathbf{s}_{\mathcal{U}}$, $\mathbf{a}$ and σ_e^2 in turn, with the other unknowns fixed at their most recent sampled values. Once convergence has been achieved, the interpolation used is the last sampled estimate from $p(\mathbf{s}_{\mathcal{U}} \mid \mathbf{s}_{\mathcal{K}}, \mathbf{a}, \sigma_e^2)$.

3.3.2 Other methods

Several transform-domain methods have been developed for click replacement. Montresor, Valière and Baudry [101] describe a simple method for interpolating wavelet coefficients of corrupted audio signals, which involves substituting uncorrupted wavelet coefficients from nearby signal according to autocorrelation properties. This, however, does not ensure continuity of the restored waveform and is not a localized operation in the signal domain. An alternative method, based in the discrete Fourier domain, which is aimed at restoring long sections of missing data is presented by Maher [92]. In a similar manner to the sinusoidal coding algorithms of McAulay and Quatieri [97], this technique assumes that the signal is composed as a sum of sinusoids with slowly varying frequencies and amplitudes (see equation 2). Spectral peak 'tracks' from either side of the gap are identified from the Discrete Fourier Transform (DFT) of successive data blocks and interpolated in frequency and amplitude to generate estimated spectra for the intervening material. The inverse DFTs of the missing data blocks are then inserted back into the signal. The method is reported to be successful for gap lengths of up to 30ms, or well over 1000 samples at audio sampling rates. A method for interpolation of signals represented by the multiple sinusoid model is given in [150][Chapter 6].

Godsill and Rayner [59, 54] have derived an interpolation method which operates in the DFT domain. This can be viewed as an alternative to the LSAR interpolator (see section 3.3.1) in which power spectral density (PSD) information is directly incorporated in the frequency domain. Real and imaginary DFT components are modelled as independent Gaussians with variance proportional to the PSD at each frequency. These assumptions of independence are shown to hold exactly for random periodic processes [136], so the method is best suited to musical signals with strongly tonal content. The method can, however, also be used for other stationary signals provided that a sufficiently long block length is used (e.g. 500-2000 samples) since the assumptions also improve as block length increases [114]. The Maximum *a posteriori* solution is of a similar form and complexity to the LSAR interpolator, and is particularly useful as an alternative to the other method when the signal has a quasi-periodic or tonal character. A robust estimate is required for the PSD, and this can usually be

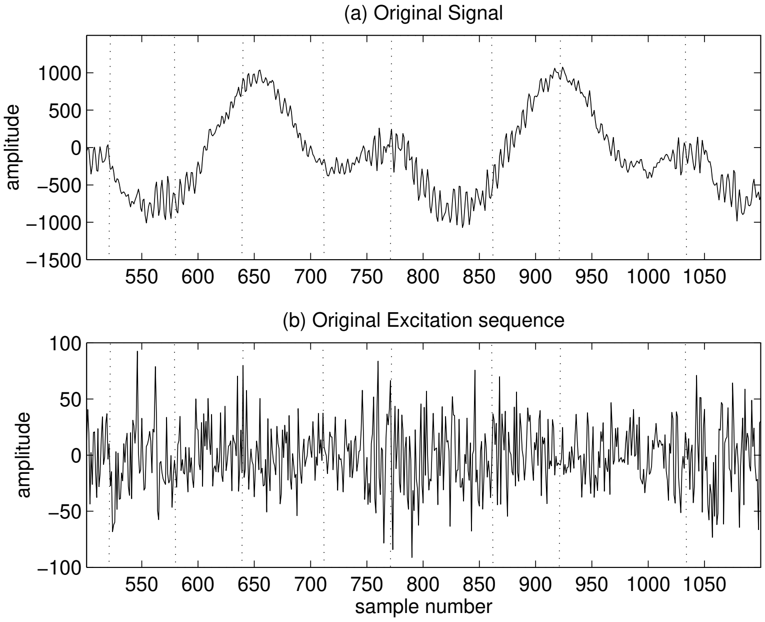

Figure 5: Original signal and excitation (P=100)

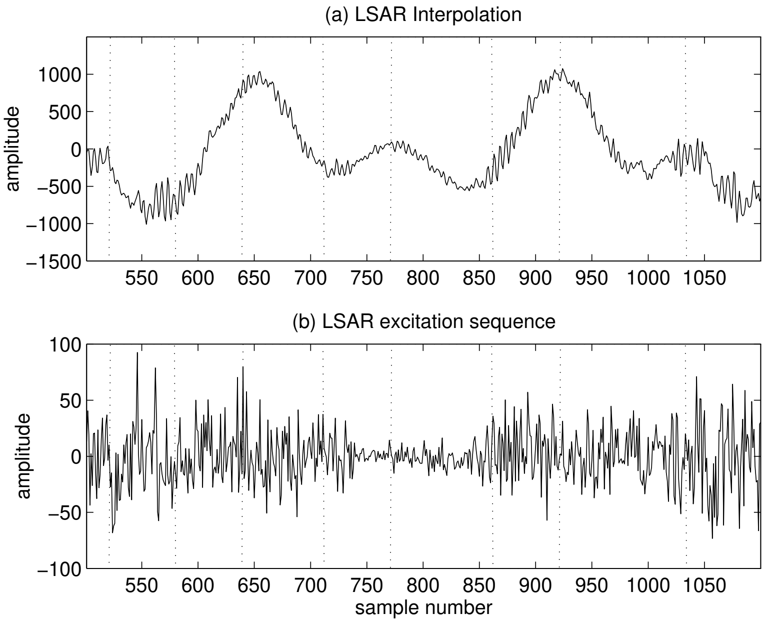

Figure 6: LSAR interpolation and excitation ($P = 100$)

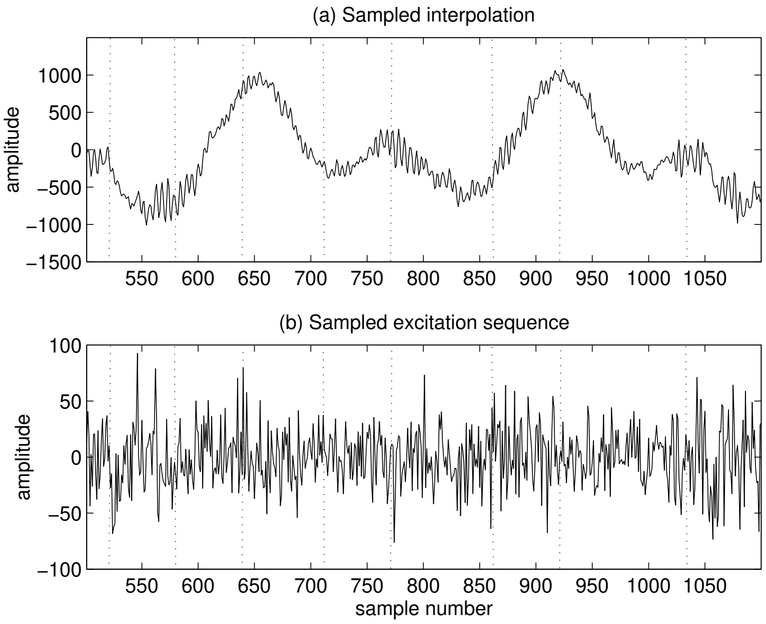

Figure 7: Sampled AR interpolation and excitation (P=100)

obtained through averaged DFTs of the surrounding data, although iterative methods are also possible, as in the case of the LSAR estimator.

Recent statistical model-based detection and interpolation methods are discussed in the next section.

3.4 Statistical methods for the treatment of clicks

The detection and replacement techniques described in the preceding sections can be combined to give very successful click concealment, as demonstrated by a number of research and commercial systems which are now used for the remastering of old recordings. However, some of the difficulties outlined above concerning the 'masking' of smaller defects by large defects in the detection process, the poor time localization of some detectors in the presence of impulse 'bursts' and the inadequate performance of existing interpolation methods for certain signal categories, has led to further research which considers the problem from a more fundamental statistical perspective.

In [58, 61, 54] click detection is studied within a model-based Bayesian framework (see e.g. [20, 12]). The Bayesian approach is a simple and elegant framework for performing decision and estimation within complex signal and noise modelling problems such as this, and relevant Bayesian approaches to the related problem of outlier detection in statistical data can be found in [19, 2, 98]. Detection is formulated explicitly as estimation of the noise 'switching' process

$i[n]$ (see section 3.1) conditional upon the corrupted data $x[n]$. The switching process can be regarded as a random discrete (1/0) process for which a posterior probability is calculated. Detection is then achieved by determining the switching values which minimize risk according to some appropriate cost function. In the most straightforward case, this will involve selecting switch values which maximize the posterior probability, leading to the maximum *a posteriori* (MAP) detection. The posterior detection probability for a block of N data points may be expressed using Bayes' rule as:

$$P(\mathbf{i} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{i})\, P(\mathbf{i})}{p(\mathbf{x})} \tag{11}$$

where all terms are implicitly conditional upon the prior modelling assumptions, $\mathcal{M}$. The prior detection probability $P(\mathbf{i})$ reflects any prior knowledge about the switching process. In the case of audio clicks this might, for example, incorporate the knowledge that clicks tend to occur as short 'bursts' of consecutive impulses, while the majority of samples are uncorrupted. A suitable prior which expresses this time dependence is the discrete Markov chain prior (see [61, 54] for discussion this point). The term $p(\mathbf{x})$ is constant for any given set of observations, and so can be ignored as a constant scale factor. Attention will thus focus on $p(\mathbf{x} \mid \mathbf{i})$, the detection-conditioned likelihood for a particular detection vector $\mathbf{i}$. It is shown in [61, 54, 58] that within the additive noise modelling framework of (3), the likelihood term is given by

$$p(\mathbf{x} \mid \mathbf{i}) = \int_{\mathbf{s}_{\mathcal{U}}} p_{\mathbf{v}_{\mathcal{U}}|\mathbf{i}}(\mathbf{x}_{\mathcal{U}} - \mathbf{s}_{\mathcal{U}} \mid \mathbf{i})\, p_{\mathbf{s}}(\mathbf{s}) \,|_{\mathbf{s}_{\mathcal{K}}=\mathbf{x}_{\mathcal{K}}}\; d\mathbf{s}_{\mathcal{U}} \tag{12}$$

where $p_{\mathbf{v}_{\mathcal{U}}|\mathbf{i}}$ is the probability density function for the corrupting noise values and $p_{\mathbf{s}}$ is the density for the underlying audio data. This formulation holds for any random additive noise process which is independent of the signal. In particular, the calculation of (12) is analytic in the case of linear Gaussian models. In [58, 61, 54] the autoregressive signal model with Gaussian corruption is studied in detail.

In order to obtain the MAP detection estimate from the posterior probability expression of equation (11) an exhaustive search over all 2^N possible configurations of the (1/0) vector $\mathbf{i}$ is necessary. This is clearly infeasible for any useful value of N, so alternative strategies must be devised. A sequential approach is developed in [61, 54] for the Gaussian AR case. This is based around a recursive calculation of the likelihood (12), and hence posterior probability, as each new data sample is presented. The sequential algorithm performs a reduced binary tree search through possible configurations of the detection vector, rejecting branches which have low posterior probability and thus making considerable computational savings compared with the exhaustive search. The method has been evaluated experimentally in terms of detection error probabilities and perceived quality of restoration and found to be a significant improvement over the autoregressive detection methods described in section 3.2.1, although more computationally intensive.

Click detection within a Bayesian framework has introduced the concept of an explicit model for the corrupting noise process through the noise density $p_{\mathbf{v}_{\mathcal{U}}|\mathbf{i}}$. Effective noise modelling can lead to improvements not only in click detection, but also in replacement, since it allows useful signal information to be extracted from the corrupted data values. This information is otherwise discarded as irrevocably lost, as in the interpolators described in earlier sections. In fact, it transpires that an intrinsic part of the likelihood calculation in the Bayesian detection algorithm (equation 12) is calculation of the MAP estimate for the unknown data conditional upon the detection vector $\mathbf{i}$. This MAP interpolation can be used as the final restored output after detection, without resort to other interpolation methods. The form of this 'interpolator' is closely related to the LSAR interpolator (section 3.3.1) and may be expressed as:

$$\mathbf{s}_{\mathcal{U}}^{\mathrm{MAP}} = -\left({\mathbf{A}_{\mathcal{U}}}^T \mathbf{A}_{\mathcal{U}} \,+\, \frac{\sigma_e^2}{\sigma_v^2}\,\mathbf{I} \right)^{-1} \left({\mathbf{A}_{\mathcal{U}}}^T \mathbf{A}_{\mathcal{K}}\,\mathbf{s}_{\mathcal{K}} \,-\, \frac{\sigma_e^2}{\sigma_v^2}\,\mathbf{x}_{\mathcal{U}} \right), \tag{13}$$

(see [61][equations (12-14)]), where σ_v^2 is the variance of the corrupting noise, which is assumed independent and Gaussian. Of course, the quality of the restored output is now dependent on the validity of the assumed noise statistics. The Bayesian detector itself shows considerable robustness to errors in these assumptions [61, 54], but the interpolator is less tolerant. This will be particularly noticeable when the true noise distributions are more 'heavy-tailed' than the Gaussian, a scenario for which there is strong evidence in many degraded audio signals. The noise modelling can in fact be generalized to a more realistic class of distributions by allowing the individual noise components $v[n]$ to have separate, unknown variances and even unknown correlation structure. We are essentially then modelling noise sources as continuous scale mixtures of Gaussians:

$$p(v[n]) = \int N(0,\lambda) g(\lambda) d\lambda$$

where $N(\mu,\lambda)$ is the Gaussian distribution with mean μ and variance λ, and $g(\lambda)$ is a continuous 'mixing' density [152]. These extensions allow for non-Gaussian defects with of widely varying magnitude and also for the noise correlation which might be expected when the signal has been played through a mechanical pick-up system followed by equalization circuitry. This noise modelling framework can be used to develop highly robust interpolators, and a Bayesian approach which requires no prior knowledge of AR parameters or noise statistics is presented in [62], using an iterative EM-based solution. Similar noise modelling principles can be used to extend the Bayesian detection algorithms, and Markov chain Monte Carlo (MCMC) methods [69, 51, 50] are presented for the solution of this problem in [63, 65]. An example of these Bayesian iterative restoration methods for removal of clicks is shown in figure 8 for a typical 78rpm recording. The same framework may be extended to perform joint removal of clicks and background noise in one single procedure, and some recent work on this problem can be found in [64] for autoregressive signals and in [56, 57] for autoregressive moving-average (ARMA) signals.

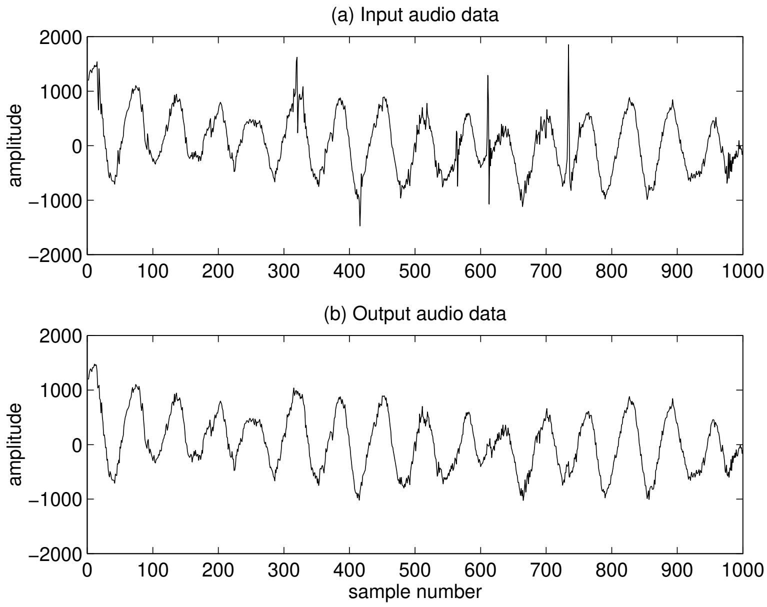

Figure 8: Restoration using Bayesian iterative methods

The statistical methods described here provide a highly flexible framework for audio restoration and signal enhancement in general. Solution for these complex models is usually of significantly higher computational complexity than the techniques described in earlier sections, but this is unlikely to be problematic for applications where restoration quality is the highest priority. The methods are still in their infancy, but we believe that future research work in the field will require sophisticated statistical modelling of signals and noise, with associated increases in solution complexity, in order to achieve improved fidelity of restoration. The Bayesian methods discussed here are likely to find application in many other areas of audio processing (see later sections).

4 Correlated Noise Pulse Removal

A further problem which is common to several recording media including gramophone discs and optical film sound tracks is that of low frequency noise pulses. This form of degradation is typically associated with large scratches or even breakages in the surface of a gramophone disc. The precise form of the noise pulse depends upon the mechanical and electrical characteristics of the playback system, but a typical result is shown in figure 9. A large discontinuity is observed followed by a decaying low frequency transient. The noise pulses appear to be additively superimposed on the undistorted signal waveform (see figure 10).

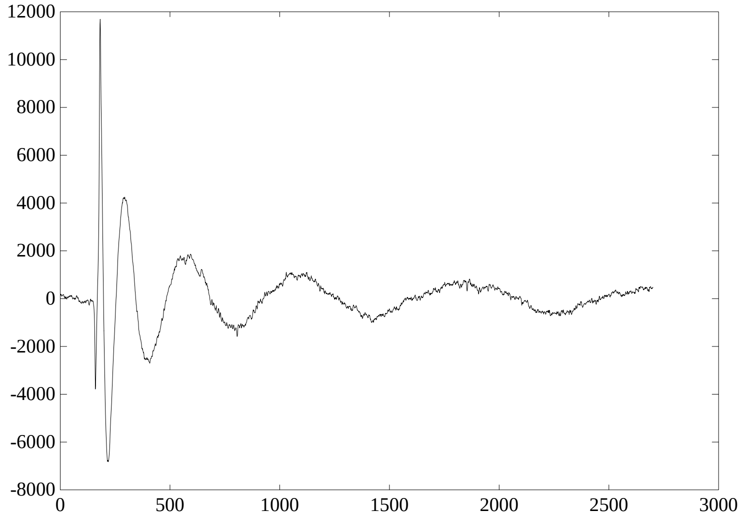

Figure 9: Noise pulse from optical film sound track ('silent' section)

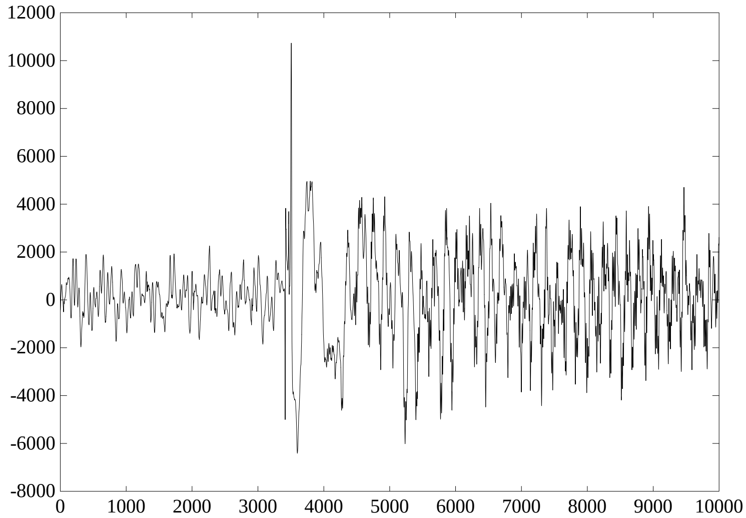

Figure 10: Signal waveform degraded by low frequency noise transient

Low frequency noise pulses appear to be the response of the playback system to extreme step-like or impulsive stimuli caused by breakages in the groove walls of gramophone discs or large scratches on an optical film sound track. The audible effect of this response is a percussive 'pop' indexNoise! Pop noise or 'thump' in the recording. This type of degradation is often the most disturbing artefact present in a given extract. It is thus highly desirable to eliminate noise pulses as a first stage in the restoration process.

The effects of the noise pulse are quite long-term, as can be seen from figure 9, and thus a straightforward interpolation using the methods of section 3.3 is not a practical proposition. Since the majority of the noise pulse is of very low frequency it might be thought that some kind of high pass filtering operation would remove the defect. Unfortunately this does not work well either, since the discontinuity at the start of the pulse has significant high frequency content. Some success has been achieved with a combination of localized high pass filtering , followed by interpolation to remove discontinuities. However it is generally found that significant artefacts remain after processing or that the low frequency content of the signal has been damaged.

It should be noted that the problem of transient noise pulses can in principle be circumvented by use of suitable playback technology. For example, in the case of gramophone disks the use of a laser-based reader should eliminate any mechanical resonance effects and thus reduce the artefact to a large click which can be restored using the methods of previous sections. Of course, this does not help in the many cases where the original source medium has been discarded after transcription using standard equipment to another medium such as magnetic tape!

4.0.1 Template-based methods

The first digital approach to this problem was devised by Vaseghi and Rayner [145, 147]. This technique, which employs a 'template' for the noise pulse waveform, has been found to give good results for many examples of broken gramophone discs. The observation was made that the resonant sections (i.e. after the initial discontinuity) of successive noise pulses in the same recording were nearly identical in shape (to within a scale factor). This would correspond with the idea that noise pulses are simply the step response of a linear time-invariant (LTI) mechanical system. Given the waveform of the repetitive section of the noise pulse (the 'template' $t[n]$) it is then possible to subtract appropriately scaled versions from the corrupted signal $x[n]$ wherever pulses are detected. The position M and scaling G of the noise pulse are estimated by cross-correlating the template with the corrupted waveform, and the restored signal is then obtained as:

$$y[n] = x[n] \; - \; G\, t[n-M], \qquad M \leq n < M + N_t \tag{14}$$

where N_t is the length of the template. Any remaining samples close to the start of the pulse which are irrevocably distorted can then be interpolated using a method such as the LSAR interpolator discussed earlier (see section 3.3.1).

The template $t[n]$ is obtained by long term averaging of many such pulses from the corrupted signal. Alternatively, a noise-free example of the pulse shape may be available from a 'silent' section of the recording or the lead-in groove of a gramophone disc.

The template method has been very successful in the restoration of many recordings. However, it is limited in several important ways which hinder the complete automation of pulse removal. While the assumption of constant template shape is good for short extracts with periodically recurring noise pulses (e.g. in the case of a broken gramophone disc) it is not a good assumption for many other recordings. Even where noise pulses do correspond to a single radial scratch or fracture on the record the pulse shape is often found to change significantly as the recording proceeds, while much more variety is found where pulses correspond to randomly placed scratches and breakages on the recording. Further complications arise where several pulses become superimposed as is the case for several closely spaced scratches. These effects may be partly due to the time-varying nature of the mechanical system as the stylus moves towards the centre of the disk, but also non-linearity in the playback apparatus. There is some evidence for the latter effect in optical film sound track readers [54], where the frequency of oscillation can be observed to decrease significantly as the response decays.

Correct detection can also be a challenge. This may seem surprising since the defect is often very large relative to the signal. However, audible noise pulses do occur in high amplitude sections of the signal. In such cases the cross-correlation method of detection can give false alarms from low frequency components in the signal; in other circumstances noise pulses can be missed altogether. This is partly as a result of the correlated nature of the signal which renders the cross-correlation procedure sub-optimal. A true matched filter for the noise pulse would take into account the signal correlations (see e.g. [143]) and perhaps achieve some improvements in detection. This issue is not addressed here, however, since other restoration methods are now available.

4.0.2 Model-based separation methods

A full study of the noise pulse mechanism would involve physical modelling of the (possibly non-linear) playback system for both gramophone systems and optical sound track readers. A full description is beyond the scope of this article (see [126] for some more detail), but can be used to shed further light upon this and other audio restoration areas including click removal and background noise reduction.

A linear modelling approach to noise pulse removal is presented in [54]. In this it is assumed that the corrupted waveform $\mathbf{x}$ consists of a linear sum of the underlying audio waveform $\mathbf{s}$ and resonant noise pulses $\mathbf{v}$:

$$\mathbf{x} = \mathbf{s} + \mathbf{v}. \tag{15}$$

We note that $\mathbf{s}$ and $\mathbf{v}$ are the responses of the playback system, including mechanical components and amplification/ equalization circuitry, to the recorded audio and noise signals, respectively. The assumption of a linear system allows the overall response $\mathbf{x}$ to be written as the linear superposition of individual responses to signal and noise components.

Here the noise pulses are modelled by a low order autoregressive process which is driven by a low level white noise excitation with variance σ_{v0}^2 most of the time, and bursts of high level impulsive excitation with variance $\sigma_{v1}^2 \gg \sigma_{v0}^2$ at the initial discontinuity of the noise transient. We can define a binary noise switching process $i[n]$ to switch between low and high variance components in a similar way to the click generation model of section 3. This modelling approach is quite flexible in that it allows for variations in the shape of individual noise pulses as well as for the presence of many superimposed pulses within a short period. The restoration task is then one of separating the two superimposed responses, $\mathbf{s}$ and $\mathbf{v}$. If the audio signal's response is also modelled as an autoregressive process then the MAP estimator for $\mathbf{s}$ under Gaussian assumptions is obtained from:

$$\left(\frac{\mathbf{A}^T\mathbf{A}}{\sigma_e^2} \; + \; {\mathbf{A}_v}^T {\mathbf{\Lambda}_v}^{-1} \mathbf{A}_v\right) \mathbf{s}^{\text{MAP}} = {\mathbf{A}_v}^T {\mathbf{\Lambda}_v}^{-1} \mathbf{A}_v \; \mathbf{x}. \tag{16}$$

Terms of this equation are defined similarly to those for the LSAR interpolator of section 3.3.1, with subscript 'v' referring to the autoregressive process for the noise pulses. $\mathbf{\Lambda}_v$ is a diagonal matrix whose mth diagonal element $\lambda_v[m]$ is the variance of the mth noise excitation component, i.e.

$$\lambda_v[m] = \sigma_{v0}^2 \; + \; i[m] \, (\sigma_{v1}^2 \; - \; \sigma_{v0}^2). \tag{17}$$

This signal separation algorithm requires knowledge of both AR systems, including noise variances and actual parameter values, as well as the switching vector $\mathbf{i}$ which indicates which noise samples have impulsive excitation and which have low level excitation. These can be treated as unknowns within a similar iterative statistical framework to that outlined for click removal in section 3.4, and this could form a useful piece of future research. In practice, however, these unknowns can usually be estimated by simpler means. The switching process can be estimated much as clicks are detected (see section 3.2), with a higher threshold selected to indicate large disturbances which are likely to be noise pulses. The autoregressive system for the noise can often be estimated from a noise pulse captured during a 'silent' section of the recording or from a similar type of pulse taken from another recording, and the very large (or even infinite) value chosen for the high level excitation variance σ_{v1}^2. The autoregressive system for the underlying audio data is then estimated from uncorrupted data in the vicinity of the noise pulses, usually in the section just prior to the start of the section to be restored.

Even with suitable estimation schemes for the unknown parameters, the separation formula of equation (16) is of relatively high computational complexity, since the noise process can affect thousands of samples following the initial impulsive discontinuity. This problem can be partially overcome by restoring samples which are fairly distant from the initial transient using a simple linear phase high-pass filter. The separation algorithm is then constrained to give continuity with this filtered signal at either end of the restored section in much the same way as the LSAR interpolator (section 3.3.1). Further computational savings can be achieved by working with a sub-sampled version of the noise pulse waveform, since it is typically over-sampled by a factor of at least one hundred for the much of its duration. This sub-sampling can be incorporated into the separation algorithm by use of an analytic interpolation operator such as the second order spline. An alternative scheme, which takes advantage of the Markovian nature of the AR models, is based on Kalman filtering [5]. This is currently being investigated and results will be reported in future publications.

Results from the model-based separation approach have demonstrated much more generality of application and ease of automation than the templating technique, which can be a highly operator-intensive procedure, and the perceived quality of output is certainly at least as good as the templating method. Figures 11-13 show the restoration of a particularly badly degraded 78rpm recording which exhibits many closely spaced noise transients. A second order autoregressive model was found to be adequate for modelling the noise transients, while the signal was modelled to order 80. The restored signal (shown on a different scale) shows no trace of the original corruption, and the perceptual results are very effective.

4.0.3 Summary

Two principal methods for removal of low frequency noise transients are currently available. The model-based separation approach has shown more flexibility and generality, but is computationally rather intensive. It is felt that future work in the area should consider the problem from a realistic physical modelling perspective, which takes into account linear and non-linear characteristics of gramophone and film sound playback systems, in order to detect and correct these artefacts more effectively. Such an approach could involve both experimental work with playback systems and sophisticated non-linear modelling techniques. Statistical approaches related to those outlined in the click removal work (section 3.4) may be applicable to this latter task.

5 Background noise reduction

Random, additive background noise is a form of degradation common to all analogue measurement, storage and recording systems. In the case of audio signals the noise, which is generally perceived as 'hiss' by the listener, will be composed of electrical circuit noise, irregularities in the storage medium and ambient noise from the recording environment. The combined effect of these sources will generally be treated as one single noise process, although we note that a pure restoration should strictly not treat the ambient noise, which might be considered as a part of the original 'performance'. Random noise generally has significant components at all audio frequencies, and thus simple filtering

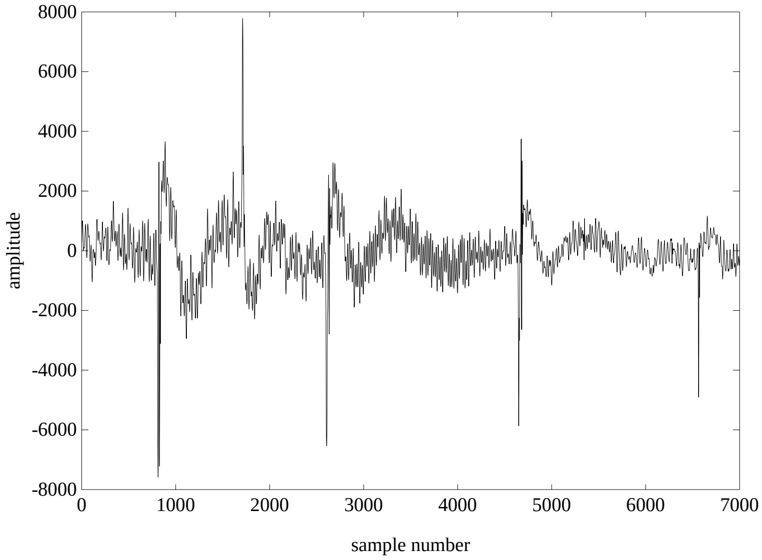

Figure 11: Degraded audio signal with many closely spaced noise transients

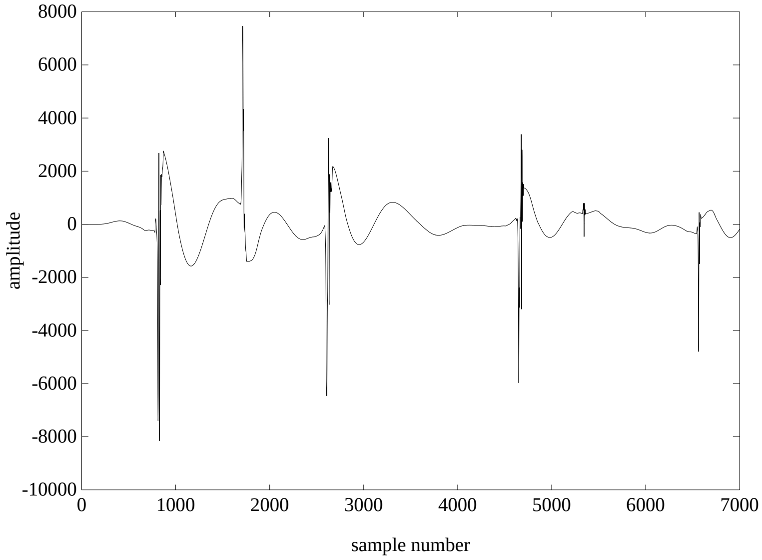

Figure 12: Estimated noise transients for figure 11

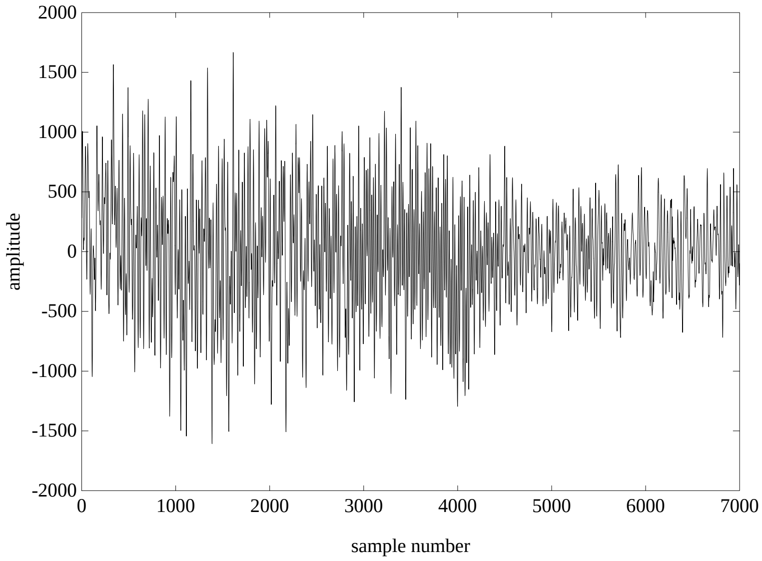

Figure 13: Restored audio signal for figure 11 (different scale)

and equalization procedures are inadequate for restoration purposes.

Analogue tape recordings typically exhibit noise characteristics which are stationary and for most purposes white. At the other end of the scale, many early 78rpm and cylinder recordings exhibit highly non-stationary noise characteristics, such that the noise can vary considerably within each revolution of the playback system. This results in the characteristic 'swishing' effect associated with some early recordings. In recording media which are also affected by local disturbances, such as clicks and low frequency noise resonances, standard practice is to restore these defects prior to any background noise treatment.

Noise reduction has been of great importance for many years in engineering disciplines. The classic least-squares work of Norbert Wiener [153] placed noise reduction on a firm analytic footing, and still forms the basis of many noise reduction methods. In the field of speech processing a large number of techniques has been developed for noise reduction, and many of these are more generally applicable to noisy audio signals. We do not attempt here to describe every existing method in detail, since these are well covered in speech processing texts (see for example [88, 85, 17]). We do, however, discuss some standard approaches which are appropriate for general audio signals and emerging techniques which are likely to be of use in future work. It is worth mentioning that where methods are derived from speech processing techniques, as in for example the spectral attenuation methods of section 5.1, sophisticated modifications to the basic schemes are required in order to match the stringent fidelity requirements and signal characteristics of an audio restoration system.

Certainly the most popular methods for noise reduction in audio signals to date are based upon short-time Fourier processing. These methods, which can be derived from non-stationary adaptations to the frequency-domain Wiener filter, are discussed fully in section 5.1.

Within a model-based framework, Lim and Oppenheim [87] studied noise reduction using an autoregressive signal model, deriving iterative MAP and ML procedures. These methods are computationally intensive, although the signal estimation part of the iteration is shown to have a simple frequency-domain Wiener filtering interpretation (see also [113, 79] for Kalman filtering realizations of the signal estimation step). It is felt that new and more sophisticated model-based procedures may provide noise reducers which are competitive with the well-known short-time Fourier based methods. In particular, modern statistical methodology for solution of complex problems (for example, the Markov chain Monte-Carlo (MCMC) methods discussed in section 3.4 for click removal) allows for more realistic signal and noise modelling, including non-Gaussianity, non-linearity and non-stationarity. Such a framework can also be used to perform joint restoration of both clicks and random noise in one single process. A Bayesian approach to this joint problem using an autoregressive signal model is described in [54][section 4.3.2] and [64] and in [56, 57] for the more general autoregressive moving average (ARMA) model,. In addition, [107, 108] present an extended Kalman filter for joint removal of noise and clicks from AR- and ARMA-modelled audio signals.

Other methods which are emerging for noise reduction include the incorporation of psychoacoustical masking properties of human hearing [23, 24, 140] and noise reduction in alternative basis expansions, in particular the wavelet domain [11] and sub-space representations [36, 45, ?]. These approaches address various short-comings of existing noise-reduction procedures, and could thus lead to improvements over existing techniques.

5.1 Background noise reduction by short-time spectral attenuation

This section deals with a class of techniques known as Short-Time Spectral Attenuation (STSA) [5]. STSA is a single input noise reduction method that basically consists in applying a time-varying attenuation to the short-time spectrum of the noisy signal. STSA techniques are non-parametric and generally need little knowledge of the signal to be processed. They rank among the most popular methods for speech enhancement and their use has been widely predominant for the restoration of musical recordings.

[5] These techniques are also often referred to as 'spectral subtraction'. We will not use this terminology in order to avoid ambiguities between the general principle and the particular technique described in [15], nor will we use the term 'spectral estimation' as quite a number of the STSA techniques are not based on a statistical estimation approach.

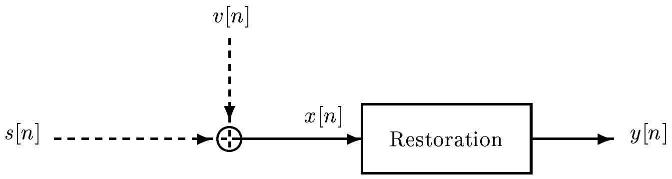

Figure 14: Modeled restoration process

5.1.1 General overview

Hypotheses Figure 14 shows the basic hypotheses common to all short-time spectral attenuation techniques. It is supposed that the original audio signal $s[n]$ has been corrupted by an additive noise signal $v[n]$ uncorrelated with $s[n]$ and that the only observable signal is the degraded signal $x[n]$ [85]. In the field of audio, restoration techniques applicable in such a situation are sometimes referred to as non-complementary [38] or one-ended [47] to differentiate them from a class of frequently used denoising methods which rely on some pre-processing of the signal prior to the degradation (see [38]).

The knowledge concerning the noise is usually limited to the facts that it can be considered as stationary and that it is possible to estimate its power spectral density (or quantities that are directly related to it) [89, 86].

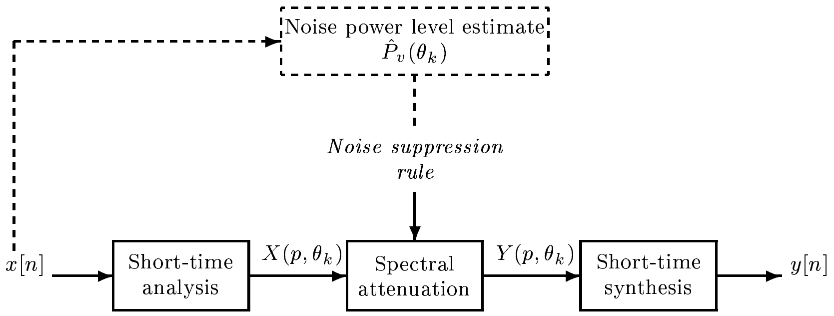

Figure 15: Background noise suppression by short-time spectral attenuation

Principle Figure 15 shows the general framework of short-time spectral attenuation: the first step consists in analyzing the signal with a (in general, multirate) filter bank, each channel of the filter-bank is then attenuated (multiplied by a real positive gain, generally smaller than 1), and finally the sub-band signals are put back together to obtain the restored signal. The time-varying gain to be applied in each channel is determined by the so called noise suppression rule [96, 144] which usually relies on an estimate of the noise power in each channel (represented by the dotted part of Figure 15). The two elements that really characterize a particular STSA technique are the filter-bank characteristics and the suppression rule.

In most STSA techniques the short-time analysis of the signal is performed by use of the Short-Time Fourier Transform (STFT) [89, 16, 43, 103], or with a uniform filter-bank that can be implemented by STFT [131, 144, 80]. Note that in such cases the two interpretations (multirate filter-bank, and short-time Fourier transform) can be used interchangeably as they are fully equivalent [33]. Examples of STSA techniques based on the use of non-uniform filter banks can be found in [115, 96].

In designing the filter-bank, it is necessary to bear in mind the fact that the sub-band signals will sometimes be strongly modified by the attenuation procedure. As a consequence, while it is of course desirable to obtain a (nearly) perfect reconstruction in the absence of modification, it is also important to avoid effects such as sub-band spectral aliasing which could create distortions in the restored signal [33]. With the short-time Fourier transform, satisfying results are obtained with a sub-band sampling rate two or three times higher than the critical-sampling rate (ie. with a 50% to 66% overlap between successive short-time frames) [25].

Historical considerations Historically, short-time spectral attenuation was first developed for speech enhancement during the 1970s [89, 16, 131]. The application of STSA to the restoration of audio recordings came afterwards [80, 103, 145, 146, 142, 47] with techniques that were generally directly adapted from earlier speech-enhancement techniques.

Prior to works such as [4] and [34], there was not necessarily an agreement about the equivalence of the filter-bank and the STFT approaches (see also [33]). Traditionally, the filter-bank interpretation is more intuitive for audio engineers [80, 103, 47] while the short-time spectrum is typically a speech analysis notion [89]. Also controversial is the problem of short-time phase: in the STFT interpretation, the short-time attenuation corresponds to a magnitude-only modification of the short-time spectrum. The fact that only the magnitude of the short-time spectrum is processed has been given various interpretations, including an experimental assessment for speech signals in [151].

The most widespread opinion is that the phase need not be modified because of the properties of the human auditory system [89]. Strictly speaking however, the assertion that the ear is "insensitive to the phase" was highlighted by psychoacoustic findings only in the case of stationary sounds and for the phase of the Fourier transform [102]. Moreover, it is well known that in the case of STFT, *phase variations* between successive short-time frames can give rise to audible effects (such as frequency modulation) [144].

It should however be emphasized that there is usually no choice but to keep the unmodified phase because of the lack of hypotheses concerning the unknown signal (recall that only the second order statistics of the signal and noise are supposed to be known). This is well known for techniques derived from Wiener filtering (time-domain minimum mean squared error filtering), and a similar result is proved in [43] for a frequency domain criterion (using strong hypotheses concerning the independence of the short-time transform bins). Although other criteria could be used, these results indicate that it may be be difficult to outperform the standard magnitude attenuation paradigm without introducing more elaborate hypotheses concerning the behavior of the signal.

Scope of the method Until now, STSA techniques have been largely predominant in the field of speech enhancement and appear to have been used almost exclusively for the restoration of musical recordings.

One of the reasons for this wide application of STSA techniques is certainly the fact that they correspond to a non-parametric approach which can be applied to a large class of signals. By contrast, considering that most music recordings contain several simultaneous sound sources, it is unlikely that some of the methods relying on very specific knowledge of the speech signal properties (such as the model-based speech enhancement techniques [16, 41]) could be generalized for audio restoration.

Another reason for the success of STSA techniques in the field of audio engineering is maybe the fact that they have a very intuitive interpretation: they extend to a large number of sub-bands the principle of well known analog devices used for signal enhancement, such as the noise gate [103] (see also [47] for a link with compandors).

5.1.2 Suppression rules

Let $X(p, \theta_k)$ denote the short-time Fourier transform of $x[n]$, where p is the time index, and θ_k the normalized frequency index (θ_k lies between 0 and 1 and takes N discrete values for $k = 1, \ldots, N$, N being the number of sub-bands). Note that the time index p usually refers to a sampling rate lower than the initial signal sampling rate (for the STFT, the down-sampling factor is equal to hop-size between to consecutive short-time frames) [33].

The result of the noise suppression rule can always be interpreted as the application of a real gain $G(p, \theta_k)$ to each bin of the short-time transform $X(p, \theta_k)$ of the noisy signal. Usually, this gain corresponds to an 'attenuation', ie. lies between 0 and 1. For most suppression rules, $G(p, \theta_k)$ depends only on the power level of the noisy signal measured at the same bin $|X(p, \theta_k)|^2$ and on an estimate of the power of the noise at the frequency θ_k, $\hat{P}_v(\theta_k) = E\{|V(p, \theta_k)|^2\}$ (which does not depend on the time index p because of the noise stationarity). In the following, the ratio $Q(p, \theta_k) = |X(p, \theta_k)|^2 / \hat{P}_v(\theta_k)$ will be referred to as the relative signal level. Note that since the noise $v[n]$ is un-correlated with the unknown signal $s[n]$, we have

$$E\{Q(p, \theta_k)\} = 1 + \frac{E\{|X(p, \theta_k)|^2\}}{\hat{P}_v(\theta_k)} \tag{18}$$

so that the expected value of the relative signal level is always larger than 1.

Standard examples of noise-suppression rules include the so-called Wiener[6] suppression rule, the power-subtraction (see Figure 16), the spectral subtraction

[6]This suppression rule is derived by analogy with the well-known Wiener filtering formula replacing the power spectral density of the noisy signal by its periodogram estimate.

[15, 89, 96, 144], as well as several families of parametric suppression curves [89, 103, 47].

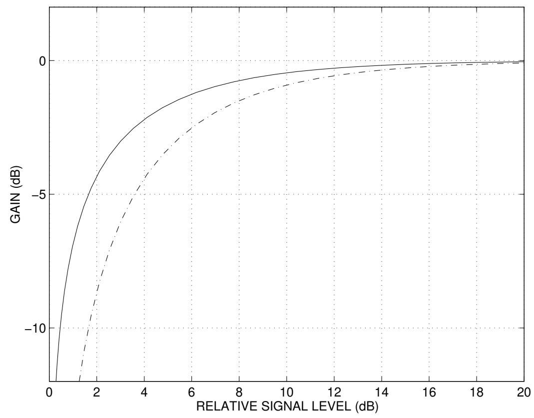

Figure 16: Suppression rules characteristics : gain (dB) versus relative signal level (dB). Solid line: Power subtraction; dashed line: Wiener.

.

All the suppression rules mentioned above share the same general behavior in that $G(p, \theta_k) = 1$ when the relative signal level is high ($Q(p, \theta_k) >> 1$), and

$$\lim_{Q(p,\theta_k) \to 1} G(p, \theta_k) = 0$$

In many cases, the noise level $\hat{P}_v(\theta_k)$ is artificially over-estimated (multiplied by a factor $\beta > 1$) so that $G(p, \theta_k)$ is null as soon as $Q(p, \theta_k) \leq \beta$ [89, 103].

Reference [16] presents a detailed review of suppression rules that are derived from a Bayesian point of view supposing a prior knowledge of the probability distribution of the sub-band signals. These suppression rules are more elaborate in the sense that they generally depend both on the relative signal level (or a quantity directly related to it) and on a characterization of the a priori information (a priori probability of speech presence in [96], a priori signal-to-noise ratio in [43]).

Finally, some suppression rules used for speech enhancement do not require any knowledge of the noise characteristics [22, 30]. These techniques, designed for improving speech intelligibility, can hardly be generalized to the case of audio recordings since they generate non-negligible distortions of the signal spectrum regardless of the noise level.

5.1.3 Evaluation

'Deterministic' analysis While it is rather difficult to analyze the results of STSA techniques in a general case, it becomes relatively simple when it is supposed that the unknown input signal is a pure tone, or more generally, a compound of several pure tones with frequencies sufficiently spaced apart. This hypothesis is pertinent since a large proportion of steady instrumental sounds can be efficiently described, both perceptively and analytically, as a sum of slowly modulated pure tones [37, 10, 68].

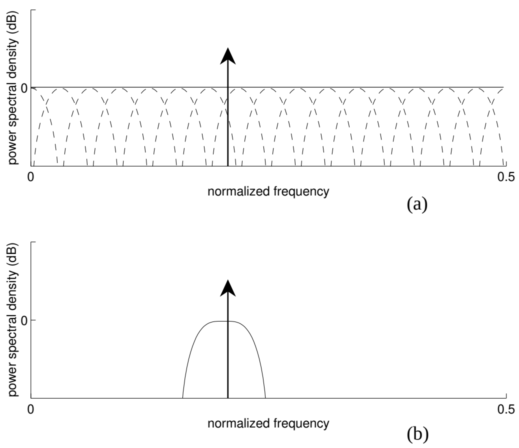

Figure 17: Restoration of a sinusoidal signal embedded in white noise (of power 0 dB). (a) The noisy signal (the dotted lines feature the filter bank characteristics); (b) The processed signal.

As pointed out in [80], short-time spectral attenuation does not reduce the noise present in the sub-bands that contain signal components. Figure 17 shows an illustration of this fact for a sinusoidal signal embedded in white noise: if the level of the sinusoid is large enough, the channels in which it lies are left unattenuated while the other channels are strongly attenuated. As a consequence the output signal consists of the sinusoidal signal surrounded by a narrow band of filtered noise. Note also that if the sinusoid level is too low, all the channels are strongly attenuated and the signal is completely cancelled.

Cancelling of the signal For a sinusoidal signal of frequency θ (which is supposed to correspond to the center frequency of one of the filters of the filter-bank), it is easily checked (assuming that the additional noise power spectral density is sufficiently smooth) that Eq. 18 becomes

$$E\{Q(p,\theta)\} = 1 + \frac{P_s}{S_v(\theta)W_\theta} \tag{19}$$

where P_s is the power of the sinusoid, $S_v(\theta)$ the power spectral density of the noise at frequency θ and W_θ is the bandwidth of the sub-band filter centered around frequency θ (see [27] for a demonstration in the STFT case).

As a consequence, *the level of the signal components that are mistakenly cancelled by the restoration process increases with the bandwidth of the analyzing* *filter-bank*. Although deceptively simple, this results nonetheless states that the signal enhancement can be made more efficient by sharpening the channel bandwidth as much as allowed by the stationarity hypothesis.

For the STFT case, the bandwidth of the filter-bank is inversely proportional to the duration of the short-time frame and it is shown in [27], using standard results concerning the simultaneous frequency masking phenomenon, that the processing can suppress audible signal components (ie. components that were not masked by the additive noise) if the short-time duration is well below 40 ms.

Audibility of the noise at the output In the case were the signal component is not cancelled, the processed signal exhibits a band of filtered noise located around the sinusoidal component. It is clear that this phenomenon, if audible, is an important drawback of the method because it makes the remaining nuisance noise correlated with the signal, which was not the case for the original broad-band noise.

It is shown in [27] for the STFT case, that this effect is only perceptible when the frame duration is short (smaller than 20-30 ms)[7].

As for the results mentioned previously concerning signal cancellation, the obtained audibility limit should only be considered as an order of magnitude estimate in real situations since it does not take into account the possible mutual masking between different signal components (a phenomenon which may prevail when the noise level is very low) [27]. These results still support several earlier experimental findings [103, 142] concerning the influence of the STFT window duration in STSA techniques. In practice, the STFT frame duration should be sufficiently long to avoid creating undesirable modulation effects [103] (audible band of noise remaining around signal components). Moreover, for audio signals, the duration of the short-time frame can generally be set advantageously to larger values than those used for speech [142] (because it lowers the limit of signal cancellation).

Transient signals The previous results are related to the case of steady portions of musical sounds, however it is well-known that musical recordings also feature many transient parts (note onsets, percussions) that play an important part in the subjective assessment of the signal characteristics [68, 37].

As with many other techniques which make use of a short-time signal analyzer, it is possible to observe specific signal distortions, generated by the restoration process, which occur when transient signals are present [75]. In STSA techniques the distortion manifests itself as a smearing of the signal waveform for low-level signal transients. This phenomenon as well as its perceptive consequences are amplified as the short-time frame duration increases [142, 27, 111].

The analysis of such transient effects is made more difficult by the fact that there is no 'prototype' transient signal as simple and as pertinent as the pure

[7] Strictly speaking, this effect could still be perceived for longer window durations when the relative signal level approaches 1. However, it is then perceived more like an erratic fluctuation of the sinusoid level which is hardly distinguishable from the phenomenon to be described in section 5.1.3.

tone was for steady sounds. However, the results obtained in a simplistic case (the abrupt onset of a pure tone) seem to indicate that the observed smearing of the transient part of low level signals is mainly due to the modification of the signal spectrum by the suppression rule [27]. This is in contrast with what happens in applications where the magnitude of the short-time spectrum is not drastically modified, such as time-scaling with STFT, where the smearing of transient signals is mostly caused by the phase distortions [67, 111].

As a consequence, methods that exploit the redundancy of the magnitude of the short-time spectra to restore a 'correct' phase spectrum [67, 106] are not efficient in eliminating the transient distortions caused by STSA.

Consequences of the random nature of the attenuation In the previous section we deliberately left apart a major problem: the fact that the attenuation is a random quantity. The randomness of the attenuation comes from the fact that it is (in general) determined as a function of the relative signal level which in turn involves the short-time transform of the noisy signal. This aspect plays a key role in STSA because the relative signal level is estimated by the periodogram (at least in the STFT case) characterized by a very high variance.

A well known result states that the values of the discrete Fourier transform of a stationary random process are normally distributed complex variables when the length of the Fourier transform is large enough (compared to the decay rate of the noise correlation function) [21]. This asymptotic normal behavior leads to a Rayleigh distributed magnitude and a uniformly distributed phase (see [96, 43] and [114]).

Using the normality assumption, it is shown in [27] that the probability density of the relative signal level Q (omitting the two indexes p and θ_k) is

$$f(Q) = e^{-[Q+(\bar{Q}-1)]}\, I_0\left(2\sqrt{Q\,(\bar{Q}-1)}\right) \tag{20}$$

where $I_o(x)$ denotes the modified Bessel function of order 0, and $\bar{Q}$ denotes the average value of the relative signal level as obtained from Eq. 18. The corresponding distributions are shown on figure 18 for 6 different average values of the relative signal level.

What is striking on figure 18 is the fact that even for signal components of non-negligible levels (such as $\bar{Q}$ = 8dB), the relative signal levels can still take very low values (below 0dB). As a consequence, the use of STSA generates strong random variations of the low-level signal components [27]. Although systematic, these variations are not always heard in practice because they are often perceptively masked either by some other signal components (especially when the noise level is low) or by the fraction of broad band noise that remains after the processing.

5.1.4 The musical noise phenomenon

What is musical noise? The other important feature of figure 18 is that when only the noise is present (when $\bar{Q} = 1$), the observed relative signal level

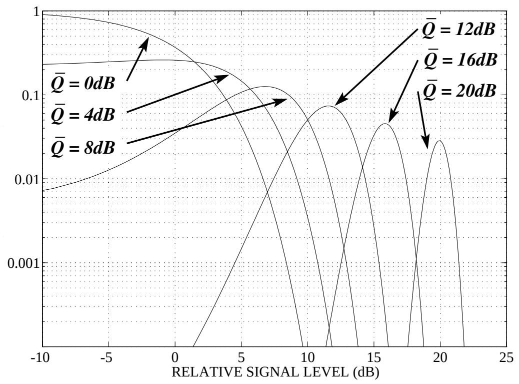

Figure 18: Probability density of the relative signal level for different mean values $\bar{Q}$ (from left to right: 0, 4, 8, 12, 16 and 20dB).

can still take high values. It is thus practically impossible to separate the noise from the low level signal components on the basis of the relative signal level. As a result, the total cancellation of the noise can only be obtained at the cost of some distortion of the low-level components.

In most STSA techniques, the noise that remains after the processing has a very unnatural disturbing quality, especially in a musical context [103]. This phenomenon is generally referred to as musical noise [43] (also as 'musical tones' [131] or 'residual noise' [47, 146]). The musical noise phenomenon is a direct consequence of the fact that the periodogram estimate used for evaluating the relative signal level yields values that are (asymptotically) uncorrelated even for neighboring bins [21]. This result, which holds for short-time transform bins belonging to the same analysis frame is complemented by the fact that bins from successive frames will also tends to be uncorrelated for frames which do not overlap in time (again, under the hypothesis of a sufficiently fast decay of the noise autocorrelation function).

Combining these two properties, it is easily seen that STSA transforms the original broad band noise into a signal composed of short-lived tones with randomly distributed frequencies. Moreover, with a 'standard' suppression rule (one that depends only on the relative signal level as measured in the current short-time frame) this phenomenon can only be eliminated by a crude overestimation of the noise level. Using the result of Eq. 20 in the case where $\bar{Q} = 1$, it is easily shown that the overestimation needed to make the probability of appearance of musical noise negligible (below 0.1%) is about 9 dB [25].

Solutions to the musical noise problem Various empirical modifications of the basic approach have been proposed to overcome this problem. A first possibility consists in taking advantage of the musical noise characteristics: more precisely, the short duration of the musical noise components (typically a few short-time frames) [15, 146] and the fact that the appearance of musical noise in one sub-band is independent of that in other sub-bands [131]. The main shortcoming of this type of approach is that, since they are based on average statistical properties, the musical noise is reduced (ie. its appearance is made less frequent) but not completely eliminated.

Another possibility is to use a smoothed estimate of the relative signal level. Time-smoothing has been considered in [15] and [47], but frequency smoothing can also be used [24, 25]. Limitations of this smoothing approach include the fact that it can generate signal distortion, particularly during transients, when time-smoothing is used. A more elaborate version of the time-smoothing approach aimed at reducing signal distortion is described in [46].

Finally, an alternative approach consists in concealing the musical noise artifact behind a sufficient level of remaining noise [103, 24]. One simple way to proceed consists in constraining the computed gain to lie above a preset threshold (which is achieved by the 'noise floor' introduced by Berouti et al. [13]).

The Ephraim and Malah suppression rule Besides these procedures specifically designed to counter the musical noise artifact, it has been noted that the suppression rules proposed by Ephraim and Malah [42, 43, 44] do not generate musical noise [43, 142, 26]. This is shown in [26] to be a consequence of the predominance of the time-smoothed signal level (the so called 'a priori signal to noise ratio') over the usual 'instantanenous' relative signal level.

A nice feature of the Ephraim and Malah suppression rule is the 'intelligent' time-smoothing of the relative signal level resulting from the use of an explicit statistical model of the sub-band noise: a strong smoothing when the level is sufficiently low to be compatible with the hypothesis that only the noise is present, and no smoothing otherwise [26]. Surprisingly, this behavior of the Ephraim and Malah suppression rule is related to the principle adopted in [46] (which consists in varying the horizon of the time-smoothing depending on the signal level). The Ephraim and Malah suppression rule therefore allows a very 'natural' means (not based on fixed thresholds) of reducing the musical noise artifact without introducing penalizing signal distortions.

When using the Ephraim and Malah suppression rule, it appears that it is still useful to limit the attenuation in order to avoid the reappearance of the musical noise phenomenon at low-levels. In practice, the average attenuation applied to the noisy part can be easily controlled via one of the parameters of the method (see [26]) in the range from 0dB to approximately -15dB (with lower values the musical noise effect can be audible in some cases). An interesting generalization consists in specifying a frequency dependent average noise reduction in order to take into account the fact that all regions of the noise spectrum do not contribute equally to the loudness sensation [102, 154].

5.1.5 Current trends and perspectives

Improving the noise characterization In many real life applications, the hypothesis that the noise is stationary is unrealistic and it is necessary to track the time-variations of the noise characteristics. For audio restoration, it seems that this aspect can play an important part in the case of old analog disk recordings. Indeed, the noise present on such recordings sounds 'less regular' than the tape hiss heard on analog tapes. It is also common to observe a discrepancy between the noise characteristics measured before and after the recorded part [25].

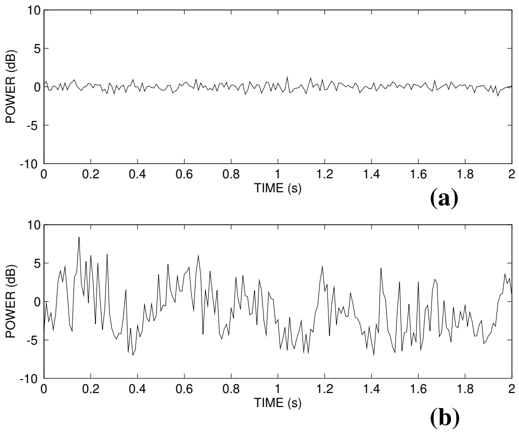

Figure 19: Short-time power variations. (a) for a standard analog cassette; (b) for a 78 rpm recording. The signal power is estimated at a 10ms rate and normalized by its average value.

An example of such a behavior can be seen on figure 19 which displays the time variations of the short-time power[8] for two noises measured on different audio recording: on a standard analog cassette for part (a), on a 78 rpm record for part (b). The sharp spikes seen on part (b) of figure 19 are simply due to the presence of impulsive degradations in the disk noise, which of course is not the case for the tape noise. However, the comparison between the two parts of figure 19 shows that the range of the power variations is much more important for the analog disk noise (part [b]) than for the tape noise (part [a]).

It is also interesting to note that the long-term power variations of the disk noise (part [b] of figure 19) seem to be related to the disk rotation period (0.77s for a 78 rpm record). This result indicates that the noise present on this particular analog disk is certainly not stationary, but that it could be cyclostationary [49]. More elaborate tests would be needed to determine if this noise is indeed

[8] More precisely, the quantity displayed is the signal power estimated from 10ms frames. As the power spectral densities of the two types of noise exhibit a strong peak at the null frequency, the two noises were pre-whitened by use of an all-pole filter [25]. This pre-processing guarantees that the noise autocorrelation functions decay sufficiently fast to obtain a robust power estimate even with short frame durations [77].

cyclostationary, and what type of cyclostationarity is actually involved (the simplest model would be an amplitude modulated stationary process) [53].

In practice, it is important to emphasize that the various procedures that have been proposed for updating the estimate of the noise characteristics in the context of speech enhancement [15, 96, 131, 46] are usually not applicable for audio signals: they rely on the presence of signal pauses that are frequent in natural speech, but not necessarily in musical recordings. The development of noise tracking procedures that are suited for an application to audio signals thus necessitates a more precise knowledge of the noise characteristics in cases where it cannot be considered stationary.

Use of perceptual noise-reduction criteria Recently, efforts have been devoted to the development of noise suppression strategies based on perceptual criteria [23, 24, 140, 105]. As of today, the proposed techniques only make use of data concerning the simultaneous masking effect in order to determine the frequency regions where the noise is most audible. A surprising side effect of these techniques is that they notably reduce the musical noise phenomenon [24]. This feature can be attributed to the strong smoothing of the frequency data in the upper frequency range performed in these techniques to simulate the ear's frequency integration properties.

Clearly more work needs to be done to take advantage of other known properties of the auditory system in the context of noise reduction. Interesting clues include the consideration of non-simultaneous masking effects that may be helpful in reducing transient distortions, as well as absolute thresholds of hearing. A troublesome point associated with the use of such perceptual criteria is that they require the knowledge of the listening acoustic intensity [102]. For most applications this requirement cannot be satisfied so that only a worst-case analysis is feasible. However, in cases where the noise reduction is performed directly at the playback level, the adaptation of the noise suppression rule to the effective acoustic intensity of the audio signal is certainly a promising aspect.

Improving the properties of the short-time transform Another interesting area of research deals with the design of the short-time transform. It is striking to see that while many efforts have been dedicated to the study of advanced suppression rules, little has been done concerning the analysis part of the noise reduction-system.

The first approach that need to be more precisely evaluated is the use of non-uniform filter banks [115, 142], especially if they are applied in connection with perceptual criteria. Indeed, non-uniform filter banks allow a frequency dependent specification of the time-resolution/bandwidth compromise which could be adapted to the known features of our hearing system. The results of section 5.1.3 show that a high frequency-resolution is needed anyway, at least in the lower part of the spectrum, to ensure a sufficient separation of sinusoidal signal components from the noise.

A complementary approach is based on the observation that the use of a fixed analysis scheme may be too constraining, which leads to the design of analysis/synthesis structures that are adapted to the local characteristics of the signal. For speech enhancement, various recent works report the successful use of subspace representations in place of the STFT [36, 45, ?, ?]. The subspace representation is still frame-based but it is characterized by a high frequency-resolution (see [36, ?] for the link with damped sinusoidal models). It has however been shown, using stationarity assumptions, that subspace approaches are asymptotically equivalent to STSA techniques for large frame durations [?]. For audio restoration, it can thus be expected that both type of methods will yield comparable results. The Adapted Waveform Analysis method described in [11] presents a different approach based on a wavelet decomposition of the signal. This promising method basically operates by determining a basis of wavelets [?] which is best adapted to the characteristics of the signal.

5.2 Discussion

A number of noise reduction methods have been described, with particular emphasis on the short-term spectral methods which have proved the most robust and effective to date. However, it is anticipated that new methodology and rapid increases in readily-available computational power will lead in the future to the use of more sophisticated methods based on realistic signal modelling assumptions and perceptual optimality criteria.

6 Pitch variation defects

A form of degradation commonly encountered in disc, magnetic tape and film sound recordings is an overall pitch variation not present in the original performance. The terms 'wow' and 'flutter' are often used in this context and are somewhat interchangeable, although wow tends to refer to variations over longer time-scales than flutter, which often means a very fast pitch variation sounding like a tremolo effect. This section addresses chiefly the longer term defects, such as those connected variations in gramophone turntable speeds, which we will refer to as wow, although similar principles could be applied to short-term defects.

There are several mechanisms by which wow can occur. One cause is a variation of rotational speed of the recording medium during either recording or playback. A further cause is eccentricity in the recording or playback process for disc and cylinder recordings, for example a hole which is not punched perfectly at the centre of a gramophone disc. Lastly it is possible for magnetic tape and optical film to become unevenly stretched during playback or storage; this too leads to pitch variation in playback. Accounts of wow are given in [7, 48].

In some cases it may be possible to make a physical correction for this defect, as the case of a gramophone disk whose hole is not punched centrally. In most cases, however, no such correction is possible, and signal processing techniques must be used. The only approach currently known to the authors is that of Godsill and Rayner [54, 60, 55], which is described here.

The physical mechanism by which wow is produced is equivalent to a non-uniform warping of the time axis. If the undistorted time-domain waveform of the gramophone signal is written as $s(t)$ and the time axis is warped by a monotonically increasing function $f_w(t)$ then the distorted signal is given by:

$$x(t) \;=\; s(f_w(t)) \tag{21}$$

If the time warping function $f_w()$ is known then it is possible to regenerate the undistorted waveform $s(t)$ as

$$s(t) = x(f_w^{-1}(t)). \tag{22}$$

A wow restoration system is thus primarily concerned with estimation of the time warping function or equivalently the pitch variation function $p_w(t) = f_w'(t)$. In the discrete signal domain we have discrete observations $x[n] = x(nT)$, where T is the sampling period. If the pitch variation function corresponding to each sampling instant, denoted by $p_w[n]$, is known then it is possible to estimate the undistorted signal using digital resampling operations.

If we have good statistical models for the undistorted audio signal and the process which generates the wow, it may then be possible to estimate the pitch variation $p_w[n]$ from the wow-degraded data $x[n]$. Any of the standard models used for audio signals (see section 2) are possible, at least in principle. However, the chosen model must be capable of capturing accurately the pitch variations of the data over the long time-scales necessary for identification of wow. One suitable option is the generalized harmonic model (see section 2, equation (2)). This represents tonal components in the signal as sinusoids, allowing for a simple interpretation of the wow as a frequency modulation which is common to all components present at a particular time.

Consider a fixed-frequency sinusoidal component $s_i(t) = sin(\omega_{0i}t + \phi_{0i})$ from a musical signal, distorted by a pitch variation function $p_w(t)$. The pitch-distorted component $x_i(t)$ can be written as (see [54]):

$$\begin{aligned} x_i(t) &= s_i(f_w(t)) \\ &= \sin\left(\omega_{0i}\int_0^t p_w(t)dt + \phi_{0i}\right), \end{aligned} \tag{23}$$

which is a frequency-modulated sine-wave with instantaneous frequency $\omega_{0i}\, p_w(t)$. The same multiplicative modulation factor $p_w(t)$ will be applied to all frequency components present at one time. Hence we might estimate $p_w[n]$ as that frequency modulation which is common to all sinusoidal components in the music. This principle is the basis of the frequency domain estimation algorithm now described.

6.1 Frequency domain estimation [54, 60, 55]

In this procedure it is assumed that musical signals are made up as additive combinations of tones (sinusoids) which represent the fundamental and harmonics

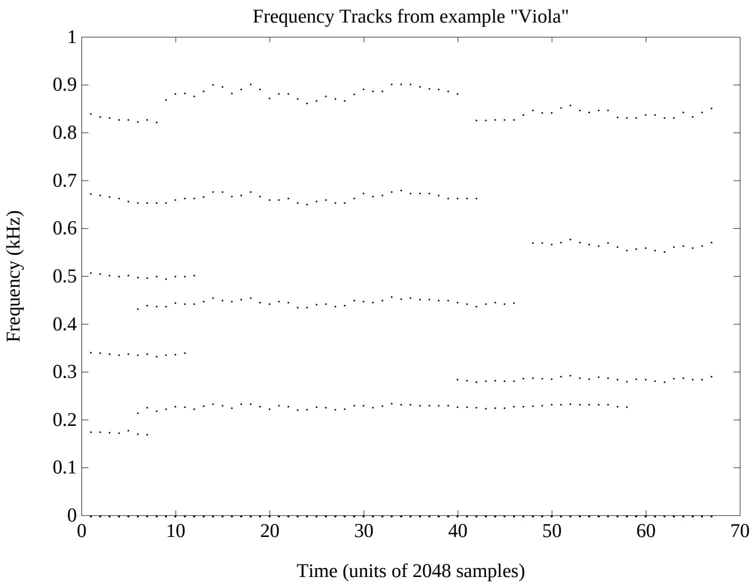

Figure 20: Frequency tracks generated for example 'Viola'.

of all the musical notes which are playing. Since this is certainly not the case for most non-musical signals, we might expect the method to fail for, say, speech extracts or acoustical noises. Fortunately, it is for musical extracts that pitch variation defects are most critical. The pitch variation process is modelled as a smoothly varying waveform with no sharp discontinuities, which is reasonable for most wow generation mechanisms.

The method proceeds in three stages. The first stage involves estimation of the tonal components using a DFT magnitude-based peak tracking algorithm closely related to that described in [97] and chapter **??**. This pre-processing stage, allowing for individual note starts and finishes, provides a set of time-frequency 'tracks' (see figures 20 and 22), from which the overall pitch variation is estimated. It is assumed in this procedure that any genuine tonal components in the corrupted signal will have roughly constant frequency for the duration of each DFT block.

The second stage of processing involves extracting smooth pitch variation information from the time-frequency tracks. For the nth block of data there will be P_n frequency estimates corresponding to the P_n tonal components which were being tracked at that time. The ith tonal component has a nominal centre frequency f_{0i}, which is assumed to remain fixed over the period of interest, and a measured frequency $f_i[n]$. Variations in $f_i[n]$ are attributed to the pitch variation value $p_w[n]$ and a noise component $v_i[n]$. This noise component is composed both of inaccuracies in the frequency tracking stage and genuine 'performance' pitch deviations (such as vibrato or tremolo) in tonal components.

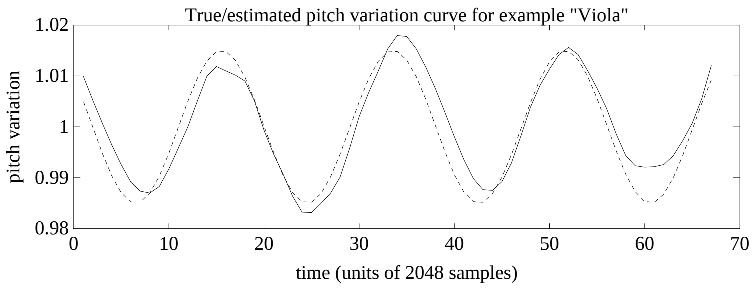

Figure 21: Estimated (full line) and true (dotted line) pitch variation curves generated for example ‘Viola’.

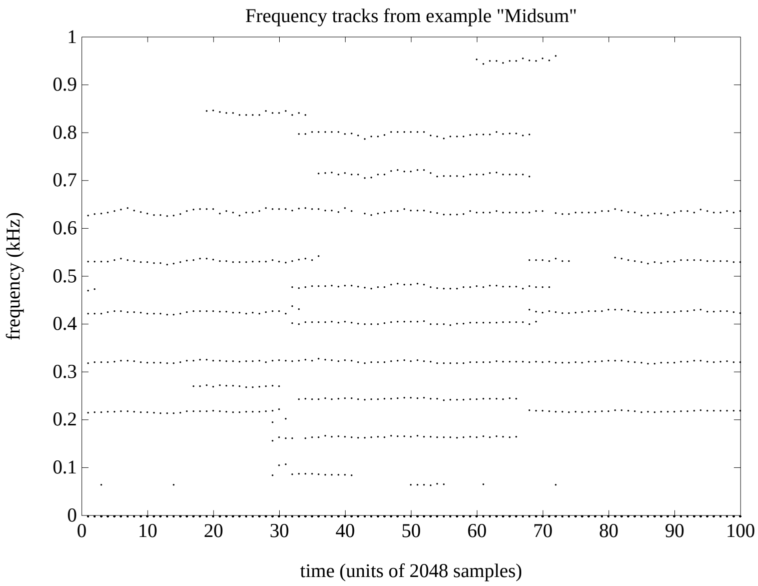

Figure 22: Frequency tracks generated for example ‘Midsum’.

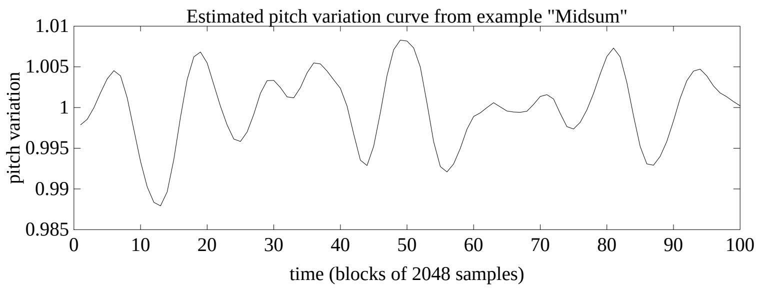

Figure 23: Pitch variation curve generated for example 'Midsum'.

Smooth pitch variations which are common to all tones present may then be attributed to the wow degradation, while other variations (non-smooth or not common to all tones) are rejected as noise, $v_i[n]$. The approach could of course fail during non-tonal ('unvoiced') passages or if note 'slides' dominate the spectrum, and future work might aim to make the whole procedure more robust to this possibility.

Each frequency track has a 'birth' and 'death' index b_i and d_i such that b_i denotes the first DFT block at which f_{0i} is present ('active') and d_i the last (each track is then continuously 'active' between these indices). Frequencies are expressed on a log-frequency scale, as this leads to linear estimates of the pitch curve (see [54] for comparison with a linear-frequency scale formulation). The model equation for the measured log-frequency tracks $f_{il}[n]$ is then:

$$f_{il}[n] = \left. \begin{array}{ll} f_{0il} + p_{wl}[n] + v_{il}[n], & b_i \leq n \leq d_i \\ 0, & \text{otherwise} \end{array} \right\}, \quad 1 \leq i \leq P_{max}, \tag{24}$$

where subscript 'l' denotes the logarithm of the frequency quantities previously defined. P_{max} is the total number of tonal components tracked in the interval of N data blocks. At block n there are P_n active tracks, and the length of the ith track is then given by $N_i = d_i - b_i + 1$.

If the noise terms $v_{il}[n]$ are assumed i.i.d. Gaussian, the likelihood function for the unknown centre frequencies and pitch variation values can be obtained. A singular system of equations results if the Maximum likelihood (ML) (or equivalently least squares) solution is attempted. The solution is regularized by incorporation of the prior information that the pitch variation is a 'smooth' process, through a Bayesian prior probability framework. A second difference-based Gaussian smoothness prior is used, which leads to a linear MAP estimator for the unknowns (see [54, 55] for full details). The estimate is dependent upon a regularizing parameter which expresses the degree of second difference smoothness expected from the pitch variation process. In [54, 55] this parameter is determined experimentally from the visual smoothness of results estimated from a small sample of data, but other more rigorous means are available for estimation of such 'hyperparameters' given the computational power (see, e.g. [124, 91]). Examples of pitch variation curves estimated from synthetic and real pitch degradation are shown in figures 21 and 23, respectively.

The estimation of pitch variation allows the final re-sampling operation to proceed. Equation (22) shows that, in principle, perfect reconstruction of the undegraded signal is possible in the continuous time case, provided the time warping function is known. In the discrete domain the degraded signal $x[n]$ is considered to be a non-uniform re-sampling of the undegraded signal $s[n]$, with sampling instants given by the time-warping function $f_w[n]$. Note, however, that the pitch varies very slowly relative to the sampling rate. Thus, at any given time instant it is possible to approximate the non-uniformly sampled input signal as a uniformly sampled signal with sample rate $1/T' = p_w[n]/T$. The problem is then simplified to one of *sample rate conversion* for which there are well-known techniques (see e.g. [33, 122]). Any re-sampling or interpolation technique which can adjust its sample rate continuously is suitable, and a truncated 'sinc' interpolation is proposed in [54, 60, 55].

6.1.1 Summary

Informal listening tests indicate that the frequency-based method is capable of a very high quality of restoration in musical extracts which have a strong tonal character. The procedure is, however, sensitive to the quality of frequency tracking and to the constant-frequency harmonic model assumed in pitch estimation. New work in the area might attempt to unify pitch variation estimation and frequency tracking into a single operation, and introduce more robust modelling of musical harmonics.

7 Reduction of Non-linear Amplitude Distortion

Many examples exist of audio recordings which are subject to non-linear amplitude distortion. Distortion can be caused by a number of different mechanisms such as deficiencies in the original recording system and degradation of the recording through excessive use or poor storage. This section formulates the reduction of non-linear amplitude distortion as a non-linear time series identification and inverse filtering problem. Models for the signal production and distortion process are proposed and techniques for estimating the model parameters are outlined. The section concludes with examples of the distortion reduction process.

An audio recording may be subject to various forms of non-linear distortion, some of which are listed below :

1. Non-linearity in amplifiers or other parts of the system gives rise to intermodulation distortion [130].

2. Cross-over distortion in Class B amplifiers [130].

3. Tape saturation due to over recording [130]: recording at too high a level on to magnetic tape leads to clipping or severe amplitude compression of a signal.

4. Tracing distortion in gramophone recordings [126]: the result of the playback stylus tracing a different path from the recording stylus. This can occur if the playback stylus has an incorrect tip radius.

5. Deformation of grooves in gramophone recordings [126]: the action of the stylus on the record groove can result in both elastic and plastic deformation of the record surface. Elastic deformation is a form of distortion affecting both new and used records; plastic deformation, or record wear, leads to a gradual degradation of the reproduced audio signal.

The approach to distortion reduction is to model the various possible forms of distortion by a non-linear system. Rather than be concerned with the actual mechanics of the distortion process, a structure of non-linear model is chosen which is thought to be flexible enough to simulate the different types of possible distortion.

7.1 Distortion Modelling

A general model for the distortion process is shown in figure 24 where the input to the nonlinear system is the undistorted audio signal $s[n]$ and the output is the observed distorted signal $x[n]$.

The general problem of distortion reduction is that of identifying the non-linear system and then applying the inverse of the non-linearity to the distorted signal $x[n]$ in order to recover the undistorted signal $s[n]$. Identification of the non-linear system takes two main forms depending on the circumstances. The first is when the physical system which caused the distortion is available for measurement. For example the recording system which produced a distorted recording may be available. Under these circumstances it is possible to apply a known input signal to the system and apply system identification techniques in order to determine the non-linear transfer function or apply adaptive techniques to recover the undistorted signal [119, 129, 81, 82]. The second, and much more common, situation is when the only information available is the distorted signal itself. The approach is now to postulate a model for both the undistorted signal

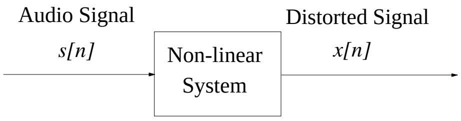

Figure 24: Model of the distortion process

and the distortion process. Time series identification techniques must then be used to determine values for the model parameters. This section will concentrate on this situation which might be called *blind identification*.

Choice of a suitable non-linear model to represent the signal and distortion processes is not a straightforward decision since there are many different classes from which to choose.

7.2 Non-linear Signal Models

A non-linear time series model transforms an observed signal $x[t]$ into a white noise process $e[t]$, and may be written in discrete form [121] as:

$$e[t] \;=\; F'\{\ldots, x[t-2], x[t-1], x[t], x[t+1], x[t+2], \ldots\}$$

where $F'\{.\}$ is some non-linear function.

Assuming that $F'\{.\}$ is an invertible function this may be expressed as:

$$x[t] \;=\; F\{\ldots, e[t-2], e[t-1], e[t], e[t+1], e[t+2], \ldots\} \qquad (25)$$

This functional relationship may be expressed in a number of different forms; two of which will be briefly considered.

7.2.1 The Volterra Series

For a time invariant system defined by equation 25, it is possible to form a Taylor series expansion of the non-linear function to give [121]:

$$\begin{aligned} x[t] \;=\; & k_0 + \sum_{i_1=-\infty}^{\infty} h_{i_1} e[t-i_1] + \sum_{i_1=-\infty}^{\infty}\sum_{i_2=-\infty}^{\infty} h_{i_1,i_2} e[t-i_1]e[t-i_2] + \\ & \sum_{i_1=-\infty}^{\infty}\sum_{i_2=-\infty}^{\infty}\sum_{i_3=-\infty}^{\infty} h_{i_1,i_2,i_3} e[t-i_1]e[t-i_2]e[t-i_3] + \cdots \end{aligned} \qquad (26)$$

where the coefficients $k_0, h_{i_1}, h_{i_1,i_2}, \ldots$ are the partial derivatives of the operator F. Note that the summation involving h_{i_1} in the discrete Volterra series corresponds to the normal convolution relationship for a linear system with impulse response $h_{i_1}(n)$. The Volterra Series is a very general class of non-linear model which is capable of modelling a broad spectrum of physical systems. The generality of the model, while making it very versatile, is also its main disadvantage : for successful modelling of an actual system, a very large order of Volterra expansion is often needed, a task which is generally not practical. In view of this, it becomes necessary to consider other representations of non-linear time series.

7.2.2 NARMA Modelling

The NARMA (Non-Linear AutoRegressive Moving Average) model was introduced by Leontaritis and Billings [83] and defined by:

$$x[n] \; = \; f\{x[n-1], \ldots x[n-P_x], e[n-1], \ldots e[n-P_e]\} \; + \; e[n]$$

Combining the terms $x[n-1], \ldots x[n-P_x]$ and $e[n-1], \ldots e[n-P_e]$ into a single vector $\mathbf{w}(n)$ and expanding as a Taylor series gives the following representation of a non-linear system [29].

$$\begin{aligned} x[n] \; = \; & a_0 \; + \sum_{i_1=1}^{P_x+P_e} a_{i_1} w_{i_1}(n) \; + \sum_{i_1=1}^{P_x+P_e} \sum_{i_2=i_1}^{P_x+P_e} a_{i_1 i_2} w_{i_1}(n) w_{i_2}(n) \; + \; \ldots \\ & \sum_{i_1=1}^{P_x+P_e} \cdots \sum_{i_l=i_{l-1}}^{P_x+P_e} a_{i_1 \ldots i_l} w_{i_1}(n) \ldots w_{i_l}(n) \; + \; s(n) \end{aligned} \tag{27}$$

where:

$$\mathbf{w}(n) \quad = \quad \begin{bmatrix} x[n-1] \\ \vdots \\ x[n-p_x] \\ e[n-1] \\ \vdots \\ e[n-p_s) \end{bmatrix}$$

The advantage of such an expansion is that the model is linear in the unknown parameters a so that many of the linear model identification techniques can also be applied to the above non-linear model. Iterative methods of obtaining the parameter estimates for a given model structure have been developed [14].
A number of other non-linear signal models are discussed by Priestley [121] and Tong [138].

7.3 Application of Non-linear models to Distortion Reduction

The general Volterra and NARMA models suffer from two problems from the point of view of distortion correction. They are unnecessarily complex and even after identifying the parameters of the model it is still necessary to recover the undistorted signal by some means. In section 2 it was noted that audio signals are well-represented by the autoregressive (AR) model defined by equation 1:

$$s[n] = \sum_{i=1}^{P} s[n-i]\, a_i + e[n].$$

Thus a distorted signal may be represented as a linear AR model followed by a non-linear system as shown in figure 25.

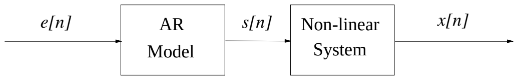

Figure 25: Model of the signal and distortion process

Two particular models will be considered for the non-linear system.

7.3.1 Memoryless Non-linearity

A special case of the Volterra system given by equation 26 is:

$$x[n] = h_0\, s[n] + h_{00}\, s^2[n] + \cdots h_{0\ldots0} s^q[n] + \cdots$$

This is termed a memoryless non-linearity since the output is a function of only the present value of the input $s[n]$. The expression may be regarded as a power series expansion of the non-linear input-output relationship of the non-linearity. In fact this representation is awkward from an analytical point of view and it is more convenient to work in terms of the inverse function. Conditions for invertibility are discussed in Mercer [99].

$$\begin{aligned} s[n] &= k_1\, x[n] + k_2\, x^2[n] + \cdots k_q x^q[n] + \cdots \\ &= \sum_{q=0}^{\infty} k_q\, x^q[n] \end{aligned}$$

An infinite order model is clearly impractical to implement. Hence it is necessary to truncate the series :

$$s[n] = \sum_{q=0}^{Q} k_q\, x^q[n] \tag{28}$$

A reasonable assumption is that there is negligible distortion for low-level signals, ie $\{x[n] = s[n];\ \text{for } s[n] \approx 0\}$ so that $k_0 = 1$. (Note that this assumption would not be valid for crossover distortion). This model will be referred to in general as the Autoregressive-Memoryless Non-linearity (AR-MNL) model and as the AR(P)-MNL(Q) to denote a AR model of order P and a memoryless non-linearity of order Q.

Note that if the non-linear parameters k_i can be identified then the undistorted signal $\{s[n]\}$ can be recovered from the distorted signal $\{x[n]\}$ by means of equation 28.

7.3.2 Non-linearity with Memory

The AR-MNL model is clearly somewhat restrictive in that most distortion mechanisms will involve memory. For example an amplifier with a non-linear output stage will probably have feedback so that the memoryless non-linearity will be included within a feedback loop and the overall system could not be modelled as a memoryless non-linearity. The general NARMA model incorporates memory but its use imposes a number of analytical problems. A special case of the NARMA model is the NAR (Non-linear AutoRegressive) model in which the current output $x[n]$ is a non-linear function of only past values of output and the present input $s[n]$. Under these conditions equation 27 becomes:

$$\begin{aligned} x[n] \quad = \quad & + \sum_{i_1=1}^{P_x} \sum_{i_2=i_1}^{P_x} a_{i_1 i_2} x[n-i_1] x[n-i_2] \ + \ \ldots \\ & \sum_{i_1=1}^{P_x} \cdots \sum_{i_l=i_{l-1}}^{P_x} a_{i_1 \ldots i_l} x[n-i_1] \ldots x[n-i_l] \ + \ s[n] \end{aligned} \tag{29}$$

The linear terms in $x[n-i_1]$ have not been included since they are represented by the linear terms in the AR model. This model will be referred to as Autoregressive Non-linear Autoregressive (AR-NAR) model in general and as AR(P)-NAR(Q) model in which the AR section has order P and only Q of the non-linear terms from equation 29 are included. Note that the undistorted signal $\{s[n]\}$ can be recovered from the distorted signal $\{x[n]\}$ by use of equation 29 provided that the parameter values can be identified.

7.4 Parameter Estimation

In order to recover the undistorted signal it is necessary to estimate the parameter values in equations 28 and 29. A general description of parameter estimation is given in many texts, e.g. Norton [110, 77].

One powerful technique is Maximum Likelihood Estimation (MLE) which requires the derivation of the Joint Conditional Probability Density Function (PDF) of the output sequence $\{x[n]\}$, conditional on the model parameters. The input $\{e[n]\}$ to the system shown in figure 25 is assumed to be a white Gaussian noise (WGN) process with zero mean and a variance of σ^2. The probability density of the noise input is:

$$p(e[n]) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{e^2[n]}{2\sigma^2}\right\}$$

Since $\{e[n]\}$ is a WGN process, samples of the process are independent and the joint probability for a sequence of data ($\{e[n]\}, n = P+1 \text{ to } N$) is given by :

$$p(e[P+1], \ldots e[N]) \;=\; \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{N-P} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=P+1}^{N} e^2[n]\right\} \tag{30}$$

The terms $\{e[1], e[2]...e[P]\}$ are not included because they cannot be calculated in terms of the observed output $\{x[n]\}$ so that, strictly speaking, the above is a conditional probability but there is little error if the number of observations $N \gg P$.

An expression for the Joint Probability Density Function for the observations $\{x[n]\}$ may be determined by transformations from $\{e[n]\}$ to $\{s[n]\}$ and from $\{s[n]\}$ to $\{x[n]\}$. This gives the likelihood function for the AR-MNL system as :

$$p(x[P+1], x[P+2], \ldots x[N]|\mathbf{a}, \mathbf{k}, \sigma) \;=$$

$$\left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{N-P} \left|\prod_{n=P+1}^{N}\{1 + \sum_{q=2}^{Q} q\, k_q\, x^{q-1}[n]\}\right| \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=P+1}^{N} e^2[n]\right\} \tag{31}$$

where $\mathbf{a}$ is a vector containing the parameters $a_1 \ldots a_P$ of the AR model and $\mathbf{k}$ is a vector containing the parameters $k_0 \ldots k_Q$. The noise sequence $\{e[n]\}$ may be expressed in terms of the observed distorted signal $\{x[n]\}$ using equations 1 and 28

The Likelihood function for the AR-NAR system is:

$$p[x[P+P_x+1], \ldots x[N]|\mathbf{a}, \mathbf{k}, \sigma] \;=\; \left[\frac{1}{\sqrt{2\pi}\sigma}\right]^{N-P-P_x} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=P+P_x+1}^{N} e^2[n]\right\}$$

where $\mathbf{a}$ is the vector of AR parameters and $\mathbf{k}$ is a vector containing the parameters a_{i_1}, $a_{i_1 i_2}$, ... of the NAR model. The noise sequence $\{e[n]\}$ may be expressed in terms of the observed distorted signal $\{x[n]\}$ using equations 1 and 29

The MLE approach involves maximising the Likelihood function with respect to $\mathbf{a}$, $\mathbf{k}$ and σ. The values of $\mathbf{a}$, $\mathbf{k}$ and σ which maximise this equation are the Maximum Likelihood estimates of the model.

7.4.1 Computational aspects

In general there is no analytic solution to maximising the Likelihood equations so that it is necessary to perform a multidimensional optimisation over the unknown model parameters. However before performing the optimisation it is necessary to select a model of appropriate order; too low an order results in a poor system which is unable to correct distortion, too high an order results in an unnecessarily complicated model which imposes a heavy computational burden in determining the optimal parameter values. Model order selection for the memoryless non-linearity is simply a matter of choosing the order of the polynomial expansion in equation 28. However the problem is more complex with the NAR model, equation 29, since the number of permutations of terms can be extremely large. There is no intuitive means for estimating which non-linear terms should be included and it is necessary to perform the Maximum Likelihood optimisation for each combination of terms in order to find an acceptable system. Such a global search over even a relatively limited subset of the possible model terms is prohibitively expensive and iterative methods have been developed to search the space of model functions to determine an acceptable, although not necessarily optimal, system [99].

In order to compare the performance of models containing different non-linear terms it is necessary to use a criterion which achieves a compromise between the overly simple model and the overly complex model. One such criterion is the Akaike Information Criterion, AIC(ϕ) (see e.g. Akaike [3]) given by:

$$\begin{aligned} AIC(\phi) &= -2\log_e \{\text{Maximised Likelihood Function}\} + \\ &\qquad \phi \times [\text{Number of Parameters}] \end{aligned} \tag{32}$$

The Akaike Information Criterion is used to select the model which mimimises the AIC(ϕ) function for a specified value of ϕ. In the original formulation of the above equation, Akaike used a value of $\phi = 2$ but an alternative selection criterion proposed by Leontaritis and Billings [84] is based on a value of $\phi = 4$.

7.5 Examples

Mercer [99] presents results for the two models discussed. For the memoryless non-linearity a section of music from a recording of a brass band was passed through the non-linearity defined by:

$$\mathbf{k} = [\ 0.00 \quad 0.30 \quad 0.00 \quad 0.50\ ].$$

An AR model of order 25 was assumed and a non-linearity with $Q \leq 9$ was allowed. Figure 26 shows a section of the original, distorted and restored signals.

In order to test the AR-NAR model a section of music was passed through the non-linear system:

$$\begin{aligned} x[n] &= 0.07\,x[n-1]x[n-4]x[n-6] + 0.05\,x[n-2]x[n-2]x[n-3] \\ &+ 0.06\,x[n-3]x[n-6]x[n-8] + 0.06\,x[n-4]x[n-7]x[n-7] \\ &+ 0.05\,x[n-8]x[n-9]x[n-9] + s[n] \end{aligned}$$

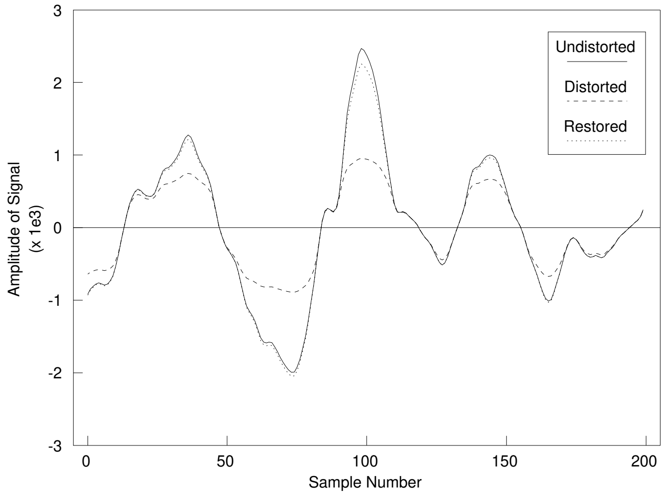

Figure 26: Typical section of AR-MNL Restoration

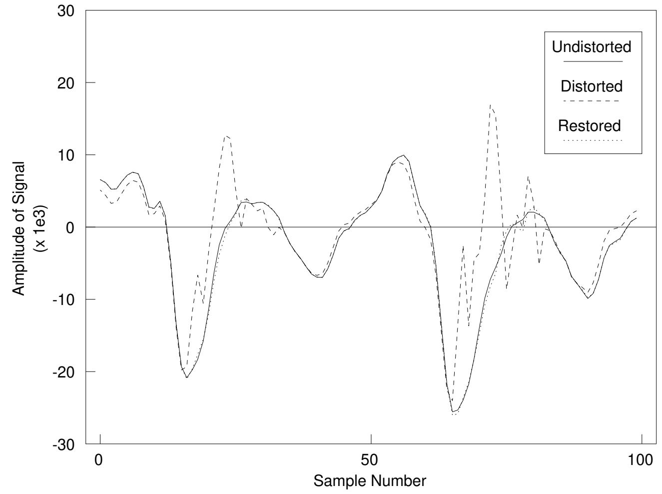

Figure 27: Typical section of AR-NAR Restoration

An AR(30)-NAR(Q) model was fitted to data blocks containing 5000 samples of the distorted data. The non-linear terms allowed in the model were of the form:

$$w(n-i)w(n-j)w(n-k)$$
$$\text{for} \quad i = 1:9,\ j = i:9,\ k = j:9.$$

and a model complexity of $Q \leq 20$ was allowed. Typical results are shown in figure 27 which shows a section of the original, distorted and restored signals.

7.6 Discussion

The techniques introduced in this section perform well on audio data which have been distorted by the appropriate model. However extensive testing is required to determine whether or not the non-linear models proposed are sufficiently flexible to model real distortion mechanisms.

Further work is required on methods for searching the space of non-linear models of a particular class (eg. AR-NAR) to determine the required model complexity. This may perhaps be best achieved by extending the Maximum Likelihood approach to a full Bayesian posterior probability formulation and using the concept of *model evidence* [118] to compare models of different complexity. Some recent work in this field [139] applies Bayesian Markov chain Monte Carlo (MCMC) methods to the problem of non-linear model term selection. It is planned to extend this work in the near future to model selection for the AR-NAR distortion models discussed earlier in this section.

8 Other areas

In addition to the specific areas of restoration considered in previous sections there are many other possibilities which we do not have space here to address in detail. These include processing of stereo signals, processing of multiple copies of mono recordings, frequency range restoration and pitch adjustment.

Where stereo signals are processed, it is clearly possible to treat each channel as a separate mono source, to which many of the above processes could be applied (although correction of pitch variations would need careful synchronization!). However, this is sub-optimal, owing to the significant degree of redundancy and the largely uncorrelated nature of the noise sources between channels. It is likely that a significantly improved performance could be achieved if these factors were utilized by a restoration system. This might be done by modelling cross-channel transfer functions, a difficult process, owing to complex source modelling effects involving room acoustics. Initial investigations have shown some promise, and this may prove to be a useful topic of further research.

A related problem is that of processing multiple copies of the same recording. Once again, the uncorrelated nature of the noise in each copy may lead to an improved restoration, and the signal components will be closely related. In the simplest case, a stereo recording is made from a mono source. Much of the noise in the two channels may well be uncorrelated, in particular small impulsive-type disturbances which affect only one channel of the playback system. Multi-channel processing techniques can then be applied to extraction of the signal from the noisy sources. A Bayesian approach to this problem, which involves simple FIR modelling of cross-channel transfer functions, is described in [54], while a joint AR-modelling approach is presented in [71]. In the case where sources come from different records, alignment becomes a major consideration. Vaseghi and Rayner [147, 145, 148] use an adaptive filtering system for this purpose in a dual-channel de-noising application.

In many cases the frequency response of the recording equipment is highly inadequate. Acoustic recording horns, for example, exhibit unpleasant resonances at mid-range frequencies, while most early recordings have very poor high frequency response. In the case of recording resonances, these may be identified and corrected using a cascaded system model of source and recording apparatus. Such an approach was investigated by Spenser and Rayner [134, 133]. In the case where high frequency response is lacking, a model which can predict high frequency components from low is required, since any low-level high frequency information in the noisy recorded signal is likely to be buried deep in noise. Such a process becomes highly subjective, since different instruments will have different high frequency characteristics. The procedure may thus be regarded more as signal enhancement than restoration.

Pitch adjustment will be required when a source has been played back at a different (constant) speed from that at which it was recorded. This is distinct from wow (see section 6) in which pitch varies continuously with time. Correction of this defect can often be made at the analogue playback stage, but digital correction is possible through use of sample-rate conversion technology (see section 6). Time-scale modification (see chapter **??**) is not required, since changes of playback speed lead to a corresponding time compression/expansion. We note that correction of this defect will often be a subjective matter, since the original pitch of the recording may not be known exactly (especially in the case of early recordings).

9 Conclusion and Future Trends

This chapter has attempted to give a broad coverage of the main areas of work in audio restoration. Where a number of different techniques exist, as in the case of click removal or noise reduction, a brief descriptive coverage of all methods is given, with more detailed attention given to a small number of methods which the authors feel to be of historical importance or of potential use in future research. In reviewing existing work we point out areas where further developments and research might give new insight and improved performance.

It should be clear from the text that fast and effective methods are now available for restoration of the major classes of defect (in particular click removal and noise reduction). These will generally run in real-time on readily available DSP hardware, which has allowed for strong commercial exploitation by companies such as CEDAR Audio Ltd. in England and the American-based Sonic Solutions in California. It seems to the authors that the way ahead in audio restoration will be at the high quality end of the market, and in developing new methods which address some of the more complex problems in audio, such as correction of non-linear effects (see section 7). In audio processing, particularly for classical music signals, fidelity of results to the original *perceived* sound is of utmost importance. This is much more the case than, say, in speech enhancement applications, where criteria are based on factors such as intelligibility. In order to achieve significant improvements in high quality sound restoration sophisticated algorithms will be required, based on more realistic modelling frameworks. The new models must take into account the physical properties of the noise degradation process as well as the psychoacousical properties of the human auditory system. Such frameworks will typically not give analytic results for restoration, as can be seen even for the statistical click removal work outlined in section 3.4, and solutions might, for example, be based on iterative methods such as Expectation-maximize (EM) or Markov chain Monte Carlo (MCMC) which are both powerful and computationally intensive. This is, however, likely to be in accord with continual increases in speed and capacity of computational devices.

To conclude, the range of problems encountered in audio signals from all sources, whether from recorded media or communications and broadcast channels, present challenging statistical estimation problems. Many of these have now been solved successfully, but there is still significant room for improvement in achieving the highest possible levels of quality. It is hoped that the powerful techniques which are now practically available to the signal processing community will lead to new and more effective audio processing in the future.

References

[1] R. A. Haddad A. N. Akansu. *Multiresolution Signal Decomposition.* Academic Press, inc., 1992.

[2] B. Abraham and G. E. P. Box. Bayesian analysis of some outlier problems in time series. *Biometrika*, 66(2):229–36, 1979.

[3] H. Akaike. A new look at statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, 1974.

[4] J. B. Allen and L. R. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, November 1977.

[5] B. D. O. Anderson and J. B. Moore. *Optimal Filtering.* Prentice-Hall, 1979.

[6] K. Arakawa, D.H. Fender, H. Harashima, H. Miyakawa, and Y. Saitoh. Separation of a non-stationary component from the EEG by a nonlinear digital filter. *IEEE Trans. Biomedical Engineering*, 33(7):724–726, 1986.

[7] P. E. Axon and H. Davies. A study of frequency fluctuations in sound recording and reproducing systems. *Proc. IEE, Part III*, page 65, Jan. 1949.

[8] V. Barnett and T. Lewis. *Outliers in statistical data, 2nd Ed.* Chichester:Wiley, 1984.

[9] R. J. Beckman and R. D. Cook. Outlier. . . . . s. *Technometrics*, 25(2):119–165, May 1983.

[10] Arthur H. Benade. *Fundamentals of musical acoustics.* Oxford University Press, 1976.

[11] J. Berger, R. R. Coifman, and M. J. Goldberg. Removing noise from music using local trigonometric bases and wavelet packets. *J. Audio Eng. Soc.*, 42(10):808–818, October 1994.

[12] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* John Wiley & Sons, 1994.

[13] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by additive noise. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 208–211, 1979.

[14] S. A. Billings and W. S. F. Voon. A prediction-error and stepwise-regression estimation algorithm for non-linear systems. *Int. J. Control*, 44(3):803–822, 1986.

[15] S. F. Boll. Suppression of acoustic noise in speech using spectral substraction. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2):113–120, 1979.

[16] S. F. Boll. Speech enhancement in the 1980s: noise suppression with pattern matching. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 10, pages 309–325. Marcel Dekker, 1991.

[17] S.F. Boll. Speech enhancement in the 1980's: Noise suppression with pattern matching. In S. Furui and M.M. Sondhi, editors, *Advances in speech signal processing*, pages 309–325. Marcel Dekker, Inc., 1992.

[18] G. E. P. Box and G. M. Jenkins. *Time Series Analysis, Forecasting and Control.* Holden-Day, 1970.

[19] G. E. P. Box and G. C. Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55:119–129, 1968.

[20] G.E.P. Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, 1973.

[21] D. R. Brillinger. *Time Series Data Analysis and Theory.* Holden-Day, expanded edition, 1981.

[22] H. T. Bunnell. On enhancement of spectral contrast in speech for hearing-impaired listeners. *J. Acoust. Soc. Am.*, 88(6):2546–2556, December 1990.

[23] C. N. Canagarajah. A single-input hearing aid based on auditory perceptual features to improve speech intelligibility in noise. *Proc. IEEE Workshop on Audio and Acoustics*, 1991.

[24] C. N. Canagarajah. *Digital Signal Processing Techniques for Speech Enhancement in Hearing Aids.* PhD thesis, University of Cambridge, 1993.

[25] O. Cappé. *Noise reduction techniques for the restoration of musical recordings* (text in French). PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1991.

[26] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Trans. Speech and Audio Processing*, 2(2), April 1994.

[27] O. Cappé and J. Laroche. Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings. *IEEE Trans. on Speech and Audio Processing*, 3(1):84–93, January 1995.

[28] M. J. Carrey and I. Buckner. A system for reducing impulsive noise on gramophone reproduction equipment. *The Radio Electronic Engineer*, 50(7):331–336, July 1976.

[29] S. Chen and S. A. Billings. Modelling and analysis of non-linear time series. *Int. J. Control*, 50(5):2151–2171, 1989.

[30] Y. M. Cheng and D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Trans. Signal Processing*, 39(9):1943–1954, 1991.

[31] C. K. Chui. *Wavelet Analysis and its Applications, Volume 1: An Introduction to Wavelets.* Academic Press, inc., 1992.

[32] C. K. Chui. *Wavelet Analysis and its Applications, Volume 2: Wavelets: A Tutorial in Theory and Applications.* Academic Press, inc., 1992.

[33] R. E. Crochiere and L. R. Rabiner. *Multirate Digital Signal Processing.* Prentice-Hall, 1983.

[34] R.E. Crochiere. A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(2):99–102, Fev 1980.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[36] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: A regenerative approach. *Speech Communication*, 10:45–57, 1991.

[37] Diana Deutsch, editor. *The Psychology of Music.* AP series in cognition and perception. Academic Press, 1982.

[38] R. M. Dolby. An audio noise reduction system. *J. Audio Eng. Soc.*, 15(4):383–388, October 1967.

[39] J. Durbin. Efficient estimation of parameters in moving-average models. *Biometrika*, 46:306–316, 1959.

[40] A.J. Efron and H. Jeen. Pre-whitening for detection in correlated plus impulsive noise. *Proc. ICASSP*, II:469–472, 1992.

[41] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proc. IEEE*, 80(10):1526–1555, October 1992.

[42] Y. Ephraim and D. Malah. Speech enhancement using optimal non-linear spectral amplitude estimation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1118–1121, Boston, 1983.

[43] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 32(6):1109–1121, 1984.

[44] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(2):443–445, 1985.

[45] Y. Ephraim and H. VanTrees. A signal subspace approach for speech enhancement. *Proc. ICASSP*, II:359–362, 1993.

[46] A.F. Erwood and C.S. Xydeas. A multiframe spectral weighting system for the enhancement of speech signals corrupted by acoustic noise. In *SIGNAL PROCESSING V: Theories and Applications*, pages 1107–1110. Elsevier, 1990.

[47] W. Etter and G. S. Moschytz. Noise reduction by noise-adaptive spectral magnitude expansion. *J. Audio Eng. Soc.*, 42(5):341–349, 1994.

[48] U. R. Furst. Periodic variations of pitch in sound reproduction by phonographs. *Proc. IRE*, page 887, November 1946.

[49] W. Gardner, editor. *Cyclostationarity in communications and signal processing.* IEEE Press, New York, 1994.

[50] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.*, 85:398–409, 1990.

[51] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[52] J. J Gerbrands. On the relationships between the SVD, KLT and PCA. *Pattern Recognition*, 14(1):375–381, 1981.

[53] N.L. Gerr and J.C. Allen. The generalized spectrum and spectral coherence of a harmonizable time series. *Digital Signal Processing*, 4(4):222–238, October 1994.

[54] S. J. Godsill. *The Restoration of Degraded Audio Signals*. PhD thesis, Cambridge University Engineering Department, Cambridge, England, December 1993. Available in A5 book format (compressed Postscript[9]) and A4 format (compressed Postscript [10]).

[55] S. J. Godsill. Recursive restoration of pitch variation defects in musical recordings. In *Proc. ICASSP*, volume 2, pages 233–236, Adelaide, April 1994.

[56] S. J. Godsill. Bayesian enhancement of speech and audio signals which can be modelled as ARMA processes. *International Statistical Review*, 65(1):1–21, 1997.

[57] S. J. Godsill. Robust modelling of noisy ARMA signals. In *Proc. ICASSP*, April 1997.

[58] S. J. Godsill and P. J. W. Rayner. A Bayesian approach to the detection and correction of bursts of errors in audio signals. In *Proc. ICASSP*, volume 2, pages 261–264, San Francisco, March 1992.

[59] S. J. Godsill and P. J. W. Rayner. Frequency-domain interpolation of sampled signals. In *Proc. ICASSP*, volume I, pages 209–212, Mineapolis, April 1993.

[60] S. J. Godsill and P. J. W. Rayner. The restoration of pitch variation defects in gramophone recordings. In *Proc. IEEE Workshop on Audio and Acoustics*, Mohonk, NY State, October 1993.

[61] S. J. Godsill and P. J. W. Rayner. A Bayesian approach to the restoration of degraded audio signals. *IEEE Trans. on Speech and Audio Processing*, 3(4):267–278, July 1995.

[62] S. J. Godsill and P. J. W. Rayner. Robust noise modelling with application to audio restoration. In *Proc. IEEE Workshop on Audio and Acoustics*, Mohonk, NY State, October 1995.

[9] http://www-com-serv.eng.cam.ac.uk/~sjg/thesis/a5thesis.zip
[10] http://www-com-serv.eng.cam.ac.uk/~sjg/thesis/thesis.zip

[63] S. J. Godsill and P. J. W. Rayner. Robust reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler. Technical Report CUED/F-INFENG/TR.233, Cambridge University Engineering Department, Cambridge, England, November 1995. Available as Postscript[11].

[64] S. J. Godsill and P. J. W. Rayner. Robust noise reduction for speech and audio signals. In *Proc. ICASSP*, May 1996.

[65] S. J. Godsill and P. J. W. Rayner. Robust treatment of impulsive noise in speech and audio signals. In *Bayesian Robustness*, volume 29, pages 315–326. IMS Lecture Notes - Monograph Series (1996) Volume 29, 1996. *(To Appear)*.

[66] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, 1989.

[67] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, 32(2):236–242, April 1984.

[68] Donald E. Hall. *Musical acoustics: An introduction*. Wadsworth, 1980.

[69] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[70] S. Haykin. *Adaptive Filter Theory, Second Edition*. Prentice-Hall, 1991.

[71] C. M. Hicks and S. J. Godsill. A 2-channel approach to the removal of impulsive noise from archived recordings. In *Proc. ICASSP*, volume 2, pages 213–216, Adelaide, April 1994.

[72] P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, (35):73–101, 1964.

[73] P.J. Huber. *Robust Statistics*. Wiley and Sons, 1981.

[74] A. J. E. M. Janssen, R. Veldhuis, and L. B. Vries. Adaptive interpolation of discrete-time signals that can be modeled as AR processes. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-34(2):317–330, April 1986.

[75] J. D. Johnston and K. Brandenburg. Wideband coding–Perceptual considerations for speech and music. In S. Furui and M. M. Sondhi, editors, *Advances in speech signal processing*. Marcel Dekker, New York, 1992.

[76] T. Kasparis and J. Lane. Adaptive scratch noise filtering. *IEEE Trans. Consumer Electronics*, 39(4), 1993.

[77] Steven M. Kay. *Fundamentals of statistical signal processing: Estimation theory*. PH signal processing series. Prentice-Hall, 1993.

[11] http://www-com-serv.eng.cam.ac.uk/~sjg/papers/95/impulse.ps

[78] G. R. Kinzie, Jr. and D. W. Gravereaux. Automatic detection of impulse noise. *J. Audio Eng. Soc.*, 21(3):331–336, April 1973.

[79] B. Koo, J. D. Gibson, and S. D. Gray. Filtering of coloured noise for speech enhancement and coding. *Proc. ICASSP*, pages 349–352, 1989.

[80] R. Lagadec and D. Pelloni. Signal enhancement via digital signal processing. In *Preprints of the AES 74th Convention*, New York, 1983.

[81] H. Landau. On the recovery of band-limited signals after instantaneous companding and subsequent bandlimiting. *Bell Sys. Tech. J.*, 39:351–364, March 1960.

[82] H. Landau and W. L. Miranker. The recovery of distorted band-limited signals. *J. Math. Anal.Appl.*, 2:97–104, 1961.

[83] I. J. Leontaritis and S. A. Billings. Input-output parametric models for non-linear systems. part i;deterministic non-linear systems; part ii: Stochastic non-linear systems. *Int. J. Control*, 41(2):303–344, 1985.

[84] I. J. Leontaritis and S. A. Billings. Model selection and validation methods for non-linear systems. *Int. J. Control*, 45(1):311–341, 1987.

[85] J. S. Lim, editor. *Speech enhancement.* Prentice-Hall signal processing series. Prentice-Hall, 1983.

[86] J. S. Lim. Speech enhancement (preconference lecture). In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3135–3142, 1986.

[87] J. S. Lim and A. V. Oppenheim. All-pole modelling of degraded speech. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-26(3), June 1978.

[88] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12), 1979.

[89] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, December 1979.

[90] G. B. Lockhart and D. J. Goodman. Reconstruction of missing speech packets by waveform substitution. *Signal Processing 3: Theories and Applications*, pages 357–360, 1986.

[91] D. MacKay. *Bayesian Methods for Adaptive Models.* PhD thesis, Calif. Inst. Tech., 1992.

[92] R. C. Maher. A method for extrapolation of missing digital audio data. *95th AES Convention, New York*, 1993.

[93] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561–580, 1975.

[94] S. Mallatt and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. Info. Theory*, pages 617–643, 1992.

[95] R.D. Martin. Robust methods for time series. *In Applied Time Series Analysis II, D.F. Findley, ed.*, pages 683–759, 1981.

[96] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(2):137–145, April 1980.

[97] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, August 1986.

[98] R. E. McCulloch and R. S. Tsay. Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis*, 15(2):235–250, 1994.

[99] K. J. Mercer. *Identification of signal distortion models.* PhD thesis, University of Cambridge, 1993.

[100] N. J. Miller. Recovery of singing voice from noise by synthesis. *Thesis Tech. Rep. , ID UTEC-CSC-74-013, Univ. Utah*, May 1973.

[101] S. Montresor, J. C. Valière, and M. Baudry. Détection et suppression de bruits impulsionnels appliqué à la restauration d'enregistrements anciens. *Colloq. de Physique, C2, supplément au no. 2, Tome 51, Février*, pages 757–760, 1990.

[102] B. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, second edition, 1989.

[103] J. A. Moorer and M. Berger. Linear-phase bandsplitting: Theory and applications. *J. Audio Eng. Soc.*, 34(3):143–152, 1986.

[104] F. Mosteller and J.W. Tukey. *Data Analysis and Regression.* Addison-Wesley, Reading, Mass., 1977.

[105] J. Mourjopoulos, G. Kokkinakis, and M. Paraskeyas. Noisy audio signal enhancement using subjective spectra. In *Preprints of the AES 92nd Convention*, Vienna, 1992.

[106] S. H. Nawab and T. F. Quatieri. Short-time Fourier transform. In J. S. Lim and A. V. Oppenheim, editors, *Adanced topics in signal processing*, chapter 6. Prentice-Hall, 1988.

[107] M. Niedźwiecki. Recursive algorithm for elimination of measurement noise and impulsive disturbances from ARMA signals. *Signal Processing VII: Theories and Applications*, pages 1289–1292, 1994.

[108] M. Niedźwiecki and K. Cisowski. Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals. *IEEE Trans. on Signal Processing*, pages 528–537, 1996.

[109] A. Nieminen, M. Miettinen, P. Heinonen, and Y. Nuevo. Music restoration using median type filters with adaptive filter substructures. *In Digital Signal Processing-87, eds. V. Cappellini and A. Constantinides, North-Holland*, 1987.

[110] J. P. Norton. *An Introduction to Identification*. Academic Press, London, 1988.

[111] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proc. IEEE*, 69(5):529–541, 1981.

[112] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. Interpolation of missing samples for audio restoration. *Electronic Letters*, 30(8), April 1994.

[113] K. K. Paliwal and A. Basu. A speech enhancement method based on Kalman filtering. *Proc. ICASSP*, pages 177–180, 1987.

[114] A. Papoulis. *Probability, random variables, and stochastic processes*. McGraw-Hill, New York, 3rd edition, 1991.

[115] T. L. Petersen and S. F. Boll. Acoustic noise suppression in the context of a perceptual model. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1086-1088, 1981.

[116] I. Pitas and A. N. Venetsanopoulos. *Nonlinear Digital Filters*. Kluwer Academic Publishers, 1990.

[117] H. J. Platte and V. Rowedda. A burst error concealment method for digital audio tape application. *AES preprint*, 2201:1–16, 1985.

[118] K. J. Pope and P. J. W. Rayner. Non-linear system identification using bayesian inference. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 457–460, 1994.

[119] D. Preis and H. Polchlopek. Restoration of nonlinearly distorted magnetic rcordings. *J. Audio Eng. Soc.*, 32(12):79–86, January 1984.

[120] M. B. Priestley. *Spectral Analysis and Time Series*. Academic Press, 1981.

[121] M. B. Priestley. *Non-linear and Non-stationary Time Series Analysis*. Academic Press, London and San Diego, 1988.

[122] L. R. Rabiner. Digital techniques for changing the sampling rate of a signal. *Digital Audio; collected papers from the AES premier conference*, pages 79–89, June 1982.

[123] L. R. Rabiner and R. W. Schafer. *Digital processing of speech signals*. Pretice-Hall, 1978.

[124] J. Rajan. *Time Series Classification*. PhD thesis, University of Cambridge, 1994.

[125] P. J. W. Rayner and S. J. Godsill. The detection and correction of artefacts in archived gramophone recordings. In *Proc. IEEE Workshop on Audio and Acoustics*, Mohonk, NY State, October 1991.

[126] H. E. Roys, editor. *Disc Recording and Reproduction*. Dowden, Hutchinson and Ross, Inc., Stroudsberg, Pennsylvania, 1978.

[127] J. J. K. Ó Ruanaidh. *Numerical Bayesian methods in signal processing*. PhD thesis, University of Cambridge, 1994.

[128] J. J. K. Ó Ruanaidh and W. J. Fitzgerald. The restoration of digital audio recordings using the Gibbs' sampler. *CUED departmental report*, (TR I34 CUED INFENG), 1993.

[129] R. W. Schafer, R. M. Mersereau, and M. A. Richards. Constrained iterative restoration algorithms. *Proc. IEEE*, 69:432–450, April 1981.

[130] I. R. Sinclair. *Audio Electronics Reference Book*. BSP, Osney Mead, Oxford, 1989.

[131] M. M. Sondhi, C. E. Schmidt, and L. R. Rabiner. Improving the quality of a noisy speech signal. *The Bell System Technical Journal*, 60(8):1847–1859, October 1981.

[132] H. Spath. *Mathematical Algorithms for Linear Regression*. Academic Press, 1991.

[133] P. S. Spenser. *System Identification with Application to the Restoration of Archived Gramophone Recordings*. PhD thesis, University of Cambridge, 1990.

[134] P. S Spenser and P. J. W. Rayner. Separation of stationary and time-varying systems and its application to the restoration of gramophone recordings. *Proc. ISCAS89, Oregon*, 1:299–295, 1989.

[135] T. G. Stockham, T. M. Cannon, and R. B. Ingebretsen. Blind deconvolution through digital signal processing. *Proc. IEEE*, 63(4):678–692, April 1975.

[136] C. W. Therrien. *Decision, Estimation and Classification*. Wiley, 1989.

[137] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, 1992.

[138] H. Tong. *Non-linear Time Series*. Oxford Science Publications, 1990.

[139] P. T. Troughton and S. J. Godsill. Bayesian model selection for time series using Markov Chain Monte Carlo. In *Proc. ICASSP*, April 1997.

[140] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos. Speech enhancement using psychoacoustic criteria. *Proc. ICASSP*, II:359–362, 1993.

[141] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, 1971.

[142] J. C. Valière. *La Restauration d'Enregistrements Anciens par Traitement Numérique- Contribution à l'étude de Quelques techniques récentes.* PhD thesis, Université du Maine, Le Mans, 1991.

[143] H. VanTrees. *Decision, Estimation and Modulation Theory, Part 1.* Wiley and Sons, 1968.

[144] P. Vary. Noise suppression by spectral magnitude estimation - Mechanism and theoretical limits. *Signal Processing*, 8(4):387–400, 1985.

[145] S. V. Vaseghi. *Algorithms for Restoration of Archived Gramophone Recordings.* PhD thesis, University of Cambridge, 1988.

[146] S. V. Vaseghi and R. Frayling-Cork. Restoration of old gramophone recordings. *J. Audio Eng. Soc.*, 40(10), October 1992.

[147] S. V. Vaseghi and P. J. W. Rayner. A new application of adaptive filters for restoration of archived gramophone recordings. *Proc. ICASSP*, V:2548–2551, 1988.

[148] S. V. Vaseghi and P. J. W. Rayner. The effects of non-stationary signal characteristics on the performance of adaptive audio restoration systems. *Proc. ICASSP*, 1:377–380, 1989.

[149] S. V. Vaseghi and P. J. W. Rayner. Detection and suppression of impulsive noise in speech communication systems. *IEE Proceedings, Part 1*, 137(1):38–46, February 1990.

[150] R. Veldhuis. *Restoration of Lost Samples in Digital Signals.* Prentice-Hall, 1990.

[151] D. L. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Processing*, 30(4):679–681, 1982.

[152] M. West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society, Series B*, 46(3):431–439, 1984.

[153] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications.* MIT Press, 1949.

[154] E. Zwicker and U. T. Zwicker. Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. *J. Audio Eng. Soc.*, 39(3):115–126, 1991.

Index