Preprint 3013 (A-6)

Problems of error masking in audio data compression systems

Michael A. gerzon
Technical consultant, Oxford, UK

Presented at the 90th Convention 1991 February 19-22 Paris

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

PROBLEMS OF ERROR-MASKING IN AUDIO DATA COMPRESSION SYSTEMS

Michael A. Gerzon

Technical Consultant, 52 Walton Crescent, Oxford OX1 2JQ, U.K.

Abstract

This paper notes the failure of spectral masking of coding errors by wanted signals when the error is highly cross-correlated with the signal. This can reduce masking thresholds by 30 dB. The existence of such cross-correlation is proved for all Shannon-efficient audio data compression systems. At low bit rates, these can exhibit up to 4dB gain modulation. An improved cross-spectral class of models for masking taking account of these effects is discussed.

1. INTRODUCTION

In recent years, there has been much interest in developing systems for data compression of very high quality audio into bit rates of less than about 4 bits/sample/audio channel. Any system operating at such low bit rates produces objectively quite large error signals, and the perceptual quality of such systems relies on the error in the coding/decoding process being masked by the wanted signal.

It is the object of this paper to note that conventional models for spectral masking of errors break down in certain situations, and that errors as much as 30 dB smaller than those deemed inaudible by conventional models of spectral masking can be heard in some situations. A further aim of this paper is to suggest precisely aspects of what causes spectral masking models to break down - briefly we suggest that masking breaks down when the error is highly cross-correlated with the wanted signal.

A specific problem that renders errors audible is identified - namely the effect of varying signal-dependent amplitude modulation of wanted signal components. Such amplitude modulation produces an "unstable" effect akin to the familiar symptoms of mistracking in noise reduction systems. By means of a general geometric argument from Shannon information theory, this paper proves that all Shannon-efficient coding systems produce highly significant amounts of amplitude modulation of the wanted signal, and that the error-signal in such coding systems is highly cross-correlated with the wanted signal. Despite the generality of this result, we are unaware of any prior account of it in the literature apart from a recent early version of this paper [1].

We go on to propose a modified model for spectral masking which involves not only the spectra of the wanted and error signals, but also their cross-spectra, which takes into account cross-correlations between the signal and the error. An appendix discusses the notion of short-term spectra and cross-spectra and their measurement. Cross-spectral ideas are also applied to the problem of directional masking in stereo signals, since conventional monophonic masking can break down if the direction of the error signal differs from that of the wanted signal.

The paper also discusses some strategies for audio data compression that can reduce or even totally eliminate amplitude modulation and error correlation effects. These all involve modifying the quantisers for signal components, and, by the earlier result, inevitably compromise the Shannon-efficiency of the coding.

While much of this paper is necessarily at a quite theoretical level, it is important not to be "blinded by the science" into not realising the empirical significance of this work in high-quality audio applications. The theoretical complexities of coding theory (see the collection of references [2]) can mean that experts in that theory can be inadequately aware of the perculiar subtleties of high quality audio. We therefore deem it appropriate to begin with a general discussion of the specific problems associated with high quality audio from the viewpoint of skilled audio professionals, since we feel that traditional engineering criteria often do not take adequate account of these.

2. HIGH QUALITY AUDIO

Traditional work on audio data compression dating from the 1960's and 1970's [2] was largely aimed at telephone-grade audio, and it is tempting to regard high-quality audio data compression as being quantitatively but not qualitatively different from these. Among the obvious quantitative differences between low- and high-quality audio are the following:

(i) improved signal-to-noise ratio (up to around 100 or 120 dB)
(ii) enlarged dynamic range
(iii) widened frequency range to beyond 20 kHz
(iv) lower required modulation noise
(v) lower required nonlinear distortion
(vi) a wider range of signal statistics
(vii) tightened frequency and phase response flatness
(viii) a wider range of transient attack and decay characteristics
(ix) a wider dynamic range between different simultaneously-occuring components of the power spectrum.

A classic text of G. Slot [3] from the early 1960's is an excellent summary of this traditional engineering approach that regards high quality audio as differing from low-quality audio mainly in these quantitative aspects.

Over the last 15 years, both among hi-fi experts and among studio professionals, it has become widely believed that these quantitative improvements are not, by themselves, enough to ensure excellent performance on high-quality program material. While this has given rise to heated debate between a so-called "objectivist" and "subjectivist" schools of thought, it is not our intention to enter into this debate, but simply to note that, compared to telephone-grade audio, there are also some profound qualitative differences in high-quality audio not encompassed by the quantitative differences listed above.

In telephone-grade audio, the signal generally consists of just a single sound source (a human voice) conveying primarily a verbal message, and also some cues conveyed by tone of voice and articulation about emotional content. In contrast, a high-quality music signal can consist of numerous separate sound sources - e.g. 100 orchestral instruments, or up to 48 mono sounds from a multitrack tape - plus additional reverberant and ambient

information derived either from a large number of reflections from the boundaries of an actual room or from a large number of effects units such as digital reverberators. Moreover, the music is derived from an interactive process whereby each musical line affects the performance of the others. Thus high-quality audio not only conveys a large number of separate messages, each of which can be followed separately, but there are numerous inter-relationships between these messages, and a listener might choose to listen at a level that concentrates on these complex inter-relationships. The process of listening, and of determining what kind of messages or inter-relationships between messages are derived by the listener, is well beyond current knowledge, and we have no good theoretical model for the listener's processing of information.

This ignorance does not preclude the attempt to find objective parameters that correlate with the subjective performance of a high-quality audio system, but it does suggest that parameters derived either from tests on simplified test stimuli or from tests on listeners with a narrow range of analytic listening modes on complex stimuli may be misleading.

Because of this difficulty, any "objective" hypothesis about acceptable signal degradation should be treated conservatively, with a constant suspicion that there may be circumstances in which much smaller degradations may still be audible. The use of "objective" hypotheses should ideally be backed up by a detailed model of how such degradations might be measured - even if the measurement is in practice difficult - in order to avoid untestable hypotheses. In the design of high-quality coding systems, one should, as far as reasonably possible, eliminate any faults capable of being avoided altogether.

Besides the qualitative difference of a multiplicity of sound sources, high-quality audio also differs from telephone-grade audio in its use of stereo. Stereo signals cannot be treated merely as two separate mono channels, but the precise relationships between the signal components in the two channels are also important.

In engineering for high-quality audio, it is a definite advantage to have extensive practical experience of the problems of recording and reproducing sound, and of the kind of cues a listener is capable of perceiving in recordings - there is much that is not in any text book or research paper that is a matter of experience among practitioners. By way of example, it is found that certain recording techniques [4] and equipment preserve cues about the distance of sounds - and it is thought that these cues involve preserving the amplitude and time delay of early reflections. It is a matter of subjective experience that much highly-specified equipment disturbs or destroys the sense of distance, whereas other more poorly specified equipment can preserve it. Although the factors involved are not well understood, it is believed that accurate preservation of signal envelope information may be important, and that signal-dependent alterations of gain too small to be audible on simple test stimuli may be enough to damage this cue.

3. SPECTRAL MASKING

It is not the intention of this paper to specify any detailed model of how spectral masking works, but merely to discuss a class of models, and

to show that one has to go outside this whole class of models to get reliable information about masking.

Spectral masking is based on the following general idea: If one has two signals, a wanted signal and an error signal, then the masking of the error signal by the wanted signal can be predicted purely from the power spectra of the two signals. From the power spectrum of the wanted signal, one computes a spectral threshold, and if the error power spectrum lies below this threshold at all frequencies, the error is presumed to be inaudible.

This basic theoretical model of spectral masking needs to be fleshed out with quite a bit of detail to be useful for predicting error audibility. First, the model only fully applies to signals with stationary statistics, and under transient signal conditions, it is important to ensure quite a high degree of temporal coincidence of the short-term power spectra of the signal and error (to with the order of 2 ms) to use spectral masking - and there is an uncertainty as to what precisely constitutes a short-term power spectrum for this application. In practice, the trade-offs for assessing spectral masking for transient signals are determined in a partly empirical way.

The actual determination of the spectral threshold for a given wanted-signal power spectrum is also not completely understood, but the general procedure used is along the following lines. One first determines experimentally the threshold for narrowband wanted and error signals as the frequency of the two signals varies. Such masking curves are very familiar [5,6] and for each masking frequency resemble the power response of a moderately high-Q resonant filter, at least until they reach down to the level of the absolute threshold of hearing.

For a complex signal with an extended power spectrum, the spectral threshold is determined by a convolution-type method - although no-one can be sure that any precise procedure is quite right. Essentially, for each frequency, one computes the spectral threshold curve at that frequency for the signal power within a critical bandwidth of that frequency, and then computes an integral, weighted by the power density of the power spectrum, of all the values of the threshold spectrum caused by different frequency components of the wanted-signal spectrum. This computation in effect computes the original power spectrum convoluted, in a frequency-dependent fashion, by the masking threshold curves for each frequency.

It is specifically found that error signals within the critical band of a masking signal are masked at levels only a few dB lower - typically between 4 and 11 dB down. Although the masking threshold falls rapidly as the difference between error and wanted signals increases, when the two signals have similar frequency content, errors only a few dB down are well masked.

Although the precise details of concrete spectral masking models may be somewhat uncertain, the above encapsulates all we need to know about these models. We shall show that this class of models, as stated above, is conceptually flawed and not in accordance with the facts, and suggest an alternative that includes the undoubted successes of spectral masking models while taking account of situations where they fail to work.

4. ERROR CORRELATION

We now demonstrate that spectral masking models fail to work in some situations by means of a concrete and well-understood example. It is well known that for middle frequencies and normal middle sound levels, the ears can typically hear gain changes of the order of 1 dB, and that at the most critical levels and frequencies, gain changes of 0.3 dB are audible.

Our example is simply to consider an amplitude modulated signal with a 0.3 dB gain change, where the gain change is sufficiently slow to prevent modulation sidebands being widely separated in frequency from the original signal - e.g. if the modulation waveform has no components above say 20 Hz, and where the extreme gains are held long enough for the ear to register the gain change - for example consider a sinewave modulated by a bandlimited step waveform, where the gain varies from 1 - 0.016 to 1 1 + 0.016, as in figure 1. The total gain variation is around 0.3 dB. However, compared to the original signal, the error signal has amplitude gain varying between -0.016 and +0.016 - i.e. is about 36 dB below the original signal. (see figure 1).

In this example, we have seen that an error signal with identical power spectrum to the wanted signal can have an audible effect even if 36 dB down - yet a spectral masking model would predict that it should not be audible until it is between 4 and 11 dB down.

This example is rather a trivial one, and commonsense would prevent anyone from attempting to use a spectral masking model to predict the audibility of the error. However, it does illustrate that certain kinds of error signal are much more poorly masked than others with an identical power spectrum. The problem arises with much more complex wanted and error signals. How can one know whether an error is of a kind to which spectral masking theory can be applied, and when it is of an unsuitable kind?

It is certainly not possible to give entirely definitive answers to this question, but there is a reasonable way of distingishing amplitude modulation type error from other errors. This is to look not just at the spectra of the wanted and error signals, but also at their cross-spectra. This is related to the generally more familiar notion of cross-correlation.

For two signals $f(t)$ and $g(t)$, functions of time t, their cross-correlation is the function

$$C_{f,g}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} f(t)g(t+\tau)\, dt \ . \qquad (1)$$

The cross-correlation of a signal with itself is termed the auto-correlation of that signal, and its Fourier transform is well-known to be the power spectrum of that signal. In a similar manner, the Fourier transform of the cross-correlation between two signals is termed the cross-spectrum of the two signals. While the power spectrum of a signal is positive at all frequencies, the cross-spectrum is a complex-valued quantity.

The theory of cross-spectra is given in the textbook of Yaglom [7]. It is often convenient to specify the spectra and cross-spectra of signals by means of the spectral matrix. Let $f(t)$ be a wanted signal, and let the

modified signal after passing through a coding and decoding system be denoted by $Qf(t)$, so that the error signal is

$$\varepsilon(t) = Qf(t) - f(t) \ . \tag{2}$$

Then the spectral matrix of the signal f and error ε is the matrix

$$\begin{bmatrix} S_{f,f}(F) & S_{f,\varepsilon}(F) \\ S_{\varepsilon,f}(F) & S_{\varepsilon,\varepsilon}(F) \end{bmatrix}, \tag{3}$$

where $S_{f,g}(F)$ as a function of frequency F denotes the cross-spectrum, i.e. the Fourier transform of equ. (1), of two signals f and g. The cross-spectral matrix is a <u>positive</u> matrix, which means that:

$$S_{f,f}(F) \geq 0 \tag{4a}$$

$$S_{\varepsilon,\varepsilon}(F) \geq 0 \tag{4b}$$

$$S_{f,\varepsilon}(F)^* = S_{\varepsilon,f}(F) \tag{4c}$$

and

$$\left|S_{f,\varepsilon}(F)\right|^2 \leq S_{f,f}(F)\, S_{\varepsilon,\varepsilon}(F) \ , \tag{4d}.$$

where * indicates taking the complex conjugate of a complex number.

If we denote the real and imaginary parts of a complex number α by $Re\alpha$ and $Im\alpha$ respectively, then we can specify the degree of cross-correlation of the signal $f(t)$ with the error signal $\varepsilon(t)$ at frequency F by the two quantities:

$$-1 \leq Re\left\{S_{f,\varepsilon}(F)\right\} / \left\{S_{f,f}(F) S_{\varepsilon,\varepsilon}(F)\right\}^{\frac{1}{2}} \leq 1 \tag{5}$$

and

$$-i \leq Im\left\{S_{f,\varepsilon}(F)\right\} / \left\{S_{f,f}(F) S_{\varepsilon,\varepsilon}(F)\right\}^{\frac{1}{2}} \leq 1 \ , \tag{6}$$

which are respectively the real and imaginary parts of the <u>correlation index</u> of the signal and error at frequency F.

If the error has the form of a random noise signal, or of an amplitude-modulated random noise signal, then the cross-spectrum, and hence cross-correlation is zero. Experiments on masking normally use signals with zero cross-spectrum. However, in the case of an amplitude-modulation error, where $\varepsilon(t)$ is an amplitude modulation of $f(t)$, then the cross-spectrum becomes nonzero, with a zero imaginary part, but with a real part that causes the correlation index to become ± 1. If instead, the error signal is caused by a (small) degree of phase modulation of $f(t)$, then it can be shown that the real part of the correlation index is zero, but that the <u>imaginary</u> part becomes near $\pm$ 1.

Thus a non-zero cross-spectrum between signal and error is indicative of the error containing components caused by gain errors (for real parts of the cross-spectrum) at frequency F, and components caused by phase errors (for imaginary parts of the cross-spectrum). In the absense of gain and phase errors, i.e. when the cross-spectrum is zero, it is reasonable to use spectral masking theory to predict error audibility.

However, if the cross-spectrum is non-zero, and especially if it corresponds to a continually fluctuating gain or phase, then it is not to be expected that spectral masking theory will necessarily be applicable.

We are unaware of any reliable data on the perceptual effect of signal-dependent gain and phase modulation errors, although it is known that errors caused by random amplitude modulation are much more audible than those of a similar energy caused by random phase or frequency modulation. Empirical experience with the design of dynamic filters, including both those with phase compensation (which do not suffer from phase modulation effects) and those without (which do), suggests that when there is signal-dependence, phase and amplitude modulation may have similar degrees of audibility, but that they have perceptually different qualities.

Noiselike errors with zero cross-spectrum tend to sound like added noises rather than as modifications of the quality of the wanted signal (although if the noise spectrum imitates that of the wanted signal closely, this partially ceases to be true, and the error then takes on a distinctive "vocoder" quality rather like the wanted signal with an added "gargling" effect). In contrast, amplitude modulation errors which have a real cross-spectrum with the signal tend to cause a distinctive quality of "instability" which is perceived not as an added sound, but as a modification of the character of the wanted sound. This effect is familiar in dynamic processors and also in drop-out effects with analogue tape. Phase-modulation errors are also heard as a modification of the quality of the wanted signal, but have a perceptual effect commonly associated with "phaser" effects devices, which is heard as pitch changes in components of the wanted sound - providing the phase modulation is not too fast.

In the absence of experimental data, it is difficult to quantify the degree to which different kinds of error are audible, but in general, the effect of errors with large cross-spectrum is quite different from that of uncorrelated errors, and the former type of errors are capable of being heard in much smaller quantities. We shall later discuss the possible uses of the cross-spectrum to study masking of errors in more detail.

However, we have already seen that amplitude modulation by around 0.3 dB is certainly audible in some situations, so that it seems wise to keep any amplitude modulation effects in coding/decoding systems at least down to this level. Experience suggests that on critical high-quality audio material containing complex distance cues, much smaller degrees of signal-dependent amplitude modulation may still have an audible effect on such cues, so that ideally, amplitude modulation effects should be virtually eliminated in systems intended for very high quality use.

The above cross-spectral account of gain and phase modulation effects has an important weakness - namely that the notion of cross-correlation (1), and hence of cross-spectrum, used involved taking an average over all time. In practice, one needs a notion of short-term spectra and cross-spectra that involve averages taken typically over time intervals of the order of 30 to 50 ms, but which preserves the positivity of the spectral matrix. The general theory of such short-term spectra and cross-spectra appears not to have been published elsewhere, and is too complex for the main body of this paper. However, we present a summary of such a theory, based on that for the Wigner distribution, in appendix A of this paper. That appendix also describes relatively simple means of measuring short-term spectra and cross-spectra of signals and errors.

5. GAIN MODULATION IN EFFICIENT CODING

This section gives a theorem that is the central theoretical result of this paper, which asserts that Shannon-efficient coding/decoding systems always have gain errors in the wanted signal, and the error-signals of such coding systems are cross-correlated with the wanted signal. The gain error is preceisely quantified and shown to be of a magnitude likely to have audible consequences in low bit rate systems.

There is a long-known basic theory, known as Shannon Rate-Distortion theory (see Davisson [8]) that, for any given wanted-signal power spectrum and an upper bound on the desired error-signal power spectrum, gives a lower bound, in principle attainable to an arbitrarily close approximation, on the bit rate required to convey the wanted signal with the des ired error spectrum. We do not need the details of this theory here, see [8], but note that this theory is strictly applicable only to wanted signals with Gaussian statistics, but actually works reasonably well also for the mildly non-Gaussian statistics of typical audio signals. Practical coding systems approximate reasonably well to the bit rates of ideal Shannon coding for a given wanted-signal power spectrum and desired error--signal power spectrum, giving of the order of 2 dB greater noise power than the ideal Shannon case.

Coding systems approximating the Shannon ideal include principal-value coding methods [9,10,11] which involve separate quantisation of a fairly large number of transform or spectral-band components, and adaptive pulse code modulation (ADPCM) or predictive coding methods [10,12], which use prediction with noise shaping [13] of the error component to achieve similar results. There also exist hybrid systems, such as the Solid State Logic apt-X 100 system, that split the audio signal into a small number of rather wide frequency bands and use predictive coding within each band. A layman's summary of the potential advantages and problems in these approaches is given by the author in ref. [14].

The Shannon theory is at the root of all modern efficient audio coding methods, and all approximations to ideal Shannon coding attempt to minimise a (suitably spectrally weighted) rms error energy for the bit--rate used within the class of coding strategies used. Although there has been much questioning of the appropriateness of wieghted rms error criteria in the coding literature (e.g. see [8]), this questioning has generally taken the form of finding an alternative error magnitude measure such as a weighted n-th root mean n-th power measure, but in both audio and image coding applications, no such measure that reliably predicts the perceptual effects of error signals has been found.

We are unaware of any study that concentrates not on the magnitude of the error itself (however weighted and computed) but on the cross-correlation of the error with the wanted signal - but we have seen earlier that this may be perceptually far more important in some situations. (Although not strictly applicable to this paper, we report that we have found that cross-correlations between the wanted and error signals are also of perceptual importance in video and image coding systems). This may be a reason why the following theorem in coding theory is not generally known. We are unaware of any account of this result in the literature, although this may simply be our own ignorance.

The definition of "efficiency" in Shannon coding theory has involved minimising a (suitably weighted) mean square error energy, and coding strategies that do not minimise error energy (with an appropriate perceptual weighting) have been regarded as "inefficient". Since these terms have a judgemental quality, it is relevant to note that such notions of "efficiency" are based on a very narrow technical criterion that in general may be perceptually inappropriate.

Let the wanted signal be f and the result of coding and decoding that signal be Qf, and denote the error signal by

$$\varepsilon f = Qf - f \ . \tag{7}$$

Moreover, denote the (weighted) r.m.s. level of a signal f by the notation $\|f\|$. Signals can be regarded as vectors in an abstract space (known technically as Hilbert Space [15]) having a geometric length $\|f\|$ in that space. Although we shall use geometrical arguments in this space in a heuristic commonsense fashion, the arguments used can be made mathematically rigorous in Hilbert Space [15].

Theorem Let f, Qf and $\varepsilon f = Qf - f$ be respectively a wanted signal, a signal that results from the wanted signal passing through a coding and decoding system, and the resulting error signal. Suppose further that we have a specified class of encoding/decoding systems such that whenever one system within the class gives result Qf, another system within the class will give results kQf for arbitrary positive gain k. Further, suppose that the coding system used is "efficient" in the Shannon rms sense, i.e. that within the specified class of coding/decoding systems and for a chosen weighted rms measure $f \rightarrow \|f\|$, the coding system minimises the rms error $\|\varepsilon f\|$.

Then

$$\|f\|^2 = \|Qf\|^2 + \|\varepsilon f\|^2 \tag{8}$$

and the component of Qf cross-correlated with f equals

$$\hat{f} = \frac{\|Qf\|^2}{\|f\|^2} \ f \ , \tag{9}$$

which has an amplitude gain

$$\|Qf\|^2/\|f\|^2 \ = \ 1 - (\|\varepsilon f\|^2/\|f\|^2) \ . \tag{10}$$

The component of the error εf cross-correlated with f equals

$$\frac{\|\varepsilon f\|^2}{\|f\|^2} \ f \tag{11}$$

and the rms magnitude of the component $\hat{\varepsilon}$ of the error εf un-cross-correlated with f is

$$\|\hat{\varepsilon}\| = \ (\|Qf\|/\|f\|)\|\varepsilon f\| \ . \tag{12}$$

In other words, this theorem asserts that "efficient" coding systems, in the Shannon rms sense described earlier, inevitably involve amplitude modulation of the wanted signal with gain $1 - (\|\varepsilon f\|^2/\|f\|^2)$. We first sketch the proof of the theorem before applying it to cases of practical consequence.

In figure 2, we show the two vectors representing the wanted signal f and the coded/decoded signal Qf, along with the vector along the third side of the triangle representing εf. If the length $\|\varepsilon f\|$ of εf is supposed to be minimised, then since kQf is, for arbitrary k, a permissible coded/decoded signal within the class of possible coding/decoding systems considered, then we conclude that the vector εf must be at right angles (or "orthogonal") to Qf as in figure 3, so that its length is minimised. By Pythagoras' theorem applied to the largest right angled triangle in figure 3, we get equ. (8) of the theorem.

The component of Qf cross-correlated with f is obtained by taking the orthogonal projection $f\hat{}$ of Qf as shown in figure 3, and simple geometry shows that $f\hat{}$ in figure 3 is given by equ. (9) of the theorem. Equ. (10) of the theorem follows from equs. (8) and (9). The component of εf cross-correlated with f is the orthogonal projection $f - f\hat{}$ of εf onto f as shown in fig. 3, and its length is $\|f - f\hat{}\| = \|f\hat{}\|(1 - [\|Qf\|^2/\|f\|^2])$ by equ. (10) $= \|f\|(\|\varepsilon f\|^2/\|f\|^2)$ by equ. (8), which gives equ. (11).

Finally, the component $\hat{\varepsilon}$ of εf un-cross-correlated with f is that component orthogonal to f shown in figure 3, and by similarity of triangles,

$$\|\hat{\varepsilon}\|/\|\varepsilon f\| = \|Qf\|/\|f\| ,$$

which gives equ. (12), completing the sketch proof of the theorem.

In principal-value coding systems [9], we can quantify the actual amplitude gain (9) and the uncorrelated (i.e. noise-like) error magnitude actually encountered in practical systems that use quantisers that minimise (weighted) r.m.s. error, such as those of J. Max [16] for Gaussian signal statistics, or those of Paez and Glisson [17] for Laplacian and Gamma probability distribution function signal statistics. In these papers, the error energy $\|\varepsilon f\|^2/\|f\|^2$ for least rms error n-level quantisers were computed, both for the cases of equilevel quantisers and for optimum (in the least error energy sense) non-uniform quantisers, and the associated optimum coding entropy (bit rate) for these quantisers was also computed by Max - we have performed a similar computation for the entropy of the quantisers of Paez & Glisson [17] for the Laplacian signal statistics case. Wood [18] has shown that, for a given error energy, equilevel optimum quantisers generally give the lowest value of coding entropy (bit rate), although this is only an approximate result.

Based on the results tabulated by Max [16] for equilevel optimum quantisers for Gaussian signal statistics, we have used equs. (10) and (12) to compute for each n the gain modulation and un-cross-correlated error signal level (in dB relative to the wanted signal level) for a Max n-level quantiser, and the results are tabulated in Table 1 and figure 4. The last column of table 1 will be explained later. The "gain" column shows the amplitude gain of that component $f\hat{}$ of Qf cross-correlated with f relative to that of f, and the "uncorrelated noise" column the level in dB of $\hat{\varepsilon}$ relative to that of f.

Table 2 shows similar results based on the data of Paez and Glisson [17] for "optimum" equilevel n-level quantisers for Laplacian signal statistics - a statistics often approximated in the coding of many audio signals such as speech and some "spikey" musical waveforms.

For Gaussian signal statistics, it will be seen from table 1 and figure 4 that the gain errors at low bit rates exceed 0.3 dB for Max quantisers with 8 levels or less, corresponding to (entropy coded) bit rates below 2.76 bits. Thus all audible signal components encoded at below this bit rate are liable to perceptually significant gain errors. These errors reach a dramatic 3.92 dB for a rate of 1 bit, 1.83 dB for 1.536 bits, and 1.10 dB even at a rate of around 2 bits. The gain error does not fall below 0.1 dB (which may well still be perceptually significant in practice) until more than 16 levels, i.e. a bit rate of over 3.6 bits, are used. The results for gain error versus bit rate are very similar for equilevel quantisers matched to Laplacian signal statistics, as shown in Table 2.

Therefore, transform or spectral-band coding systems using "optimum" Max-type quantisers matched to signal statistics in each band at a bit rate of less than about 3 bits in any perceptually significant band at any time are liable to suffer from perceptually significant gain modulation effects and errors.

Unfortunately, even this still understates the true gain errors encountered in encoding systems using Max-type quantisers, for several reasons:-

(i) in practice, not the whole bit rate allocated to each sub-band or transform component is allocated to conveying just the quantiser outputs - a significant overhead is taken up by the error protection redundancy required with entropy coding, and by additional information required to convey the quantiser gain, quantiser statistics and bit-allocation data.

(ii) the results of tables 1 and 2 assume that the quantiser is perfectly matched to the instantaneous signal statistics. In practice, it is impossible to have certain knowledge of the actual signal statistics at each moment due to statistical fluctuations in the means used to estimate them. This results in an additional, generally very significant, unpredictable gain fluctuation of the quantised signal component due to the mismatch.

(iii) if n stages of coding and decoding are used in series, then random-type noise errors build up in power proportional to n, but correlated (e.g. gain) errors build up in amplitude proportional to n, since they add up coherently, giving an error power proportional to n^2. Thus the effect of cascading coding/decoding systems affects gain errors much more severely than random (uncorrelated) noise-type errors. If, for professional uses, one requires that the results of up to 10 stages of coding and decoding to be permissable (as suggested in [14], and often required by broadcasters), with a combined gain error of 0.3 dB, then one has to use a bit rate of not less than 4.45 bits for coded signal components according to table 1 - which would mean in practice using systems coding at a rate of at least 5 or 6 bits per sample per audio channel.

(iv) There is also a distinct possibility that for critical audiophile material, the ears might actually hear degradations due to signal-dependent fluctuating gain errors well below 0.3 dB. Most workers aware of this possibility believe that gain fluctuations should be below 0.1 dB, and a few have suggested that much smaller fluctuation - of the order of 0.01 dB or even 0.001 dB might have

audible significance for critical listening over good audio equipment to very high quality programmes. Such a worst-case situation would require the use of Max quantisers with over 160 levels for a 1-pass encoding/decoding system, and not less than 500 levels for a 10-pass professional coding/decoding requirement. Such systems, coding at 9 or more bits per sample per audio channel, would no longer qualify for the appelation "low bit rate"! Admittedly, this is a worst-case extreme, but it does illustrate the potential difficulties encountered in audiophile applications with gain modulation effects.

6. REDUCING GAIN MODULATION EFFECTS

The above results suggest that, in some situations, serious signal--dependent gain modulation effects of some signal components will be encountered with all practical low-bit-rate coding systems using bit allocation among transform or spectral band signal components, if the systems use Max-type optimal quantisers. We therefore seek strategies for reducing or eliminating such gain modulation effects.

There are two broad strategies that can be used: the first strategy is to continue to use Max quantisers but to increase the reproduced gain of the quantised signal to compensate for the computed gain loss in the quantiser. Such "gain compensated" Max quantisers produce a coded/decoded signal $Q'f$ and error $\varepsilon'f$ as shown in figure 5, and here the error vector is at right angles to the wanted signal f, so that $\varepsilon'f$ is un-cross-correlated with the wanted signal f.

One price to be paid is that the error energy is now increased by the square of the amplitude gain $\|f\|^2/\|Qf\|^2$, as shown in the last columns of tables 1 and 2. The resulting noise energy is larger than that of the original quantisers without gain compensation listed in refs. [16] and [17], since the gain compensated quantiser is no longer designed to minimise mean square error energy without constraint, but rather to minimise the mean square error energy subject to the constraint that the error have zero cross--correlation with the wanted signal. Inevitably (according to the theorem of the last section), this reduces the Shannon coding efficiency, to the extent that a 3-level gain compensated Max quantiser has about the the same un-cross-correlated noise level as a two-level Max quantiser (see table 1). This increased noise level may require some alteration of the bit-allocation strategy in coding systems using low bit rates in order to retain masking of noise.

However, while gain compensated Max quantisers will reduce systematic gain errors, they will not overcome the problem that one does not have a priori knowledge of signal statistics, and that, especially under transient signal conditions, there may be a marked mismatch between assumed and actual signal statistics, resulting in unpredictable gain errors (item (ii) above).

If a signal indeed has Gaussian statistics (and even this is

uncertain), then a quantiser optimised for a different r.m.s. signal level σ to that σ_A of the actual signal will suffer from a gain error equal to the ratio of the length of the orthogonal projection of the gain-compensated coded signal $Q'f$ onto f (see fig. 6) to that of f. While we have not systematically computed the graph of this gain error against σ/σ_A , it is very dependent on the precise quantiser used. For example, 2-level and 3-level gain-compensated Max quantisershave a gain curve of the general form shown schematically in figure 7.

It is possible to design quantiser characteristics using 3 or more levels that minimise the gain error caused by errors in σ/σ_A near $\sigma/\sigma_A = 1$, but such quantisers will no longer be gain-compensated Max quantisers, and will suffer an even larger loss of Shannon coding efficiency. While we believe that the theoretical optimisation of such quantisers which are gain-insensitive to level mismatch will be a worthwhile excercise, one then also has the additional problem of minimising gain errors also for departures from Gaussian statistics. The more one attempts to make the quantiser gain insensitive to mismatches in both level and statistics, the poorer is the Shannon efficiency, and the larger the number of quantisation levels required.

Pending the results of detailed numerical design studies, it is our best guess that the gain-compensated quantiser route to reducing signal-dependent gain modulation effects is likely to result in poor coding efficiency in the quantiser, and we are sceptical that this approach is the best way forward.

A second, and much more promising, route for reducing gain modulation effects is to replace deterministic quantisers by stochastic quantisers - i.e. those whose performance is probabilistic. Such a quantiser was first described by Roberts [19], namely the subtractively dithered quantiser [20] shown in fig. 8. In this quantiser, a pseudo-random noise ("dither") is added to the signal which is then quantised by an n-level equilevel quantiser, and is decoded by subtracting a synchronised reconstructed replica of the dither from the output of the quantiser. The optimum dither noise in this application is that with a uniform probability distribution function (pdf) with peak-to-peak level difference equal to the step size of the quantiser.

Such a subtractive uniformly dithered quantiser gives an output which equals the input plus an uncorrelated uniform pdf noise signal, provided only that the input signal level does not exceed the + or - peak quantiser levels. Larger signals suffer from clipping distortion. (See [19-20].

So, provided that one chooses an n-level quantiser whose peak levels exceed that of the signal component to be quantised, a Roberts subtractively dithered quantiser gives an error that has zero cross-correlation with the wanted signal, and no gain error (or, indeed, any nonlinear distortion) of the wanted signal.

However, subtractive dither only works correctly with equilevel ("uniform") quantisers, although entropy coding of the output of such a quantiser can be used to reduce the transmitted bit rate.

Unfortunately, unlike the case of undithered equilevel quantisers, one cannot accept occasional large-level signal excursions beyond the peak levels of the subtractively dithered quantiser, since these cause clipping and amplitude modulation effects. This means that much greater care has to be taken to scale the gain of the signal to be quantised to avoid clipping, and for a given signal-to-noise ratio one has to use quantisers with many more levels - which requires the use of more complicated entropy coding strategies to bring the bit-rate back down again.

Thus the use of subtractive dither somewhat complicates the practical design of quantisers and their associated entropy coding, bit rate allocation and gain ranging. In general, the use of subtractive dither round an n-level quantiser reduces its signal -to-noise ratio to that associated with an (n-1)-level undithered equilevel quantiser, resulting in a significant loss of Shannon efficiency at low bit rates.

However, in compensation, the subtractively dithered equilevel quantiser is insensitive to signal statistics, since such mismatches do not cause any error cross-correlation or gain modulation effects, but only an increase in the bit rate required for a given signal- -to-noise ratio.

We conclude that, by replacing the quantisers in existing systems with subtractively dithered quantisers, and modifying the associated bit-allocation and entropy-coding strategies accordingly, it is possible to design coding systems free of any gain modulation or error/wanted signal cross-correlation effects - and indeed free of any nonlinear distortion effects. The price to be paid is some increase in bit rate and a possible increase in quantiser complexity. In a future publication, we hope to present detailed design methods for optimising subtractively-dithered coding systems entirel free of error/signal cross-correlation effects and without any non- linearity. We are somewhat more pessimistic about the prospects for systems using modified and gain-compensated undithered quantisers.

7. CROSS-SPECTRAL MASKING

Although we have just shown that it is possible to design coding systems avoiding cross-spectral error components altogether, not all coding systems use subtractively dithered quantisers or may have other sources of nonlinearity. Thus it is important to have some models for studying the masking of errors when cross-spectral errors do occur. This section is devoted to a speculative generalisation of spectral masking models that incorporate such effects.

Also, it is not entirely certain, in the presence of uncorrelated noiselike error components, that a zero cross-spectrum is necessarily subjectively optimal. For example, it may conceivably be that the ear could prefer a coded/decoded signal that neither minimises (weighted) error energynor one which eliminates error/signal cross-

correlation, but might instead prefer, for example, a coded signal that has identical spectral energy content to the original signal. If Qf is the Max-quantised signal, then this "unchanged spectral power" stategy would imply that the ear might prefer a decoding that reconstituted

$$\frac{\|f\|}{\|Qf\|}\, Qf \tag{13}$$

rather than Qf (Max quantisation) or $(\|f\|^2/\|Qf\|^2)Qf$ (gain-compensated Max quantisation).

Interestingly, both the gain-compensated Max quantisation strategy that minimises cross-correlation between the error and wanted signal, and the unchanged spectral power strategy retain the reqired property after any number n of stages of coding/decoding, whereas even just two stages of coding/decoding using Max quantisers no longer preserves the Max quantisation property, as seen from figure 9.

The class of models we shall propose for the audibility of cross--spectral error components has the following general form. For a given wanted-signal power spectrum, instead of producing a single "spectral thresshold" curve, we propose that 3 separate curves $T_1(F)$, $T_2(F)$, $T_3(F)$ be derived, one $T_1(F)$ describing the masking of the error power spectrum $S_{\epsilon,\epsilon}(F)$ as in conventional spectral masking theory, the second $T_2(F)$ describing the masking of the real part $\mathrm{Re}S_{f,\epsilon}(F)$ of the cross-spectrum, and the third $T_3(F)$ describing the masking of the imaginary part $\mathrm{Im}S_{f,\epsilon}(F)$ of the cross-spectrum..

We suggest that the requirements for masking be

$$\begin{aligned} S_{\epsilon,\epsilon}(F) &< T_1(F)\ , \\ |\mathrm{Re}S_{f,\epsilon}(F)| &< T_2(F)\ , \\ |\mathrm{Im}S_{f,\epsilon}(F)| &< T_3(F)\ . \end{aligned} \tag{14}$$

It is not suggested that such a model is the most general possible, since one can consider a more general model in which one has a 'threshold region' R(F) derived from the spectrum $S_{f,f}(F)$ which is a solid region in 3-dimensional space, depending on frequency F, and have masking if and only if the vector

$$(S_{\epsilon,\epsilon}(F),\ \mathrm{Re}S_{f,\epsilon}(F),\ \mathrm{Im}S_{f,\epsilon}(F)) \tag{15}$$

lies within the threshold region R(F) for all frequencies F. However, equ. (14) is at least a starting point not introducing too many variables for experimental investigations.

The earlier results suggest that $T_2(F)$ and $T_3(F)$ will generally have much lower values than the conventional masking threshold $T_1(F)$. For example, a -6 dB spectral masking threshold corresponds to $T_1(F) = \frac{1}{4}S_{f,f}(F)$, and a -36dB threshold for gain-type errors corresponds to $T_2(F) = 0.016S_{f,f}(F)$.

In general, looking at masking of cross-spectral as well as spectral components experimentally should allow spectral masking models to have an improved predictive value. However, one needs to look not

only at the case where the spectra and cross-spectra are stationary, but where they vary with time. For this purpose, one will need to use a notion of short-term spectra and cross-spectra of the type discussed in appendix A.

Nevertheless, a spectral matrix approach including cross-spectra as well as spectra of error signals provides a unified theoretical framework for investigating the simultaneous audibility of simultaneous gain errors, phase errors and uncorrelated noise errors, errors which hitherto have been regarded as isolated and distinct in nature. By representing such errors as components of the spectral matrix, one has a space of parameters which can be subjected to experimental investigation.

However, investigations of the case where the spectral and cross--spectral components are stationary with time will probably not be adequate to reveal the audible effects of fluctuating and signal-dependent errors. There is some evidence that fluctuating errors can be significantly more audible than stationary errors.

Additionally,it is now well established (see Moore [21]) that if the errors are un-crosscorrelated, but if the wanted signal has fluctuating levels in different frequency bands that vary in a mutually correlated fashion, then this results in a marked reduction of the masking thresholds. This suggests that for complex signals, nearly all signal-dependent errors may prove to be more audible in some circumstances than suggested by traditional masking theory [5,6]. This is in addition to the effects of amplitude modulation discussed in this paper.

8. STEREO SPECTRAL MASKING

Cross-spectral ideas are particularly valuable for conceptualising the problems of directional masking of errors in stereo systems. At an informal level, it is known that if a first signal masks a second one in monophonic reproduction, then such masking can cease to work in stereo when the two signals are reproduced from two different perceived stereophonic directions. Indeed, it is precisely this "directional unmasking" that encouraged the standardisation of stereo rather than mono in domestic audio, since this greatly enhanced intelligibility of multi-source programme material is probably more significant than directional effect per se.

Informal listening tests on complex musical material suggest that directional unmasking reduces masking thresholds by at least 6 dB in a worst-case situation - i.e. lines remain audible at a level at least 6 dB lower in stereo than in mono - but that the degree of directional unmasking can be much higher in some situations - possibly up to the order of 25 dB.

Most conventional audio data compression systems are mono, and simply use separate mono coding of the two stereo channels. There is a clear risk here that, even if each channel separately monophonically masks coding/decoding errors, such errors may nevertheless be unmasked in stereo or binaural reproduction. This is because

the perceived direction of coding errors may well be different from that of the wanted signal. Spectral matrix theory can be used to design stereo coding/decoding systems that avoid such directional discrepancies.

The spectral matrix

$$\begin{bmatrix} S_{L,L}(F) & S_{L,R}(F) \\ S_{R,L}(F) & S_{R,R}(F) \end{bmatrix} \tag{16}$$

of respective left L and right R stereo channel signals is a positive matrix that can be written in the form

$$\tfrac{1}{2}(S_{L,L}(F) + S_{R,R}(F)) \begin{bmatrix} 1 + x(F) & y(F) + jz(F) \\ y(F) - jz(F) & 1 - x(F) \end{bmatrix}, \tag{17}$$

where positivity of the matrix (see equs. (4)) can be shown to be equivalent to

$$x(F)^2 + y(F)^2 + z(F)^2 \leqslant 1 \quad . \tag{18}$$

The coordinates

$$(x(F),\ y(F),\ z(F)) \tag{19}$$

thus lie within a sphere of unit radius. This sphere has been termed the energy sphere by the author - see ref. [22] for a full account of its properties and uses in connection with 2-channel stero and surround-sound systems. The position of a spectral matrix frequency component within this sphere thus describes the distribution of energy within the stereo channels at frequency F. The point (19) within the unit sphere may thus be taken to describe the stereo position of the signal at frequency F - although if quadrature phase aspects of localisation are ignored, one can suppress z(F) and describe localisation by the point (x(F), y(F)) inside the unit circle.

In a similar way, if the left and right error signals are respectively denoted by $\lambda(t)$ and $\rho(t)$, then the spectral matrix of the stereo error signal may be written in the form

$$\tfrac{1}{2}(S_{\lambda,\lambda}(F) + S_{\rho,\rho}(F)) \begin{bmatrix} 1+x_E(F) & y_E(F)+jz_E(F) \\ y_E(F)-jz_E(F) & 1-x_E(F) \end{bmatrix} \tag{20}$$

where the point

$$(x_E(F),\ y_E(F),\ z_E(F)) \tag{21}$$

within the unit-radius sphere describes the stereo position at frequency F of the stereo error signal.

In general, directional masking of a stereo error signal is likely to be best masked if the stereo position of the error signal, as defined by the energy sphere point (21), roughly coincides with the stereo position of the wanted signal as defined by the energy sphere point (19).

The stereo situation is in general much more complicated than the mono, since the total spectral matrix of the 4 signals L, R, λ, and ρ is a 4×4 positive complex matrix containing 16 real components at each frequency F, of which we have considered only 8 above. The other 8 components describe cross-correlations between the error signals and the wanted signals, and describe gain, phase and stereo positioning errors. For simplicity in this paper, we shall not attempt to describe this most general case (it is said that an art is a science with more than 7 variables - and the error signals contribute to 12 variables in the 4×4 spectral matrix!). Rather, we shall assume that a coding/decoding system is used that is designed to avoid these cross-correlations between error and wanted signals - e.g. by the use of quantisers with subtractive dither.

Even so, we are still left to consider the 8 spectral matrix components in equs. (17) and (20). We thus adopt a simplified directional masking strategy, which may not actually be optimal, by imposing the requirement that a coding/decoding system be designed to ensure that the energy sphere points of the error and wanted stereo signals be substantially the same at all frequencies. With such a restriction, we can be fairly sure that if monophonic masking works, then it will still work for the stereo signal. Such a strategy is very likely to give much better directional masking than independent monophonic coding of the two channels.

We now examine practical means of achieving this identity of energy sphere points (19) and (21). In a transform coding [9,11] or sub-band coding system, one can use pricipal-value quantisation of the stereo signal within each transform or frequency band component - i.e. one determines (from the eigenvectors of the correlation matrix or otherwise) the two orthogonal stereo directions that contain maximum and minimum signal energy in that frequency band (shown in terms of an XY oscilliscope display [23] in figure 10), and quantises these two components separately. If the two quantisers have the same number of bits but have levels adapted to each of the two components separately, then the error signal at each frequency or in each transform component will have the same stereo distribution as the wanted stereo signal, as required. This method is easiest to implement for the real components of stereo position ignoring $z(F)$ and $z_E(F)$, and complex principal- component quantisation of the stereo signal will require the use of 90^o phase shifts applied to the signals.

If instead of using the same number of bits to quantise the two principal stereo components in each frequency band, one were to quantise them using bit-allocation [9-11], one would get lower objective error energy in each band for a given bit rate (perhaps typically reducing the number of bits per stereo sample by $\frac{1}{2}$ bit), but one would pay the price of increased directional unmasking. It is possible that a compromise strategy allocating a number of bits to each principal stereo component intermediate between an equal number and mean-square-optimal bit allocation might work better subjectively than either.

Therefore we suggest that, as a starting point, stereo principal component quantisation in each frequency band or transform component be used for transform or sub-band coding of stereo signals, possibly using an identical number of bits for both components, and using subtractively dithered quantisers.

Rather than using two independent monophonic quantisers, it is possible to use a stereo vector quantiser using, for example, a regular hexagon quantisation region, and using a subtractive two-dimensional dither signal with a uniform probabilty distribution function over the hexagonal region. For signals using more than two channels, similar higher-dimensional subtractive dither quantisation strategies can be used, based, for example, on a rhombidodecahedral region in 3 dimensions or a regular 24-hedroid quantisation region in 4 dimensions [24]. Such vector quantisers improve the Shannon-efficiency of a joint quantisation of variables, reducing the 1.53 dB loss caused by separate quantisation even with ideal entropy coding. The signal-to-noise ratio gains, however, are not large and might not justify the extra complication - and significant gains might require, say, the use of the close-sphere -packing region in 23 dimensions.

Improved direction masking of stereo error signals can also be achieved using predictive coding (DPCM) systems, by replacing the monophonic prediction filter

$$P(z^{-1}) = a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n} \ , \qquad (22)$$

where z^{-1} represents a 1-sample time delay, by a stereo prediction matrix

$$\mathbf{P}(z^{-1}) = \begin{bmatrix} P_{LL}(z^{-1}) & P_{LR}(z^{-1}) \\ P_{RL}(z^{-1}) & P_{RR}(z^{-1}) \end{bmatrix} \ , \qquad (23)$$

where each of the 4 filters P_{LL} , P_{LR} , P_{RL} , and P_{RR} is of a form similar to (22). Such prediction filters can be derived by a natural extension of the monophonic case discussed by Makhoul [12] if one simply wished to minimise objective means square noise. In this case, the matrix filter $\mathbf{I} - \mathbf{P}$ would be a whitening filter for the stereo signal, and the stereo quantiser (whether a pair of mono subtractively dithered quantisers or a stereo vector quantiser) could use principal-component coding of the resulting prediction error signal, which in general will have more energy in some stereo directions than others, such as shown in fig. 10.

Without going into too much detail here, figures 11 and 12 show in schematic form the basic algorithms for stereo prediction and stereo noise shaping that can be used for stereo predictive coding. To avoid error/wanted signal cross-correlations, the stereo quantiser should be subtractively dithered as in the monophonic case, and the predictor matrix, noise-shaping matrix and stereo quantiser can all be made adaptive as in monophonic ADPCM. We postpone detailed consideration of stereo ADPCM to a future paper or papers - the details are far from trivial, although ultimately, the theory is a natural generalisation of the monophonic case.

Thus we have shown that it is possible to design coding systems that achieve good directional masking using any of the philosophies of transform coding, sub-band coding or predictive coding by the use of principal-component matrix methods to replace monophonic coding. It is evident from the above that such schemes are more complex than doubled-up mono coding. Nevertheless, stereo matrix coding schemes will in general be capable of better results than bimonophonic coding schemes in that they take automatic account of channel cross-correlations, and are likely to retain the masking of errors even if subject to subsequent rematrixing operations such as those involved in stereo width enhancement, surround-sound or Ambisonic decoding or multispeaker stereo reproduction. It is also obvious that the above schemes generalise to cases with more than 2 related channels (e.g. multichannel stereo or surround-sound systems) without any conceptual change.

9. PROBLEMS

We have almost reached the end of our tour of error-masking problems in high-quality audio coding systems. We have two remaining areas of unresolved difficulty besides that of the practical design of systems of coding and decoding that minimise unmasked errors.

First, our notion of cross-spectra involves averaging over all time, and to measure cross-spectral effects, we need to use a relatively short-term averaging, certainly not generally exceeding 50 ms. While we have been able to summarise the general theory of this in Appendix A, the result of taking finite time averages over a limited number of samples is that we are measuring a statistical quantity concerning signals from limited data - which inevitably involves fluctuations in the measurements for purely statistical reasons. These fluctuations are rather larger than might at first be imagined. Both theoretical analysis and measurement of practical cross-correlators used to derive control signals in surround-sound logic decoders have shown a measured correlation index that varies by $\pm \frac{1}{2}$ approximately for reasonable values of time constants.

Such large fluctuations in the measured cross-spectrum mask the ability to measure whether or not the cross-spectrum of an error with a wanted signal is substantially zero or not. Since transient cues are important in interpreting sounds, this makes direct measurement of cross-spectral errors on actual transient signals effectively impossible. The only way to get round this problem is to generate a large ensemble of transient signals having the same non-stationary statistics, and to average the measured cross-spectrum over all these measurements. This is a very lengthy and time-consuming process if one wishes to measure, say, transient gain errors of the order of 0.1 dB, since one has to average over around 2000 measurements to reduce fluctuations below the desired level.

Thus, although in principle it is possible to measure signal-dependent gain and phase errors as a function of frequency and time for arbitrary

encode/decode systems using the short-term cross-spectrum, in practice this is an involved, time-consuming and expensive process. It is thus probably better to adopt the attitude that, if one can design encoding/decoding systems free of cross-spectral errors (and we have shown that it is possible by using subtractively dithered quantisers), one should, in order to prevent the possibility of severe transient distortions.

Of course, this immediately raises the question that, if it is so difficult to measure cross-spectral error components of transients, then how does the ear manage to hear the effect? After all, no matter how intelligent the signal processing in the ears and brain, ultimately, all they can do is to analyse the incoming signal. Our conjecture is that the brain performs a similar averaging of data over multiple similar sound events as does the averaging method of making measurements suggested above. If this is the case, it implies that this type of information involves gathering data over a period of time in an acoustic environment, and so will not operate when there is a sudden change of acoustical environment such as encountered in blind AB testing. We hope to explore the consequences for the design of "objective" psychoacoustic testing of audio equipment elsewhere in another publication.

A second point arising from the cross-spectral model is that it is not evident why the ears and brain should only be sensitive to spectral matrices, which are quadratic in the input signal (see [25] and Appendix A), and not to higher order nonlinearities [25]. It is possible to extend the theory of spectral matrices to higher-order polyspectral tensors, as did the author in an unpublished but quite widely circulated paper [26]. The advantage of using subtractively dithered quantisers in coding/decoding systems is that, because they eliminate all nonlinear distortions in the quantisation process, and give noise that is un-cross-correlated with every nonlinear function of the wanted signal, they also avoid all higher order cross-polyspectral correlations between the error and the wanted signal. A full discussion of this point is beyond the scope of this paper - but it is worth noting that subtractive dither has advantages beyond those predicted by cross-spectral analysis.

10. CONCLUSIONS

This paper has explored a number of mechanisms whereby coding/decoding errors in audio data compression systems may have significant audible effects despite satisfying the requiements of naive spectral masking theory.

In particular, we have shown that all Shannon-efficient coding systems, and those that minimise a weighted rms error over a restricted class of coding strategies, automatically produce errors cross-correlated with the wanted signal, and we have given evidence that the resulting varying gain errors are likely to be audible at much lower levels than uncorrelated noiselike errors.

The failure of conventional spectral masking theory in this situation is explained by introducing the cross-spectra of signals and error. The resulting spectral matrix can be used to devise more widely applicable models for masking. Cross-spectral models are also useful for discussing the directional unmasking of errors in stereo systems.

These concepts have been used to suggest coding/decoding strategies that avoid these failures of masking. Specifically, we noted that replacing Max quantisers by subtractively dithered quantisers with entropy coding avoids error cross-spectral effects, and that the use of principal-component coding of stereo transform or sub-band components or of stereo matrix predictors can improve directional masking in the stereo case.

This paper contains many sketchy technical details that will require further publications to flesh out, but we believe that we have introduced here most of the basic tools required to improve subjective results by closing the gap between objective criteria and subjective perceptions a little.

The use of cross-spectral ideas in a general Shannon coding theory context, and in particular our Theorem about the need to depart from "ideal" Shannon coding if one is to avoid error/signal cross-correlations, is, we believe, an important conceptual clarification of the foundations of the Shannon theory, since it places emphasis on the statistics of the 2-vector signal whose components are the wanted signal and the error signal rather than merely their separate statistics. We hope in future to present an abstract formulation of these results with applications to other kinds of signals - indeed, we first discovered our theorem in connection with image data compression systems, where error/wanted signal cross-correlations also have a marked psychovisual effect.

In conclusion, we hope that designers of commercial audio data compression systems feel able to take on board the results of this paper, although it may require some additional work on their part to modify their systems to take account of the effects discussed here. Although the paper involves a lot of theoretical concepts, our aim has been to bridge the gap between the perceptions of the skilled practitioners of high quality audio and those of engineers trained in coding theory, by giving the latter theoretical tools that take on board many of the subjective perceptions of the practicing audio professional and audiophile.

To my ears, some of the low-rate prototype audio data compression systems that have been demonstrated clearly suffer from audible "pumping" and gain modulation of the wanted signal of the kind predicted by the work of this paper. If the communication gap between skilled audio professionals and engineers persists, there is a risk that audio data compression systems might come into widespread use that will be found to be seriously flawed by users despite meeting previously-defined engineering criteria such as the criteria of spectral masking theory. We hope that an improved understanding of the problems, and their solutions, will prevent this undesirable situation from occuring.

APPENDIX A. SHORT-TERM CROSS-SPECTRA

A short-term spectral analyser is a device whose outputs, one for each frequency F, are functions of time t that estimate the spectrum of a signal f(t) at frequency F around a time t by means of processing signal information covering a restricted time interval around time t. These outputs are required to be positive.

A basic short-term spectral analyser can be achieved as in figure A1 in which to each frequency F one has an associated linear filter with impulse response $\phi_F(t)$ (which may be real or complex-valued) centred around frequency F, followed by a squaring of the absolute value of the output of the filter to ensure a positive output

$$Q_{F,t}(f) = \left|\int \phi_F(t-\tau) f(\tau)\, d\tau\right|^2 . \qquad (A1)$$

If all the filters have identical bandpass shapes apart from centre frequency F, then for every frequency F :

$$\phi_F(t) = e^{2\pi jFt}\, \phi_0(t) . \qquad (A2)$$

The output of a basic analyser as in figure A1 and equ. A1 with complex-valued filter impulse responses $\phi_F(t)$ can be realised as the sum of the outputs of two basic analysers with real-valued filter impulse responses, since for all real-valued signals f(t) :

$$\left|\int \phi_F(t-\tau) f(\tau)\, d\tau\right|^2$$

$$= \left|\int [\mathrm{Re}\phi_F(t-\tau)] f(\tau) d\tau\right|^2 + \left|\int [\mathrm{Im}\phi_F(t-\tau)] f(\tau) d\tau\right|^2 , \qquad (A3)$$

where the integrals are evaluated over all time, i.e. from $-\infty$ to ∞.

Fluctuations in the output of such an analyser having frequencies of the form F_1+F_2 for frequencies F_1 and F_2 in the input signal f(t) can be avoided if $\phi_F(t)$ is an analytic signal, i.e. of the form $\psi_F(t) - jH\psi_F(t)$ where H is the Hilbert transform (see refs. [27-29]) or 90^o phase shift.

An alternative to the use of an analytic filter (i.e. one with an analytic-signal impulse response) to reduce high frequency ripples is to introduce low-pass filters at the outputs of the basic spectral analyser as shown in fig. A2; in order to preserve positivity, these filters must have a convolution kernel (impulse response) that is non-negative - e.g. a first-order low pass filter or running-average filter. Yet another stategy is to average the outputs of several basic spectral analysers having different filter characteristics, so as to average out output fluctuations. The problem here is simply that one has a wide variety of possible short-term spectral analysers, and to understand their behaviour, it is useful to have a unifying theory of all of them independent of the actual practical mode of realisation used in any particular case. This Appendix provides that unifying theory, based on the spectral theory of positive operators discussed in ref. [15] in abstract Hilbert space language, We then go on to modify the theory to include short-term cross-spectra.

A general spectral analyser is a quadratic functional $Q_{F,t}(f)$ acting on signals $f(t)$, defined for each frequency F and time t, given by an equation of the form :

$$Q_{F,t}(f) = \iint k_F(t-\tau_1,\ t-\tau_2) f(\tau_1) f^*(\tau_2)\ d\tau_1 d\tau_2 \ , \qquad \text{(A4)}$$

where the function $k_F(t_1,\ t_2)$ of 2 variables satisfies the hermitian property :

$$k_F(t_2,\ t_1) = k(t_1,\ t_2)^* \ . \qquad \text{(A5)}$$

For such a quadratic [25] analyser, we can write:

$$Q_{F,t}(f) = \int (A_{F,t} f)(\tau_2) f^*(\tau_2) d\tau_2 \ , \qquad \text{(A6)}$$

where $A_{F,t}$ is the linear operator acting on signals defined by

$$(A_{F,t} f)(\tau_2) = \int k_F(t-\tau_1, t-\tau_2) f(\tau_1) d\tau_1 \ . \qquad \text{(A7)}$$

the function $k_F(t-\tau_1,\ t-\tau_2)$ of τ_1 and τ_2 is termed the kernel of the linear operator $A_{F,t}$ defined by equ. (A7). In order to keep the output of the analyser finite for finite energy signals, we in general require that the operator $A_{F,t}$ be bounded (defined in ref. [15] whether or not the kernel k_F can be defined as an explicit function of two variables. An important special case of practical interest is when k_F is square-integrable, i.e. when

$$\int |k_F(t_1,\ t_2)|^2\ dt_1 dt_2 < \infty \quad ; \qquad \text{(A8)}$$

in this case, the linear operator $A_{F,t}$ is said to be a Hermitian Hilbert-Schmidt operator. In this case, we have the following important and non-trivial theorem (see [15]), which is a version of the spectral theorem (warning: the term spectrum here has nothing to do with frequency spectrum; it refers to what is termed the "spectrum" of a linear operator - i.e. the set of its eigenvalues, although the frequency spectrum is actually the "spectrum" in this sense of the Fourier transformation operator).

Spectral Theorem

Let $A_{F,t}$ be a linear operator defined by equ. (A7), where the kernel k_F is hermitian (A5) and square-integrable (A8), i.e. where $A_{F,t}$ is a Hermitian Hilbert-Schmidt operator. Then there exist a unique sequence of real numbers

$$\lambda_{F,1} \geqslant \lambda_{F,2} \geqslant \lambda_{F,3} \geqslant \ \ldots \geqslant \lambda_{F,n} \geqslant \ \ldots . \qquad \text{(A9)}$$

and a sequence of square-integrable functions $\phi_{F,n}$ $(n = 1,2,3,\ldots)$ of a real variable, satisfying :

$$\text{(i)} \quad \sum_{i=1}^{\infty} (\lambda_{F,i})^2 < \infty \quad . \qquad \text{(A10)}$$

(ii) the functions $\phi_{F,n}$ for a fixed F form an orthonormal basis for the square integrable fuctions, i.e.

$$\int \phi_{F,i}(\tau)[\phi_{F,k}(\tau)]^* d\tau = \delta_{ik} \ , \qquad \text{(A11)}$$

where $\delta_{ik} = 0$ if $i \neq k$ and $\delta_{ik} = 1$ if $i = k$, and where every square-integrable function can be expanded uniquely as a convergent linear combination of the $\phi_{F,n}$'s,

and

$$\text{(iii)} \quad k_F(t_1, t_2) = \sum_{i=1}^{\infty} \lambda_{F,i}\, \phi_{F,i}(t_1)[\phi_{F,i}(t_2)]^* \ . \qquad \text{(A12)}$$

Moeover, for all (square-integrable) signals f,

$$(A_{F,t}f)(\tau_2) = \sum_{i=1}^{\infty} \lambda_{F,i} \int \phi_{F,i}(t-\tau_1)\phi^*_{F,i}(t-\tau_2)f(\tau_1)d\tau_1 \ , \qquad \text{(A13)}$$

and

$$Q_{F,t}f = \sum_{i=1}^{\infty} \lambda_{F,i} \left| \int \phi_{F,i}(t-\tau)f(\tau)d\tau \right|^2 . \qquad \text{(A14)}$$

Moreover, the $\lambda_{F,i}$'s are the eigenvalues of the linear operator $A_{F,t}$ associated with the eigenvector $\phi_{F,i}(t-\tau)$, i.e.

$$(A_{F,t}\phi_{F,i}(t-.))(\tau) = \lambda_{F,i}\phi_{F,i}(t-\tau) \ . \qquad \text{(A15)}$$

The most important aspect of the above spectral theorem is that it asserts that <u>all</u> quadratic spectral analysers (satisfying the purely technical Hilbert-Schmidt condition (A8)) giving a real analyser output (this is the content of the hermition condition (A5)) can be expressed as linear combinations (A14) of the outputs of basic spectral analysers as described earlier in equ. (A1). Moreover, knowing the eigenvalues $\lambda_{F,i}$ and eigenvectors $\phi_{F,i}$ of a specific quadratic spectral analyser gives us a powerful theoretical tool for understanding its properties.

Although the case where some eigenvalues are negative is actually of practical interest in some measurement applications, we are mostly interested in the case that the spectral analyser has non-negative output, i.e. when the analyser is <u>positive</u> , i.e. when for all signals f(t)

$$Q_{F,t}f \geqslant 0 \ . \qquad \text{(A16)}$$

<u>Theorem A2</u>

Let $Q_{F,t}$ be a quadratic spectral analyser with associated linear operator $A_{F,t}$ as in equ. (A6) and kernel $k_F(t_1, t_2)$ as in equ. (A4) such that $A_{F,t}$ is Hermitian and Hilbert-Schmidt.

Then $Q_{F,t}$ is a positive analyser if and only all of the eigenvalues $\lambda_{F,i}$ of the Spectral Theorem above are non-negative.

This result is a direct consequence of the theorem found in [15] that the spectrum of a positive linear operator is positive.

It is thus possible to specify a positive short-term spectral analyser either by specifying for each frequency F a kernel $k_F(t_1,t_2)$ of a positive operator or a sequence of non-negative eigenvalues $\lambda_{F,i}$ satisfying equs. (A9) and (A10) and a sequence of orthonormal filter impulse responses $\phi_{F,i}(t)$, which for the larger eigenvalues should be designed only to let through frequencies close to F. If one requires that all analyser outputs have the same shape of "bandpass" characteristic appart from a frequency translation, then analogously to equ. (A2), one should put

$$k_F(t_1,t_2) = e^{2\pi jF(t_1-t_2)}k_0(t_1,t_2) \ , \qquad \text{(A17)}$$

which ensures that all the eigenvalues become independent of frequency F and that all the eigenvectors satisfy equ. (A2). This frequency-translation invariant case is of particular interest

since it only requires the specification of the eigenvalues λ_i and eigenvectors ϕ_i at frequency 0 to specify them at <u>all</u> frequencies via equs. (A2) or (A17).

We now relate this work to the Wigner Distributon method of simultaneous time-frequency analysis of signals (also termed the time-varying spectrum by Mark [30] who rediscovered it independently). The Wigner distribution was first published in connection with quantum statistical mechanics by Wigner [31] in 1932, and some of its more abstract theory was developed in an abstract mathematical language by Pool [32], whose methods we have borrowed here in more concrete notations. The audio applications of the Wigner distribution were first noted by de Bruijn [33] (in Dutch - which may explain the spate of Audio publications using it emerging from Holland), and it is available on some commercial signal analysis packages such as MLSSA - however, it does not present information in a form accessible to the eye, having a large amount of confusing visual clutter. It does have a useful role, however, in understanding short-term spectral analysers as we shall now see.

The <u>Wigner Distribution</u> of a signal f(t) is defined by

$$W_{f,f}(F,t) = \int f(t+\tfrac{1}{2}\tau) f^*(t-\tfrac{1}{2}\tau) e^{-2\pi j F\tau} d\tau \qquad \text{(A18)}$$

which can be written in a more abstract notation similar to that of Pool [32]

$$W_{f,f}(F,t) = \mathfrak{F}_1 \rho (f \otimes f^*)\,(F,t)\ , \qquad \text{(A19)}$$

where we define

$$\rho k(t_1,t_2) = k(t_2+\tfrac{1}{2}t_1,\ t_2-\tfrac{1}{2}t_1) \quad , \qquad \text{(A20)}$$

$$\mathfrak{F}_1 k(F,t_2) = \int k(t_1,t_2) e^{-2\pi j F t_1}\, dt_1\ , \qquad \text{(A21)}$$

and

$$(f \otimes g)(t_1,t_2) = f(t_1)\, g(t_2) \qquad \text{(A22)}$$

for all functions f of 1 variable and k of 2 variables.

Now, both $\mathfrak{F}_1$ and ρ are "unitary" (see [15]), i.e. they preserve square-integrals, orthogonality and inner products of functions in 2 variables. By using this unitary property, we can write a quadratic (Hilbert Schmidt) short-term spectral analyser (A4) in the form

$$\begin{aligned} Q_{F,t}f &= \iint k_F(t-\tau_1,\ t-\tau_2)(f \otimes f^*)(\tau_1,\tau_2) d\tau_1 d\tau_2 \\ &= \iint W_{k_F}(-F',\ t-\tau)\, W_{f,f}(F',\tau)\, dF' d\tau\ , \end{aligned} \qquad \text{(A23)}$$

where we define

$$\begin{aligned} W_{k_F}(F',t') &= (\mathfrak{F}_1 \rho k_F)(F',t') \\ &= \int k_F(t'+\tfrac{1}{2}\tau',\ t'-\tfrac{1}{2}\tau') e^{-2\pi j F\tau'}\, d\tau'\ . \end{aligned} \qquad \text{(A24)}$$

This means that the output of a short-term spectral analyser can be obtained by convoluting (in time and frequency) the Wigner distribution (A18) and (A19) of the signal with the Wigner distribution (A24) of the kernel k_F of the spectral analyser, and evaluating the result along the time axis (i.e. F'=0) of the Wigner distribution plane.

In the case (A17) where the properties of the analyser are frequency-invariant, we can rewrite (A23) in the convenient form

$$Q_{F,t}f = \iint W_k(F-F',t-t')\, W_{f,f}(F',t')\, dF'dt'$$
$$= W_k * W_{f,f}\,(F,t)\ , \qquad \text{(A25)}$$

where $k(t_1,t_2) = k_0(t_1,t_2)$ and $*$ indicates convolution - in this case in two variables.

Thus the output of an arbitrary frequency-invariant short-term spectral analyser is simply obtained by convoluting the Wigner Distribution of the signal f by the Wigner Distribution of the analyser's DC kernel $k(t_1,t_2)$. The actual computation of (A25) in signal analysis software can most easily be done by computing the Fourier transform (in both variables) of both Wigner distributions, i.e.

$$\hat{W}_k(t'',F'') = (\mathfrak{F}_2^* \rho\, k)(t'',F'') \qquad \text{(A26)}$$

and

$$\hat{W}_{f,f}(t'',F'') = (\mathfrak{F}_2^* \rho(f \otimes f^*))(t'',F'')\ , \qquad \text{(A27)}$$

where $\mathfrak{F}_2^*$ indicates the inverse Fourier transform in the 2nd variable, multiplying the two together, and then taking the Fourier transform back again to the (F,t) plane. This is not an excessively complex computation to do using the Fast Fourier Transform. Moreover, the form of W_k or of its Fourier transform gives an easily visualised picture of how much "smearing" in the time and frequency domains an analyser gives. The larger the area of the time-frequency domain over which the smearing occurs, the poorer the resolution of the analyser, but the lower the level of statistical fluctuations in the outputs. A 2-dimensional Gaussian convolution kernel for W_k may be a good choice for many applications, with an appropriate choice of standard deviations σ_F and σ_T along the time and frequency axes - a special case of this is when one has a basic analyser whose filters have Gaussian impulse responses, when one gets what is termed the Husimi Transform discovered by Husimi in 1940 [34].

For an arbitrary (Hilbert-Schmidt) frequency-invariant analyser, one can use the spectral theorem along with (A25) to write

$$Q_{F,t}f = \sum_{i=1}^{\infty} \lambda_i\, W_{\phi_i,\phi_i}(F,t) * W_{f,f}(F,t)\ , \qquad \text{(A28)}$$

which expresses the output, when the analyser is positive, as a positive linear sum of the smoothing of the Wigner distribution of the signal f by a kernel that is the Wigner distribution of a DC convolution kernel ϕ_i.

The spectral theorem expansion can be used to estimate the fluctuations in the outputs of a positive short-term spectral analyser. One can define such an analyser to have "unit gain" if

$$\sum_{i=1}^{\infty} \lambda_{F,i} = 1\ . \qquad \text{(A29)}$$

In this case, one is averaging the outputs, with weights $\lambda_{F,i}$, over orthogonal (and hence, for white noise input signals, statistically independent) components. Thus if one has signals that within

a region of the time frequency plane (i.e. for a duration and within a given frequency range) have a locally white spectrum, , the standard deviation of the fluctuations caused by the quadratic spectral analyser are a factor

$$\sum_{i=1}^{\infty} (\lambda_{F,i})^2 \qquad \text{(A30)}$$

times smaller than for a basic analyser of unit gain, i.e. the rms fluctuation is proportional to

$$\left\{\sum_{i=1}^{\infty} (\lambda_{F,i})^2\right\}^{\frac{1}{2}} \quad . \qquad \text{(A31)}$$

This result shows, for example, that an analyser that averages over 4 orthonormal vectors ϕ_i will have half the amplitude of fluctuations in its output than a basic analyser.

This averaging, of course generally reduces the available time--frequency resolution. If the time smear is Δt and the frequency smear is ΔF, then by the uncertainty principle

$$\Delta F \times \Delta t \geq \tfrac{1}{2} \qquad \text{(A32)}$$

for a basic analyser , whereas for the general case satisfying (A29):

$$\Delta F \times \Delta t \geq \frac{1}{2 \sum_{i=1}^{\infty} (\lambda_{F,i})^2} \qquad \text{(A33)}$$

for other analysers. We shall not attempt to make these results precise here, but this gives a general idea of the tradeoffs between fluctuations and time/frequency resolution.

For signals whose bandwidths near a frequency F are narrower than ΔF, then the fluctuations will be larger than indicated above, because fewer of the eigenvectors $\phi_{F,i}$ are excited by the signal and the averaging is thus over fewer components.

We now extend the above methods to short-term cross-spectra. The above results all extend to cross-spectral analysis simply by replacing f* throughout by a second signal g*, where g(t) is a second square-integrable signal. By way of example, we have the short-term cross-spectrum

$$Q_{F,t}(f,g) = \iint k_F(t-\tau_1,\ t-\tau_2) f(\tau_1) g^*(\tau_2) d\tau_1 d\tau_2 \qquad \text{(A34)}$$

$$= \iint (A_{F,t} f)(\tau_1) g^*(\tau_2) d\tau_1 d\tau_2 \qquad \text{(A35)}$$

$$= \int W_{k_F}(-F', t-\tau)\ W_{f,g}(F', \tau)\ dF' d\tau\ , \qquad \text{(A36)}$$

where

$$W_{f,g}(F,t) = \int f(t+\tfrac{1}{2}\tau) g^*(t-\tfrac{1}{2}\tau) e^{-2\pi j F\tau}\, d\tau \qquad \text{(A35)}$$

$$= W_{f \otimes g^*}(F,t)\ ,$$

and equ. (A14) becomes

$$Q_{F,t}(f,g) = \sum_{i=1}^{\infty} \lambda_{F,i}[\int \phi_{F,i}(t-\tau)f(\tau)d\tau][\int \phi_{F,i}(t-\tau)g(\tau)d\tau]^* . \quad (A36)$$

All the earlier results can, with this modification, be applied to describing not just the short-term spectrum of a single signal, but also the short-term cross-spectrum $Q_{F,t}(f,g)$ of two signals f and g. It is then quite easy to prove:

Theorem A3
Let $Q_{F,t}$ be a positive short-term spectral analyser. Then the short--term spectral matrix of two signals f(t) and g(t) :

$$\begin{bmatrix} Q_{F,t}(f,f) & Q_{F,t}(f,g) \\ Q_{F,t}(g,f) & Q_{F,t}(g,g) \end{bmatrix} \quad (A37)$$

is a positive matrix for all frequencies F and time t, i.e.

$$Q_{F,t}(f,f) = Q_{F,t}f \geqslant 0 \quad (A38a)$$

$$Q_{F,t}(g,g) = Q_{F,t}g \geqslant 0 \quad (A38b)$$

$$Q_{F,t}(f,g)^* = Q_{F,t}(g,f) \quad (A38c)$$

and

$$|Q_{F,t}(f,g)|^2 \leqslant (Q_{F,t}f)(Q_{F,t}g) . \quad (A38d)$$

This theorem allows one to handle the short-term spectral matrix in very much the same way as the ordinary spectral matrix in the main body of this paper.

While in principle this has given us all the basic tools of short--term cross-spectral analysis, if we wish to use the imaginary part of the short-term cross-spectrum in a meaningful way, it is necessary to restrict the kind of analysers we consider to those that handle 90^o-phase shifted components of signals well. This is best done by replacing real signals in the above by analytic signals [27] of the form $\mathbf{f} = f - jHf$, where H is the Hilbert transform or 90^o phase shift operator. We define a short-term cross-spectral analyser $\mathbf{Q}_{F,t}$ to be analytic if and only if there is a second short-term spectral analyser $Q_{F,t}$ such that for all real signals f and g :

$$\mathbf{Q}_{F,t}(f,g) = Q_{F,t}(f-jHf, g-jHg) . \quad (A39)$$

For an analytic short-term cross-spectral analyser $\mathbf{Q}_{F,t}$, one then has

$$\begin{aligned} \mathbf{Q}_{F,t}(Hf,g) &= Q_{F,t}(H(f-jHf), g-jHg) = Q_{F,t}(j(f-jHf), g-jhg) \\ &= j\mathbf{Q}_{F,t}(f,g) , \end{aligned} \quad (A40)$$

so that , as one would require, a Hilbert transform (90^o phase shift) acting on the signal f has the effect of multiplying the short-term cross-spectrum by $j = \sqrt{-1}$ when the analyser is analytic. It is not difficult to prove that the eigenvectors $\boldsymbol{\phi}_{F,i}$ of an analytic spectral analyser are also analytic signals, but in general they will not be equal to $\phi_{F,i} - jH\phi_{F,i}$ where $\phi_{F,i}$ are the eigenvectors of the non-analytic analyser $Q_{F,t}$ of equ. (A39).

The general analytic positive quadratic analyser can be expessed as before by the spectral theorem and theorem A2 as a positive sum of basic analysers based on orthonormal analytic eigenvectors - the

proof of this and other theorems about analytic analysers lies in the replacement of the Hilbert space of square-integrable complex--valued signals by the Hilbert space of analytic (i.e. positive frequency) signals.

We therefore describe one implementation of a basic analytic cross-spectral analyser by way of example, since all others are positive linear combinations of such basic analysers. One can also use the Wigner distribution method of computing the output of analytic cross-spectral analysers as described earlier of the spectral analyser case.

Figure A3 shows the implementation of a single frequency-band output of an analytic cross-spectral basic analyser, using a Hilbert transform (90^o phase shift) network block: the basic concepts used here were described, in a more general setting by the author in refs. [28] and [29]. The implementation shown in fig. A3 requires 4 real multiplications of pairs of signals per analyser output - this representing one multiplication of a pair of complex signals. We shall not go into a complete analysis of fig. A3's operation here, leaving it to the reader to verify that the output is merely the product of the result of the analytic signals f-jHf and $(g-jHg)^* = g+jHg$ passing through the filters ϕ_F and then being multiplied together as complex signals.

Using fig. A3, one can implement basic spectral and cross-spectral analytic analysers directly, and more complex analytic analysers by taking positive linear combinations of the outputs of several such analysers. The Wigner distribution method of computation, using the FFT, is more appropriate when the non-analytic filters $\phi_{F,i}$ are frequency translations of a DC filter characteristic, using convolution with the Wigner distribution of the kernel of the analytic analyser, for reasons of computational efficiency.

We omit further details here, since there is detailed material for many possible future publications. However, our main aim has been to show that a comprehensive theory of short-term cross-spectral analysis is possible maintaining the main features (e.g. positive spectral matrix responding appropriately to the effect of 90^o phase shifts) of the long-term cross-spectral analysis used in the main body of the paper. We have also given basic information allowing implementation and computation of short-term spectra and cross-spectra, and also a basic analysis of the fluctuations in their outputs and the compromises involved (in time/frequency resolution) when attempting to reduce these fluctuations. This all makes it possible to give a systematic analysis of cross-spectral errors encountered when actual systems respond to transient or non--stationary signals and to measure such effects - a necessary prerequisite if the theory of this paper is to be useful.

However, we make no claim that such measurements are <u>easy</u> : section 9 of the main paper discusses why such measurements are likely to be time-consuming - basically because one needs many repeated measurements to reduce the statistical effects of output fluctuations.

REFERENCES

[1] M.A. Gerzon, "Masking of Coding/Decoding Errors in Audio Data Compression Systems", Proc. Inst. Acoustics, Vol.12, Part 8 pp. 175-182 (1990 Nov.)
[2] ed. N.S. Jayant, Waveform Quantization and Coding, IEEE Press New York 1976
[3] G. Slot, Audio Quality, Philips Technical Library, Eindhoven 1962
[4] M.A. Gerzon, "Blumlein Stereo Microphone Technique", J. Audio Eng. Soc., Vol.24, No.1, pp.36-37 (1976, Jan/Feb)
[5] R. Plomp, "Aspects of Tone Sensation", Academic Press, London 1976, pp. 1-167

[6] B.C.J. Moore, "An Introduction to the Psychology of Hearing Third Edition", Academic Press, London 1989

[7] A.M. Yaglom, "An Introduction to the Theory of Stationary Random Functions", Prentice-Hall 1962 and Dover 1973, chapter 3.
[8] L.D. Davisson, "Rate Distortion Theory and Application", Proc. IEEE, vol.60, pp.800-808 (1972 July) - also in ref. [2].
[9] J.J.Y. Huang & P.M. Schultheiss, "Block Quantization of Correlated Gaussian Random Variables", IEEE Trans. Commun. Syst., vol. CS-11, pp.289-296 (1963 Sept.) - also in ref [2].
[10] A. Habibi & R.S. Hershel, "A Unified Representation of Differential Pulse-Code Modulation (DPCM) and Transform Coding Systems", IEEE Trans. Commun., vol. COM-22, pp.692-696 (1974 May) - also in ref. [2]
[11] R.J. Clarke, "Transform Coding of Images", Academic Press, 1985
[12] J. Makhoul, "Linear Prediction : A Tutorial Review", Proc. IEEE vol.63, pp.561-580 (1975 April)
[13] M.A. Gerzon & P.G. Craven, "Optimal Noise Shaping and Dither of Digital Signals", Presented at the 87th Convention of the Audio Engineering Society, New York, 1989 Oct., Preprint 2822
[14] M.A. Gerzon, " The Gentle Art of Digital Squashing", Studio Sound, vol.32 no. 5, pp.68-76 (1990 May)
[15] P.R. Halmos, "An Introduction to Hilbert Space and the Theory of Spectral Multiplicity", Chelsea, New York, 1957,
[16] J. Max, "*Quantizing for Minimum Distortion*", *IRE Trans. Inform. Theory*, vol. IT-6, pp.7-12 (1960 Mar.) - also in ref. [2]
[17] M.D.Paez & T.H. Glisson, "Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems", IEEE Trans, Commun., vol. COM-20, pp.225-230 (1972 April) - also in ref. [2]
[18] R.C. Wood, "On Optimum Quantization", IEEE Trans. Inform Theory, vol. IT-15, pp. 248-252 (1969 Mar.) - also in ref. [2]
[19] L.G. Roberts, "Picture Coding Using Pseudo-Random Noise", IRE Trans. Inform. Theory, vol. IT-8, pp. 145-154 (1962 Feb.) - also in ref. [2]
[20] L. Schuchman, "Dither Signals and Their Effect on Quantization Noise", IEEE Trans. Commun. Technol., vol. COM-12, pp. 162-165 (1964 Dec) - also in ref. [2]
[21] B.C.J. Moore, "Co-modulation Masking Release: Spectro-Temporal Pattern Analysis in Hearing", Brit. J. Audiology, vol. 24, pp. 131-137 (1990)

[22] M.A. Gerzon, "A Geometric Model for Two-Channel Four-Speaker Matrix Stereo Systems", J. Audio Eng. Soc., vol.23, no.2, pp. 89-109 (1975 March)
[23] B.B. Bauer & G.W. Sioles, "Stereophonic Patterns", J. Audio Eng. Soc., vol.8, no.2 pp.126-129 (1960 April)
[24] H.M. Coxeter, "Regular Polytopes", Macmillan, London 197?
[25] M.A. Gerzon, "Mathematics and Sound Perception", J. Audio Eng. Soc., vol.26, nos. 1/2, pp.46-50 (1978 Jan/Feb)
[26] M.A. Gerzon, "Nonlinear Models for Auditory Perception" Unpublished but privately circulated, 1974
[27] J. Ville, "Theorie et Applications de la Notion de Signal Analytique", Cables et Transmission, vol. 2, pp.61-74 (1948)
[28] M.A. Gerzon, "Decomposition of Nonlinear Operators into 'Harmonic' Components, With Applications to Audio Signal Processing", Electonics Letters, vol.12, pp.23-24 (1976)
[29] M.A. Gerzon, "Effect of Gain Control and Modulation on Harmonic Nonlinear Operators, With Applications to Audio Signal Processing", Electronics Letters, vol.13,pp.118-120 (1977)
[30] W.D. Mark, "Spectral Analysis of the Convolution and Filtering of Non-Stationary Stochastic Processes", J. Sound Vib., vol. 11, pp.19-63 (1970)
[31] E. Wigner, "On the Quantum-Correction for Thermodynamic Equilibrium", Phys. Rev., vol.40 pp.749-759 (1932)
[32] J.C.T. Pool, "Mathematical Aspects of the Weyl Correspondence", J, Math, Phys., vol.7, pp.66-75 (1966)
[33] N.G De Bruijn, "Wigner-Distributie als Muzikaal Notenschrift bij de Fourier-Analyse van Signalen", Kon. Ned. Akad. Wet., vol. 76, pp. 19-23 (1967)
[34] K. Husimi, Proc. Phys. Math. Soc. Japan, vol (3)22 pp.264-314 (1940)

ACKNOWLEDGEMENTS

I would like to thank Dr. J.P. Wilson of the Dept. of Communication and Neuroscience at the University of Keele for drawing my attention to the important ref. [21], and also to thank Drs. Peter Craven and Geoffrey Barton for their encouragement and discussions over several years on digital signal processing and coding systems.

no. of levels	Entropy (bits)	Gain dB	uncorrelated noise dB	gain-compensated noise dB
1	0.000	$-\infty$	$-\infty$	$+\infty$
2	1.000	-3.92	-6.36	-2.44
3	1.536	-1.83	-8.12	-6.29
4	1.904	-1.10	-9.80	-8.70
5	2.183	-0.745	-11.22	-10.48
6	2.409	-0.543	-12.44	-11.90
7	2.598	-0.417	-13.50	-13.08
8	2.761	-0.331	-14.43	-14.10
9	2.904	-0.271	-15.27	-14.99
10	3.032	-0.224	-16.02	-15.79
11	3.148	-0.192	-16.70	-16.51
12	3.253	-0.165	-17.33	-17.16
13	3.350	-0.144	-17.91	-17.77
14	3.440	-0.127	-18.45	-18.32
15	3.524	-0.113	-18.95	-18.84
16	3.602	-0.101	-19.43	-19.33
20	3.876	-0.069	-21.08	-21.01
25	4.146	-0.047	-22.74	-22.69
32	4.449	-0.030	-24.59	-24.56
36	4.594	-0.025	-25.47	-25.45

Table 1. Performance of n-level equilevel Max quantiser for Gaussian signals, based on data of J. Max [16] and equations (9) and (10), and showing the gain of the wanted-signal component of the coded signal, the level of the un-cross-correlated component of the error signal, and the level of the error for a gain-compensated quantiser.

no. of levels	Entropy (bits)	Gain dB	uncorrelated noise dB	gain-compensated noise dB
1	0.000	$-\infty$	$-\infty$	$+\infty$
2	1.000	-6.02	-6.02	0.00
4	1.751	-1.90	-8.02	-6.12
8	2.392	-0.646	-11.77	-11.12
16	3.077	-0.223	-16.06	-15.84
32	3.776	-0.076	-20.64	-20.57
4*	1.742	-1.69	-8.38	-6.69
8*	2.610	-0.490	-12.86	-12.37

Table 2. As for table 1, but for optimum equilevel quantiser for a Laplacian pdf statistics, based on the data of Paez and Glisson [17], except for * which is data for "optimum" non-uniform quantisers for Laplacian pdf signals.

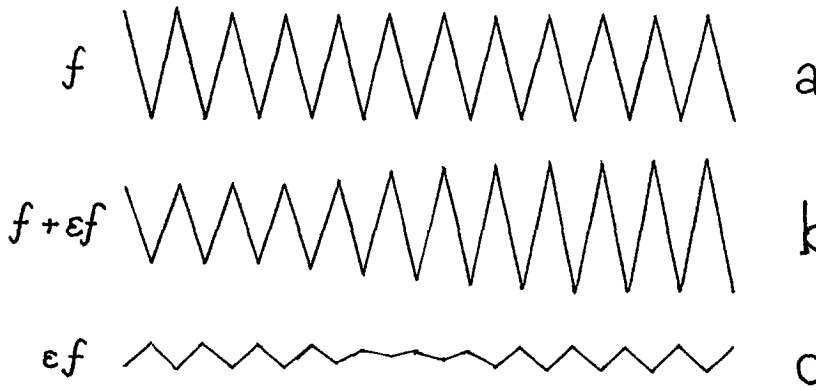

Figure 1. Effect of small amplitude modulation (b) of wanted signal (a), showing the error signal (c). The gain change is exaggerated for illustrative clarity.

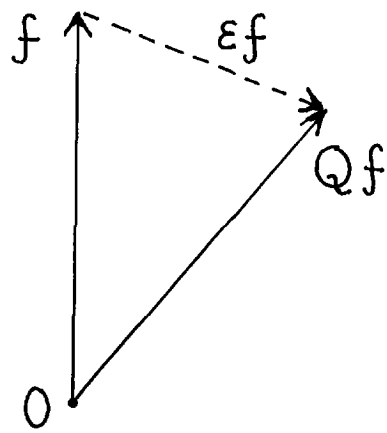

Figure 2. Triangle consisting of vectors representing the wanted signal f, the coded/decoded signal Qf and the error signal εf.

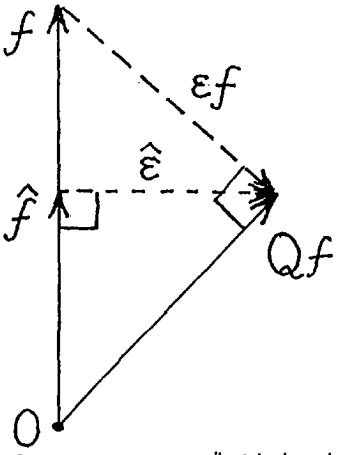

Figure 3. The case when r.m.s. error $\|\varepsilon f\|$ is minimised, showing the orthogonal projection $\hat{f}$ of Qf onto f.

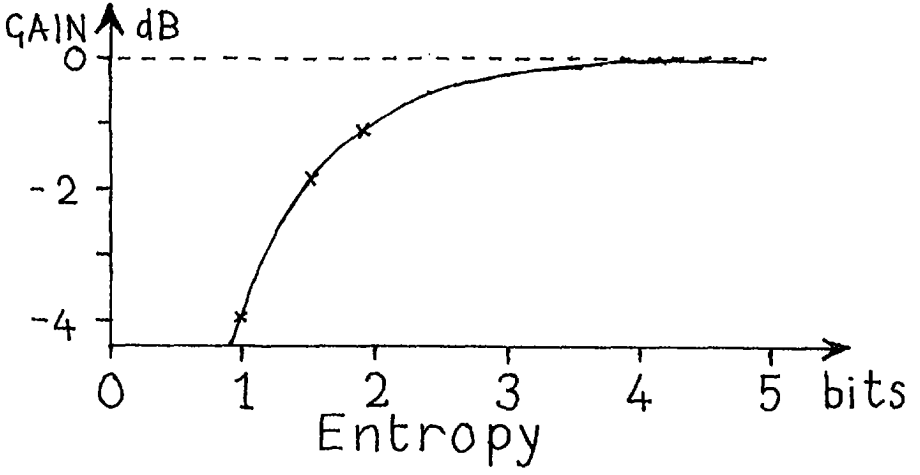

Figure 4. Gain in dB of the correlated component of the coded/decoded signal via equilevel Max quantisers for signals with Gaussian statistics, plotted against the Entropy-coded bit rate as shown intable 1.

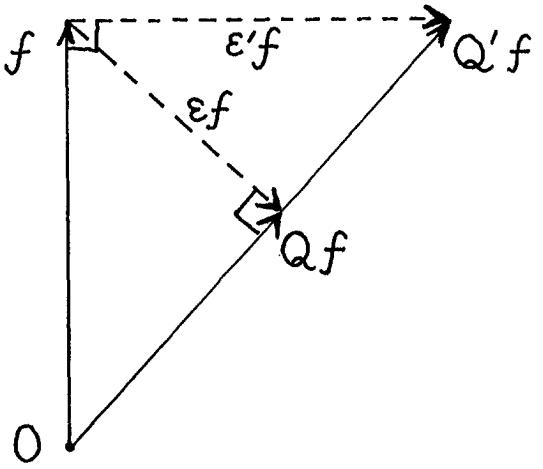

Figure 5. Gain-compensated Max quantiser signals Q'f.

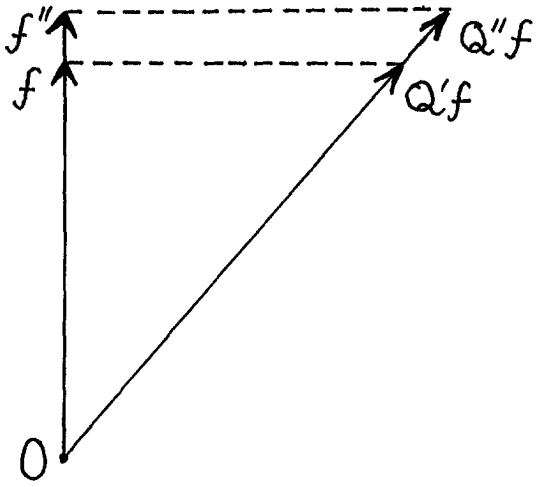

Figure 6. Gain modulation effects caused by erroneous gain-compensated quantisation Q"f caused by signal/quantiser gain mismatch.

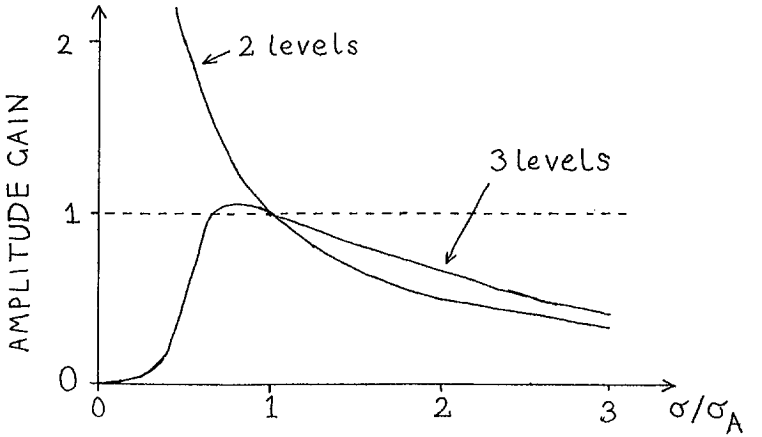

Figure 7. General form of the gain errors of 2-level and 3-level gain-compensated Max quantisers optimised for r.m.s. level σ when actual r.m.s. signal level is σ_A. The 3-level graph is only a rough sketch not based on precise computations.

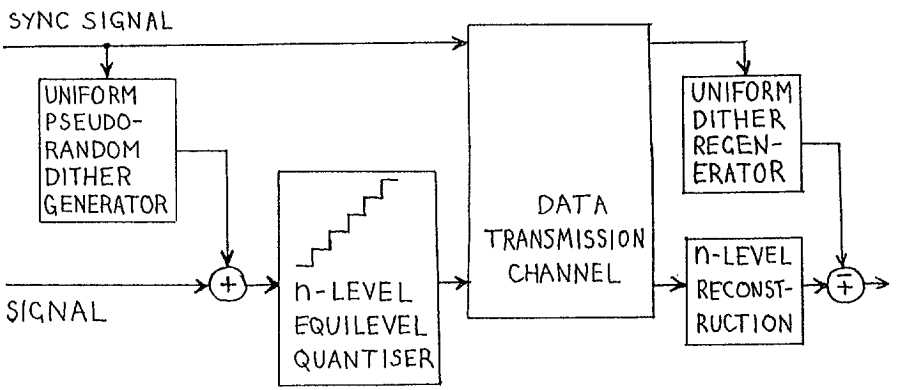

Figure 8. Subtractively dithered quantiser (after [19] and [20]).

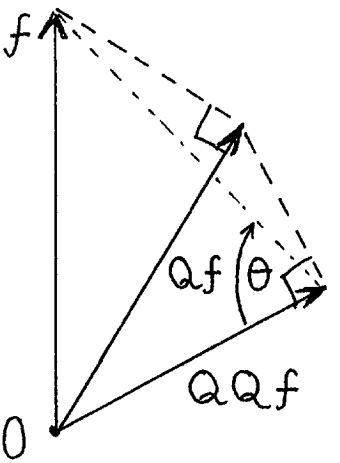

Figure 9. The Max quantisation QQf of a Max quantisation Qf of a signal f is not a Max quantisation of f, since in general, the angle $\theta \neq 90^{\circ}$. The figure should be viewed as a vector diagram in 3 dimensions.

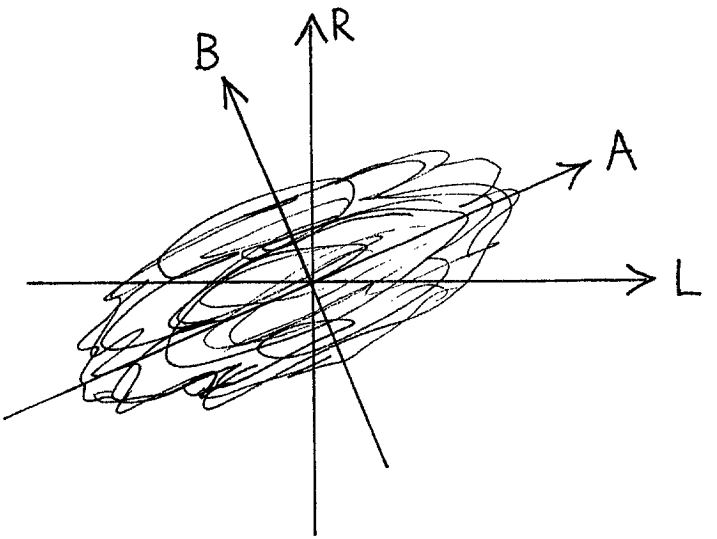

Figure 10. Principal-component axes A and B for quantising a stereo signal component whose "XY" oscilliscope display has the form shown.

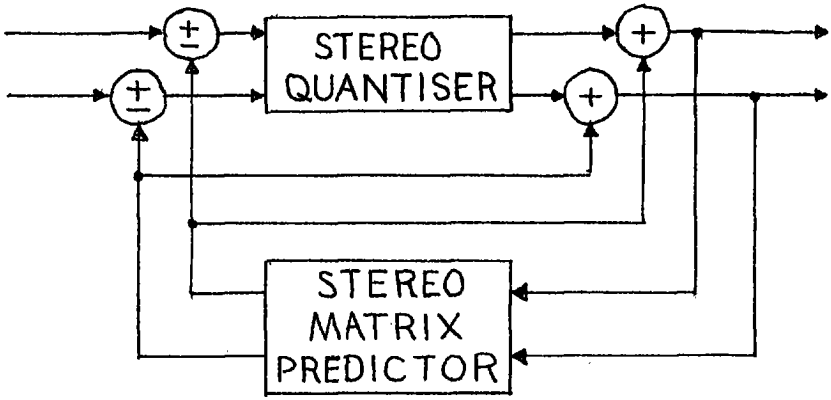

Figure 11. Schematic of a stereo prediction system, with predictor as in equ. (23). The stereo quantiser can be equpped with stereo noise-shaping as in figure 12.

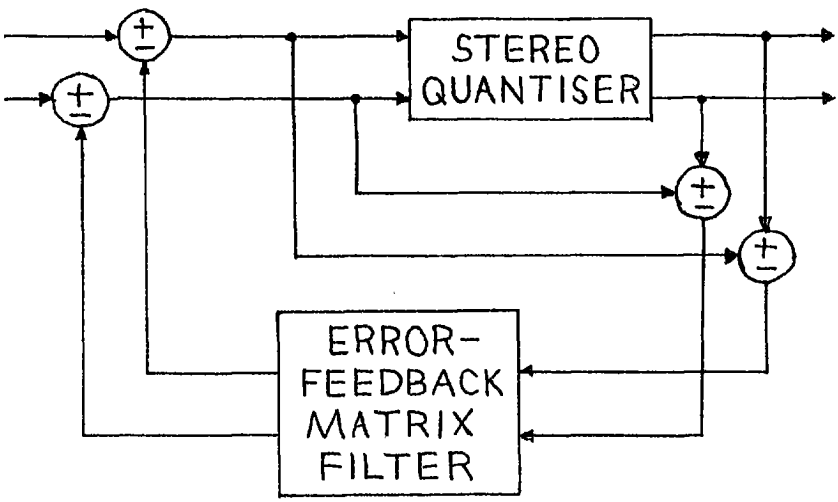

Figure 12. Schematic of stereo noise shaping around a stereo quantiser to modify the spectral matrix content of the stereo quantisation noise. This is a stereo version of ref. [13]. The stereo quantiser may be subtractively dithered as in figure 8.

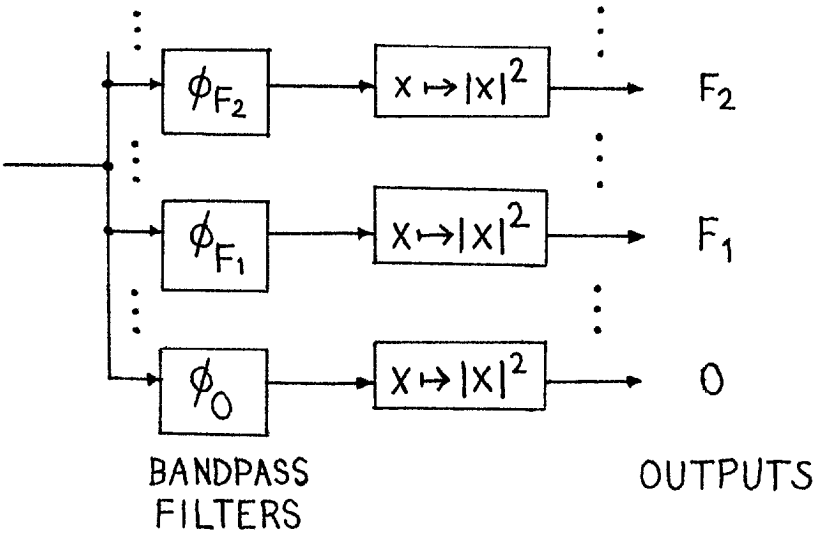

Figure A1. Basic real-time spectral analyser using multiple bandpass filters ϕ_F, one centred at each frequency F, followed by a square-of-absolute value operation to ensure a positive output.

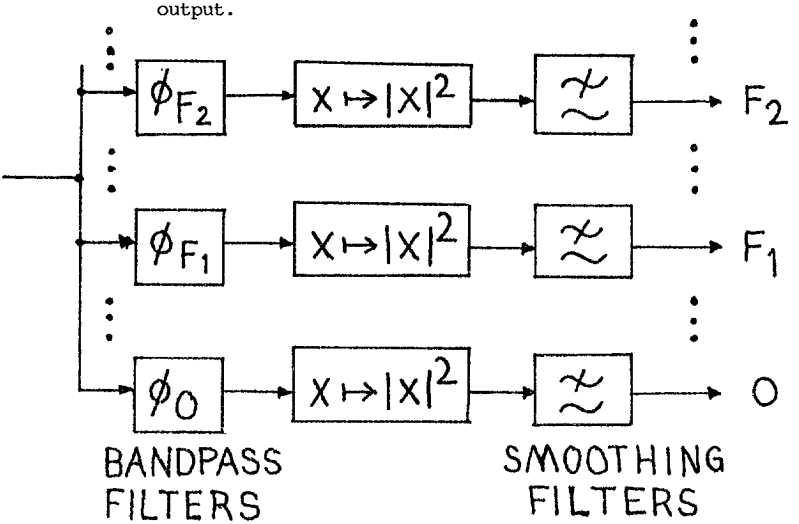

Figure A2. The basic filter-bank spectral analyser of fig. A1 with additional output low-pass filters to reduce output fluctuations.

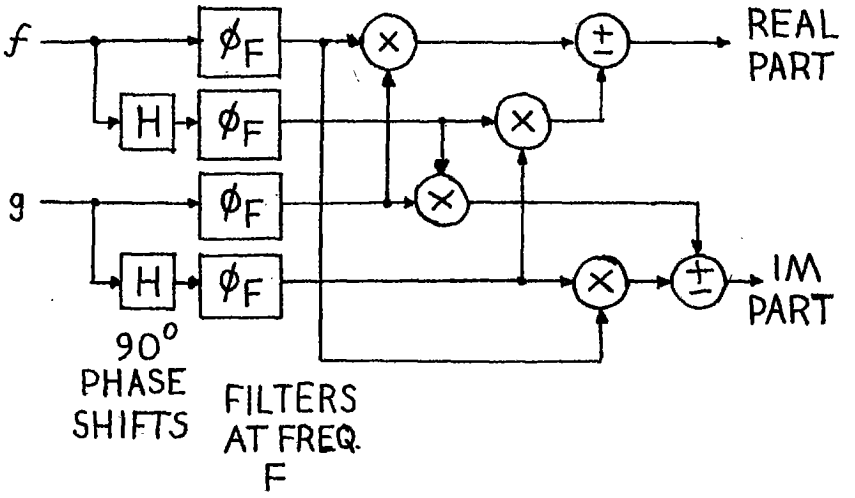

Figure A3. Derivation of the real and Imaginary components of the cross-spectral output at frequency F of a basic analytic spectral analyser based on bandpass filters ϕ_F.