Proceedings of the Institute of Acoustics

PSYCHOACOUSTIC MODELS FOR EVALUATING ERRORS IN AUDIO SYSTEMS

J. R. Stuart

Technical Director, Boothroyd Stuart Ltd, Huntingdon, England

1. INTRODUCTION

This paper outlines an approach to examining distortion and noise (considered as errors) in a way that allows a more accurate relationship to be established between these errors and the human hearing capability of detecting or discriminating them. As a first step a method is evolved to relate various sounds to a scale of pure-tone equivalence, using an improved model to calculate the *masked subjective loudness*, detectability and *aural patterns* of a stimulus based on recently established data of the auditory filter.

A fundamental assumption in the work is that if an error can be shown to have a positive partially masked loudness, then it will be a candidate for changing the sound.

Of necessity the bulk of this paper is given to introducing the psychoacoustic ideas and phenomena on which the model for loudness calculation is based.

2. SYSTEM CHARACTERISATION

Speech and music are extremely complex signals. In order to reduce the number of variables, both audio engineers and psychoacousticians have evolved methodologies which investigate different aspects of system behaviour in response to relatively simple stimuli – a small number of tones or noises.

In psychoacoustics this approach has led to considerable understanding of many aspects of human auditory behaviour including loudness, pitch, discrimination, threshold, masking etc. The audio engineer uses similar stimuli to investigate frequency response, distortion, self-noise and so on for equipment. The prime difficulty for audio engineers is then to interpret their results in useful ways.

The most common tools for audio-system measurement are tone generators and rms-calibrated mean-reading frequency-selective meters. Digital techniques have added to that armory, particularly spectrum analysis by FFT. To the extent that the human hearing system has been calibrated for pure tones, it is possible to make reasonable comparisons between, for example, parameters like frequency response and audible frequency range.

The process of predicting the sound from objective measurements becomes significantly more complex when more than one tone is present in the output signal, or when that signal combines different types of sound, e.g. noise bands and/or tones.

3. ABSOLUTE LEVELS

To estimate the audibility of a stimulus, we have to establish its fundamental acoustic parameters. Although this may seem obvious, it is not in fact common practise for equipment measurements to be traced forward to sound-pressure level (spl) at the listener. Such a connection is vital since the human hearing system is quite non-linear in many respects.

Many audio components whose distortion we wish to characterise, will precede the system volume control. In these situations we have to determine the effective gain of the system. The author has considerable experience of listening levels through using calibrated digital active loudspeakers, and has proposed previously [41] that three general situations are of interest. For domestic listening to

CD material a typical gain setting is that which would produce 105dB spl @ 1m per speaker if the CD were encoded with a full level sinewave (0dBFS). Demonstration or review situations tend to use somewhat higher levels which can be close to the gain settings that permit 'natural' volume levels at the listening position. This second gain setting is equivalent to 112dB spl @ 1m per speaker for 0dBFS. This average setting has been determined from a survey of a large number of classical-music CDs. The peak level of course is almost never achieved – the system may not even be capable of delivering the 0dBFS level at *any* frequency. The third gain setting of interest is found in professional monitoring where gains up to 122dB for 0dBFS are used.

These levels, although independently derived, are similar to those proposed by Fielder [10]. This issue is addressed further from a practical viewpoint in [43]. All examples in this paper treat levels as final spl.

4. HUMAN HEARING

It is helpful to consider human hearing as a two-level hierarchy; perception and cognition. Here, perception is considered to be the low-level peripheral process where coding and relatively simple decisions are made relatively independent of context or meaning. By contrast cognition is a high-level process which invokes many signal and context-dependent decisions. It is a late process in which external noises are grouped into perceived external objects [44].

In the main, psychoacoustic experiments are designed to minimise the cognitive aspects of the tests. By creating simple stimuli the perceptual limits and performance can be determined. This section reviews aspects of hearing perception that are relevant to constructing the model.

4.1 Absolute threshold

Fig. 1 shows the the average minimum-audible field (MAF) in spl for frontally incident tones for audiometrically normal subjects – the so-called threshold of hearing.

Since the data is an average, 50% of the population can detect sounds below this level – not necessarily at all frequencies. It is not uncommon in young listeners to see thresholds at some frequencies 10dB below this curve, and thresholds 20dB lower are not unknown [28]. In any individual, the microstructure of the threshold can vary by up to 10dB between frequencies only 10% apart [48]. See also [22], [43] and [6].

4.2 Masking

Masking is the process by which the threshold for a sound is raised by the presence of another sound. The degree of masking depends upon the intensity of the masker and its proximity in frequency to the maskee. The masker effectively redefines the hearing threshold. Fig. 2 shows an example of the masked threshold resulting from a 90dB spl narrow band noise centred on 1kHz.

It is quite helpful to see that the absolute threshold is also in fact a masked threshold. It is thought that internal physiological and neural noise place a limit on the detectability of quiet sounds – particularly at low frequencies [36], [9].

4.3 Loudness

A tone can be described by its objective acoustic parameters of spectrum; frequency and intensity. The subjective correlates are pitch and loudness. Loudness is the most basic impression since all audible sounds have loudness, whereas they may not all have a pitch, chroma, timbre or position.

Loudness varies with the stimulus according to intensity, frequency, temporal variation and the loudness of other simultaneous sounds – including room noise. Loudness also varies with the listener. Variables include temporary threshold-shift (TTS), monaural/binaural listening, attention and of course general condition of the listeners hearing.

ISO131 defines two physical magnitudes for sound, sound pressure level and sound intensity level, as follows:

$$spl = 20\log(p/p_0) \quad dB \tag{1}$$
$$L_i = 10\log(I/I_0) \quad dB \tag{2}$$

where p is the rms sound pressure at a point in a field of intensity I, and p_0 is the reference pressure of 20uPa equivalent to an intensity I_0 of $1pW/m^2$.

Excitation level L_e, sometimes called sensation level. L_e is the stimulus level compared to the relevant hearing threshold expressed in dB, and is defined in (3).

$$L_e = 20\log(p/p_t) \quad \text{where } p_t \text{ is the spl at the relevant threshold} \tag{3}$$

There are two widely used subjective magnitudes for sound, and both are required in the model.

- Phon. The loudness level of any sound, L_n, is n phon when it is judged by normal listeners to be equally loud to a plane-progressive frontal-incidence pure tone at 1kHz whose spl is n dB.
- Sone. The sone scale is a numerical designation of the loudness of a sound that is proportional to its subjective magnitude N estimated by normal observers. By standard agreement one sone (N=1) is the loudness of a sound whose loudness level L_n is 40 phon. By definition an inaudible sound has a loudness of 0 sones and so the sone scale is discontinuous at threshold, there being no basis for judging inaudible stimuli. From ISO226 we see that 0 sones occurs on average below 4.2 phons.

For an audio engineer the phon, being expressed in dB and easily traced back to a 1kHz tone is intuitively more useful. However in the analysis of a sound we have first to consider its subjective magnitude, the sone. Obviously a scale based on subjective magnitude will exhibit variability – not least due to experimental circumstances. Stevens [37 38] was largely responsible for proposing the standard loudness function given in ISO131 [20] which for a 1kHz tone is shown in Fig 3. An important property of hearing (shared by other sensory modalities) is that the subjective response follows a power-law relationship with the stimulus – equal stimulus ratios producing equal sensation ratios (above threshold) [6]. In the case of loudness there is a good fit to a simple power-law behaviour above 40 phon. The standard loudness function – shown dotted in Fig 3 – is defined only above 40 phon and is expressed by:

$$N = 0.01P^{0.6} \quad \text{where P is the sound pressure in uPa} \tag{4}$$

There has been considerable debate about the correct value for the exponent, values up to 1 being proposed [47] [39]. These differences, giving a range between 6 and 10dB for a doubling of loudness are on the face of it disturbing. The author's view is that for this investigation, absolute adherence to a sone-scale is not as important as internal consistency in the model nor as important as producing accurate results in phons.

4.3.1 Loudness variation near threshold

It is a characteristic of sensory systems that the power-law behaviour does not hold in the region of threshold. Typically, just above threshold, sensation will initially rise more sharply – almost linearly with intensity. In Fig 3 the solid line shows Zwislocki's data for loudness. He approximated it by the formula:

$$N = 0.01(P - P_0)^{0.6} \quad \text{where } P_0 \text{ is the threshold sound pressure in uPa.} \tag{5}$$

The initial steepness of the loudness function is very similar to the characteristic observed for loudness in the presence of masking. Fig 4 (from [40]) shows the loudness function for a 1kHz tone in the presence of various levels of white noise. Although the form is the same, loudness grows somewhat more rapidly out of masking than out of the absolute threshold. Although in many respects the absolute threshold can be modelled as one of a general set of maskers, it should be borne in mind that neural input at these very low spls is very small, particularly in the middle frequencies, and a successful loudness model should take some account of this. Fig 5 illustrates the unmasked difference limen for intensity (DLI) for various pure tones as a function of sensation level L_e.

4.3.2 Loudness variation with frequency

Fig 6 gives a familiar set of equal-loudness contours based on ISO 226 and extended with data from [1] and [31]. Each contour describes the variation of the phon with spl and frequency. This figure illustrates several important properties of hearing:

- Loudness is frequency dependent
- The frequency dependence seems to have two components; one level dependent and one level-independent.
- Growth of loudness with spl is frequency-dependent, or more probably dependent on both frequency and threshold.

In producing an improved model for loudness the author had to incorporate these features which are now considered in a little more depth.

The loudness for frontally incident plane-waves (as shown in Fig. 6), combines the acoustic response of the head/pinna/meatus with the ear system considered from eardrum onwards. Fig 7 plots the difference between data for 100 phon equal-loudness measured using headphones (adapted from Fletcher-Munson) and the 100phon curve of Fig 6. Some general features of the equal-loudness contours may be ascribed to the acoustic path to the eardrum. This assertion is strengthened by the uniformity of the dynamic range above 500Hz. Fig 8 plots the dynamic range interpreted as the excitation or sensation level for 100 phons.

A model that estimates loudness ideally needs to separate the performance of the cochlea and subsequent neural system from all prior acoustic and mechanical responses. It is thought that at least between 300Hz and 12kHz, the general shape of the 100 phon curve basically describes the external acoustic response while the deviation below 300Hz reflects stiffness loss in the eardrum/ossicular chain/round window system.

These considerations led to the adoption of a frequency-dependent but level independent response shape shown in Fig 9 which is in fact the difference between the 100phon equal-loudness contour and 100dB spl, for use as a frequency-response correction.

By identifying the probable mechanical/acoustical response element of the loudness function we can deduce more clearly the power input to the cochlea. It is in the cochlea and subsequent neural processing that the core features of frequency selectivity and neural responses need to be closely modelled.

Fig 10 represents the data of Fig 6 modified by the function of Fig 9 showing in effect the variation of loudness with normalised cochlear input. Having identified the level-independent frequency-response component of the loudness function, we now consider the level-dependency. Fig 11 presents this data for several frequencies showing the growth of loudness with L_c which is defined as the normalised cochlear input expressed in dB. Fig 11 clearly shows the 'masked' form of behaviour at low frequencies.

At low-frequencies the rapid growth of loudness between threshold and what appears to be an efficiency approaching that of mid-range processing in the region of 100 phon, reflects both physiological masking and neural inefficiency. Since below 2kHz individual neurones produce a broadly phase-locked response, it is plausible to suggest a 1/f characteristic of neural efficiency based purely on reduced data. Whilst understanding more of the mechanism is helpful in evolving a model which reflects the real situation, in this case there is no difference between masking and loss of efficiency due to reduced data. Therefore we can successfully separate the loudness function into a frequency-dependent but level-independent component as shown and a threshold dependent function. The advantage of a threshold-dependent function is that it is much more useful in a practical tool, where in fact masked thresholds are an important variable.

4.3.3 Loudness variation for noises

When a second, different tone is added to an existing tone stimulus, the loudness will usually increase. The increase depends on the frequency and level difference between the two tones. Unfortunately the change will depend on whether the cognitive response combines the two tones into a single percept. See [43]. For example a two-tone presentation results in three general loudness perceptions namely:

- the overall loudness of the combination,
- a separate loudness for each tone (each now partially masked by the other).

These three loudnesses will generally be different from the loudness of each tone taken separately and the overall will rarely be equal to their sum.

These factors give rise to two considerations in predicting loudness, namely:

- the variation of loudness with spectrum content,
- mutual masking and loudness under masking.

If a tone stimulus is exchanged for a narrow band stimulus of the same intensity – for example by presenting narrow-band noise or several tones close together in frequency, then for very narrow combination widths, the loudness of the complex will be judged equal to the single tone. As the bandwidth of the complex is increased, then its loudness at mid phon levels, tends to remain constant up to a certain bandwidth, above which the loudness increases with bandwidth. The bandwidth at which loudness starts to increase was called the 'critical bandwidth for loudness summation' by Zwicker [49]. Fig 12 shows data for loudness growth with bandwidth from Scharf [33].

For stimuli whose bandwidths are are significantly wider than a critical band, e.g. pink or white noise, then loudness variation with intensity also varies as more or fewer frequencies become involved due to the shape of the absolute threshold [23]. Fig 13 shows the growth of loudness for a 20kHz bandwidth white noise compared with a pure-tone of 1kHz at the same overall spl with data taken from [34]. Note the following features.

- Near threshold the wideband noise is less loud because the energy is divided over more than one critical band and therefore each band is stimulated even closer to threshold.
- Above threshold the growth of loudness for the noise is steeper as more bands come together well out of threshold.
- At high stimulus levels the loudness of the noise grows less quickly than the pure tone, a more appropriate exponent being 0.5.
- These functions cross twice.

4.4 Loudness, masking and frequency selectivity

The summation of loudness with complex stimuli has for some time been explained by the critical-band model [45]. The frequency selectivity of the auditory system is clearly not infinite, there being (among other considerations) a balance to be made between frequency and temporal resolution in the cochlea. A pure tone stimulus does not give rise to the excitation of a single neurone on the basilar membrane; rather it creates a spread of excitation along the membrane – a spread in mapped frequency space – in a way that reflects the selectivity of the cochlea/neural complex. It was first suggested by Fletcher [11], that:

- The masked threshold produced by a masker reflected this spread of excitation. See Fig 2. The underlying assumption is that detectability of a signal relates directly to the neural excitation it creates relative to that of the masker, and the DLI.
- The loudness of a sound was related to the total excitation pattern on the basilar membrane.

Obviously we have no sensible means of determining directly the excitation pattern of the basilar membrane itself, but since loudness reflects the *total* neural excitation, this can be determined by psychophysical experiments.

4.5 Frequency selectivity and the auditory filter

A widely used measure of selectivity for determining masking and loudness has been the critical bandwidth proposed by Zwicker et al [49]. His assertion was that the frequency selectivity was demonstrated by the loudness-summation-with-bandwidth effect described earlier. His critical-bandwidth function is plotted in Fig 14. This measure of frequency selectivity has given rise to considerable controversy, particularly for frequencies below 1kHz, since it deviates from selectivity estimates such as those of Greenwood [13 14] which are derived from neural mapping. The low frequency selectivity suggested by the critical-band function also seems implausible from a musical viewpoint.

Patterson and Moore [30] have shown that the frequency selectivity is in fact as shown in Fig 14. By an entirely different route they have reached a measure of selectivity based on detection in masking that separates selectivity and neural efficiency. Patterson and Moore called this selective element the auditory filter and defined for it an equivalent rectangular noise bandwidth, the ERB. The author is particularly attracted to this measure of frequency selectivity because it is based on modelling at the physical rather than psychophysical level, and because it agrees more closely with rational measures such as those of Greenwood. Within extremely close limits the ERB represents a distance of 0.9mm along the basilar membrane [29]. See Fig 14.

Recently estimates for the width of the auditory filter have extended the measure from below 100Hz to 10kHz [24, 35]. This model uses an equation for the equivalent bandwidth ERB in terms of frequency f in Hz that was fitted by Moore to Greenwood's data:

$$ERB = 19.5 + 0.117897f \tag{6}$$

The integral of the reciprocal of this expression gives a scale for frequency based on the ERB which effectively maps frequency as it lays out on the basilar-membrane. The resulting function, the E-scale for frequency is necessary when total excitation is calculated and is analogous to, but different from, Zwicker's z or bark scale of frequency.

$$E = 18.31\log(0.006046f + 1) \tag{7}$$

A question arises in this work about the highest frequency auditory filter. Clearly at 20kHz, or wherever the highest audible tone is, there cannot exist a symmetrical auditory filter since there is no sensitivity to higher frequencies. Examination of the threshold and of the filter shape at high frequencies leads to the conclusion that the highest auditory filter is between 14kHz and 15kHz, certainly from the point-of-view of masking or loudness summation. This end-of-cochlea hypothesis is confirmed in [3].

Choosing the auditory filter model for calculating loudness of error signals is clearly more appropriate since it reflects a greater selectivity – and therefore gives more sensitivity to the measure.

All models of frequency selectivity propose a continuous set of filters, that is, within a possible limit of approximately 11000 innervations, a filter can be identified at any point on the basilar membrane, and the detection of a signal involves a process that selects the filter that produces the best signal/noise ratio. To calculate the spread of excitation resulting from an acoustic stimulus we need to know the response of the auditory filter, the masking pattern being related to the combination of the outputs of all auditory filters. According to a review of Patterson and Moore [29], the filter is a bandpass with the following properties:

- the response shape is level dependent,
- the filter operates down to the absolute threshold,
- the shape is asymmetric for frequencies above and below the centre,
- the *normalised* shape is similar at all frequencies when plotted on an E scale of frequency,
- the shape can be described mathematically in a rounded-exponential form.

Fig 15 illustrates the response of the auditory filter at various levels for a 1kHz centre frequency.

5. MODELS FOR EXCITATION AND LOUDNESS

The two most used methods for calculating loudness are defined in ISO532 [21]. Method A is based on a pragmatic model by Stevens while method B is based on work by Zwicker. In many situations Stevens model will produce a more exact result, however Zwicker's model, being lower-level in the sense of attempting to model the perceptual process more closely is theoretically more pleasing.

Glasberg in [26] presented a FORTRAN program which allowed an excitation pattern to be produced from a series of input spectral lines. This original program was improved to show how excitation patterns could be extended to specific loudness patterns at mid frequencies [27].

The model derived here is an attempt to move forward in accuracy, scope and range from the principles of method B which are best described in [49] while moving from a critical-band to an auditory filter selectivity criterion.

6. SPECIFIC LOUDNESS

If loudness is related to the total spread of excitation on the basilar membrane then this loudness will be the sum of the neural response at all points along the membrane. This response at each point, an inferred parameter called specific loudness by Stevens, is defined as the loudness-per-unit distance (distance representing the mapping of frequency along the basilar membrane). This hypothesis has been confirmed by Cacace and Margolis [5]. The best mapping of distance is given by the E scale of frequency. A suitable expression to calculate specific loudness as a function of excitation was derived by Zwicker and Scharf. In a generalised form it is as follows:

$$S = D.B(f,th).C(X_e) \qquad (8)$$

where S is the specific loudness (then in sones/bark, in this case in sones/ERB). D is a constant chosen ultimately to give a result of 1 sone at 40 phon. B(f,th) is a factor which is either frequency and/or threshold dependent and which describes for each frequency the rate at which loudness grows. $C(X_e)$ describes the growth of loudness with the input signal excitation X_e (in response to input intensity I). This model follows Zwicker's example in making B(f,th) a function of threshold – this makes masked calculations more straightforward. For an input of L_e there is a corresponding power excitation of X_e. There is also a threshold excitation X_t, and an absolute excitation X_0 equivalent to $1pW/m^2$. The author has further modified B(f,th) taking account of the level-independent frequency function of Fig 9, X_d being the modified threshold excitation. Equation 8 then takes the form

$$S = D.(X_d/X_0)^k.((bX_e/X_t + (b-1))^k - 1) \qquad (9)$$

where b is used here as a frequency-dependent factor describing neural efficiency initially set to 0.5, and k is the exponent relating excitation to specific loudness.

Earlier in this paper it was indicated that exponents for pressure were 0.6 for a pure tone at 1kHz and 0.5 for white noise. Experiments by Zwicker to find k used uniformly exciting noise (albeit based on a critical band criterion) found an exponent of 0.46 to fit the data. According to basic tests by the author the loudness difference between uniformly exciting noise based on critical-band or auditory-filter criteria is very small. In the final model, fitting gave a power exponent of 0.205 and a value for D of 0.11.

Expressing B(f,th) as $(X_d/X_0)^k$ may seem counter-intuitive since this part of the equation should relate to the available dynamic range as shown in Fig 9. In fact the two are fairly closely related, and the expression used here is more convenient when masked thresholds are used and gives improved accuracy.

7. MODEL FOR EXCITATION, MASKING DETECTION AND LOUDNESS

The objective was to create a series of simple computer programs which could be used to make several estimates from input data [42]. The input data include:

- signal data defining the input sound,
- a masked threshold curve which may have been calculated previously by the model,
- parameters to define the acuity of hearing.

Output from the model is in data files for graphing or in single number form, and include the following 'modes' controlled from the command-line.

- The normalised form of the masked excitation pattern produced on the basilar membrane, as a function of frequency in Hz.
- An estimate of the simultaneous masking threshold resulting as a function of frequency in Hz in the cochlea
- An estimate of the simultaneous masking threshold resulting as a function of frequency in Hz expressed in 'external' spl.
- The specific loudness pattern as a function of E
- The specific loudness pattern as a function of Hz
- A numeric result for masked loudness in sones and phons
- An estimate of the detectability of a sub-threshold combination using independent-threshold statistics

Input data is taken from an ASCII list file whose default format is that of the data files produced by the *Audio Precision* measuring systems. Output data is produced in the same format so that a result can be post-processed and either graphed or analysed later. These formats are also readable by popular spreadsheet programs such as *Microsoft Excel* or *Lotus 123*.

In the input data, tones are represented as line frequencies with a signal magnitude while noise is represented by a series of tones of appropriate level with sufficiently close spacing. For many measurements a spacing of 50Hz is more than adequate, at lower frequencies a smaller spacing down to even 1Hz is preferable. In addition to constructed data, a common source is the output of an FFT program and in these cases an offset and/or adjustment has to be made for the data to relate the windowing of the FFT and the bin width to 1Hz for noise signals. The method chosen is to pass gain offset as a command-line parameter and for mixed noise/spectral line data these files are first preprocessed to correct the level differences caused by bin width.

Several programs have been written, some optimised for batch file use; other more complex versions handle masking data files and are able to produce 'negative' loudness estimates – see [43]. Essential differences from earlier published models are:

- use of auditory filter rather than critical band concepts,
- extension of auditory filter analysis from 20Hz to 15kHz,
- ability to accept masking profile and give masked rather than absolute-threshold-based results,
- specific-loudness calculations based on inferred round-window input using the weighting function of Fig 9 for much improved accuracy,
- level-dependence of auditory filter handled using iteration,
- Oriented to audio engineers and audio measurements.

These changes have resulted in a model which when tested against a wide variety of established data gives very good results. More details of the models are given in [42].

7.1 Model operation

Fig 16 shows a block diagram of the major data processes. The essential method is to pass each data point in turn through an array of auditory filters (default E of 0.6 to 35 0.1ERB apart – 25Hz to 13.5kHz). For each filter the shape is level-dependent, and the appropriate level is calculated for this equalised stimulus after ERB averaging. In this case the stimulus is weighted by the external–cochlea acoustic/mechanical response. The appropriate level is that which produces the highest output from that filter. The model starts with a 30phon excitation and is modified through a number of converging iterations using the input data.

To develop the excitation pattern, the input data is first converted to excitation level L_e using internal or variable masked threshold data. Then for each filter, the excitation pattern is calculated by passing each adjusted input data point through it and calculating the total intensity of the response. The resulting curve shows the spread of excitation along the basilar membrane.

From this excitation pattern, equations describing the DLI are used to produce masked thresholds, both for the cochlea and in terms of externally applied acoustic signals.

The excitation pattern is then transformed into a pattern of specific loudness, using a relation in the form of equation (9). When required the total loudness is calculated by summing the specific loudness on the E-scale to give a result in sones. This is then converted to phons using the formula (10) described below.

As an indication of the models behaviour in response to simple stimuli Fig 17 shows the loudness results for a 1kHz tone. The model is within 0.5 phon over the range of the ISO131 standard. Below 40phon the behaviour is similar to the data of Zwislocki shown in Fig 4. These results are more accurate than those of Moore and Glasberg in [27] primarily because of the iterative approach to defining filter shape, the modified form of B(f,th) in equation (8) and the use of masked-threshold functions. Internal consistency was given much more priority than conformance to precise sone numbers in the region up to 20 phon, so the output function of this model at 1kHz in Fig 17 was fitted using a formula given in [27] as a prototype, the result relating phons L_n to sones N – which is correct within 0.5dB above 6 phon was:

$$L_n = 10\log(8000N^{3.35} + 3200N^{1.8} + 9) \qquad (10)$$

7.2 Verification

The model was tested on pure tones at intensities between 0 and 110dB spl. Up to 100dB the loudness result was within 1 phon of the standard between 100Hz and 10kHz, and within 2 phon otherwise.

Figs 18 19 and 20 show the excitation, specific loudness and masked threshold patterns for a 1kHz tone presented at several levels. The characteristic upward spread of masking with increasing level is well illustrated. Figs 21 and 22 show how the excitation and specific loudness patterns develop for a signal centred on 1kHz, of constant 60dBspl but increasing bandwidth – modelled as a series of closely spaced tones. The specific loudness result shows how the critical-band phenomenon of loudness summation occurs. As the complex widens the specific loudness pattern widens, but the peak reduces in a way that leads to a nearly constant area, as the overall phon results show.

An further test of the success of the model is shown in Fig 23 where the loudness of complexes is investigated by width at different intensities. It can be seen that the model predicts the observed results shown in Fig 12.

The model has also been extensively verified in determining its ability to reproduce masked loudness results like those of Fig 4, and in the growth of loudness for noise as in Fig 13.

8. SOME RESULTS

Fig 24 shows the excitation patterns for a flat triangular-probability-distribution 16-bit dither signal calculated at the three replay levels proposed in section 3. The data comes from an FFT measurement made on the *Audio Precision* dual domain analyser. The model clearly gives a new perspective on the way this noise may be perceived, for the three cases the full-band noise spl would be 12dB, 19dB and 29dB; the calculated loudnesses are quite different – as we would expect – at 0, 14 and 36 phons.(0, .08 and .75 sones)

Fig 25 shows an FFT plot of the output of a digital/analogue converter reproducing a dithered -90dB at 1kHz and illustrating a typical result for a multibit converter. When analysed for the situation 0dBFS = 112dB spl, the excitation pattern is as shown in Fig 26. Whilst this distortion may be considered insignificant at this level, the model clearly predicts that the error will be detectable. It is interesting that the loudness of the complex of Fig 25 is 35phon (.52 sone), whereas for an undistorted tone we get 23 phon (.23 sone); that is to say the distorted tone will sound twice as loud!

This distortion can then be variously described as; −95dBr, 50% THD, masked loudness of 30 phon (0.49 sone), or twice as loud as the original!

Fig 27 shows a result, post-processing a test on a data-reduction system. In this case 0dBFS is set to 110dB spl. The three curves show the masked threshold for the original 50dB spl tone and the calculated excitation curves for both the error signal and the no-signal background-noise of the device. We see that the model predicts an audible error in this case. Fig 28 shows a result predicting clearly audible modulation-noise for the same device in response to a very modest 100Hz -50dB tone.

9. PERCEPTION OF ERRORS

The model described in section 7 is based on psychoacoustic data and detection models appropriate to very simple stimuli. Before considering changes to our criteria for error perception in signals of *cognitive* consequence (e.g. with meaning) we first need to review relevant discrepancies of the *perceptual* process for simple signal groups.

9.1 Signal detection

The detection criterion for data produced by the programs are based on the energy in each auditory filter. As a starting point, if an error signal is calculated to have a masked loudness greater than zero then it will be audible to a degree indicated by its loudness in that context. Similarly if adding an error signal changes the graphic result for excitation at any frequency then it will be detectable. (Strictly we look for a change that exceeds the appropriate DLI; we can often use a 0.5dB criterion).

9.2 Threshold function

Auditory signal-detection is a stochastic process, the 'detector' operates near threshold in the presence of significant noise, and thresholds as such are points determined on a psychometric function by statistical testing. The probability of detecting a signal is low (but finite) below threshold and high above it. Fig 29 shows a representative psychometric function based on a very useful form derived in [2] for relating probability of detection to signal level. This underlines the point that unlike a comparator threshold, there will be circumstances in which signals below nominal threshold (absolute or masked) may be detected, and of course the reverse.

Since one purpose of this study is to help to identify the most rigorous criteria for error detection, we need to review circumstances that lead to effective threshold reduction.

9.3 Below 'threshold' detection

Schafer [32] showed that tone complexes were detectable in circumstances where each component in isolation was below threshold.

This is not surprising for signals within one auditory filter since the *total* energy of the combination controls detection.

It has also been shown that combinations of otherwise separately inaudible tones can be detected when the complex is *wider* than an auditory filter, and that indeed there is a mechanism whereby the likelihood of detection increases as more auditory filters are excited below threshold.

Buus [2] shows that the shape of the psychometric function described in Fig 29 holds for single tones and for complexes, the function moving to the left for complexes. He gives the basic result of importance here – namely that the auditory detection system can be described by an independent-thresholds model. According to this model detection occurs when a stochastic variable in any one or combination of channels exceeds threshold and that a number of channels (or filters in this case) can be monitored simultaneously. From [2] the threshold reduction for n equally detectable but individually subthreshold signals is:

$$10\log(n)^{0.5}\ \mathrm{dB} \tag{11}$$

In the example given by Buus this reduction is 6dB for 18 components. Since there are 35 ERBs in the auditory range, the maximum masked threshold reduction we could expect for a uniformly exciting complex may be close to 8dB.

The investigator therefore needs to be alive to the possibility that combinations of errors may arise in the perithreshold region which together, or when combined with other non-stationary noises become detectable. As a rule-of-thumb one would cap this effect for errors 6dB below the nominal threshold. Some general points are summarised.

- The nominal masked threshold for a tone is determined in a test which teaches the listener what the frequency will be. From a signal detection point of view this means $n = 1$.
- When the listener is directed by instruction (or in a cognitive test – by internally driven expectation) to listen for a tone complex then the threshold will be lower than for a single tone.
- Wideband noises are governed in this respect by the rules for multiple complexes.
- When the listener does not know where the signal will be, as in tests presenting randomly chosen tones, the threshold will be higher since more than one filter needs to be monitored while only one is capable of exceeding the threshold.

The author contends that in general, music listening generates a situation based on expectation and feedback where the threshold to errors will more often be lowered than raised.

Often when the detection threshold for a wide-band noise signal is determined by a listening test, an analysis of the noise using either critical-band, auditory-filter or even 1/3 octave summation models give results that indicate the noise to be below threshold. Two published examples are in [10] and [41] where discrepancies between 3 and 5dB are noted. Fig 30 shows an example. The gain of a digital active-speaker was set in an experiment to determine the just-audibility of its dither noise at 1m. The acoustic spectrum was determined by FFT analysis. When this data was post-processed by the programs described earlier, the noise was predicted to be inaudible and had to be raised 3dB to achieve a positive loudness.

There are two reasonable explanations for this. The region between 1kHz and 5kHz which is the first to exceed threshold as the noise is raised, covers 12ERBs, so we could expect an average reduction in threshold of 5dB. The second explanation is that the microstructure of each individual's threshold can vary by this amount, leading to at least one filter operating ahead of the others.

9.4 Threshold reducing phenomena

There are several mechanisms, now clearly understood, which result in masked thresholds being lower than those indicated by the energy or excitation criterion of simultaneous masking. It is impossible to review these in this paper, however the reader is referred to [43] for a suitable starting point.

For completeness, the following list is given.

- Below-threshold beats. [15] [45]
- Distortion in the ear itself.
- Tonal masking reduction. Fig 31 shows tone and noise masking at 1kHz.
- Beat formation.
- Binaural masking differences. [7] [45] [25] [18]
- Comodulation masking release. [16] [17] [4].
- Spatial summation.

10. NON-LINEAR DISTORTIONS

Fig 32 gives collected results from the model of the way the masked-threshold for a second harmonic tone varies with frequency and spl of the fundamental. It can be seen that the minimum sensitivity occurs at different frequencies depending on the spl of the fundamental. Fig 33 shows the loci of the minima as frequency functions for the 2nd, 3rd, 6th and 8th harmonics. Obviously the numbers are comfortably in line with accepted wisdom on harmonic distortion.

It should be recognised that energy-based detection is not the only criterion we should apply here. Adding audible or just-audible harmonics to a fundamental also changes its timbre. The most benign addition we can make to a musical tone is to add the second harmonic. As we rise through the harmonics a significant change occurs at the 7th where the distortion product is no longer consonant.

Of course the total-harmonic distortion (THD) criterion is often used in audio, not because it describes the audibility of harmonics, but because it indicates system linearity. A much more relevant mechanism in systems exhibiting non-linearity is intermodulation distortion.

A system that produces second harmonic distortion will also produce a first order difference tone (3dB below the harmonic level) when stimulated by two tones.

Fig 34 illustrates the predicted threshold for this mechanism as a function of difference-frequency and spl (20–90dB) when the lower-frequency tone is 10kHz. It is clear that a higher sensitivity to the difference product is indicated. The curves above 80dB spl should be treated with caution since above this level the ear starts to introduce a difference-tone of its own. However below that level IM distortions as low as 0.01% are indicated as significant. This is an important result.

10.1 Maximum error criterion

The previous section included a list of mechanisms and circumstances that may lead to underestimation of errors using models of masking and excitation based on steady-state energy criteria.

If we desired to create a guard-band in a design that gave reasonable confidence that for all listeners the error mechanism were not significant on program material, then two things are required.

- A modification to allow for hearing differences between the averages presented in ISO 226 and in the auditory-filter bandwidths to represent the most 'keen-eared' subjects.
- A guard-band to allow a distance below detection to allow for the release mechanisms discussed and any that are not known or considered such as differences arising from temporal and pitch characteristics which have been ignored so far.

The first is easily accommodated by changing parameters in the model and this is currently under development by the author. The second is a judgment, one would have to chose a value between 10 and 15dB to be reasonably certain that an error would not be released. The combination of these considerations could lead to errors being sought 20dB or more below nominal audibility. In some cases this will be easily achieved, in others it may lead to significant over-engineering.

10.2 Minimum noise criterion

Recently there have been suggestions for modifying the noise-spectrum of a dither signal to make the resulting wideband noise minimally audible [12] [46].

The general approach has been to attempt to create a noise that follows the hearing threshold. Obviously if this were the objective then the spectral density of the dither would have to be shaped to the contour, taking account of the summation bandwidth of the auditory filter. Having done this we would then have a uniformly exciting noise. For interest Fig 35 shows the specific loudness result for the function proposed by Wannamaker in [46]. It can be seen that it does not quite achieve a flat result, but could easily be modified to do so. For this noise, the model indicated that the intensity of the noise-shaped dither had to be increased 15dB above the unshaped-dither for both to be audible. Detectability mode, using a the independent thresholds criterion gave the result to be 8dB. Comparing the dither of Fig 35 with flat dither, at the high listening level, the loudnesses compared at 42 phon (1.2 sone) for the flat dither, and 14.3 phon (0.04 sone) for the shaped dither. This result shows very clearly the usefulness of a model like this to the audio engineer.

11. COGNITION OF ERRORS

So far we have considered the basic perceptual aspects of sounds and in particular of certain peri-threshold signals within an overall presentation. This class of sounds is already very complex to interpret. For example the simple process of determining the masked loudness of a tone will depend on whether it becomes incorporated in an existing percept – as for example will happen adding a second harmonic – or whether it is separated in the cognitive process into a separate sound with its own loudness.

The difficulty we face is to take this understanding forward to begin to explain the kinds of things we hear when system errors are modified. For example, an improvement in a high-performance system may well result in the separation of one instrument into two, or change markedly the impression of the acoustic space in which the recording was made. Both effects can be caused by objectively small differences making a significant change to the impression or cognitive response.

Space does not permit developing this discussion here. The reader is however referred to [43], which develops this area in more detail.

12. CONCLUSIONS

This paper has introduced the idea of examining errors in audio systems by first establishing the basic perceptual parameters for the signal. In summary.

- The inherent non-linear behaviour of human perception requires us to interpret system errors in the final acoustic domain and at the correct listening spl. Distortions or noises cannot be finally assessed in abstract as a ratio of some parameters – e.g. signal/noise ratio.
- By transforming the spectral content of the signals under examination onto a form resembling the frequency selectivity of the cochlea/neural system of the basilar membrane, we can more readily see a picture of the excitation produced by the signal or error.
- In examining this spectral spread of excitation for a signal, we can make fairly sophisticated estimates of the simultaneous masking it produces. By so doing, a first-order estimate can be obtained for the 'detectability' i.e. audibility or otherwise of an added error.
- A model was evolved that also allows excitation and loudness of a signal to be estimated fairly accurately for both absolute-threshold and other simultaneously masked situations. The frequency range was extended to cover 20Hz to 15kHz. The model was fitted for amplitudes between 0 and 100dB spl with particular emphasis on accuracy in the peri-threshold to 80 phon region.
- The model has been verified by comparing its results to several sets of published data and by its ability to explain critical-band phenomena and several types of masking anomalies. Particular attention is paid to determining the audibility of different noise-bands, especially near absolute and masked thresholds.

Of necessity a large part of the paper deals with the basic psychoacoustic concepts and also with the considerations leading to a series of computer programs that will give graphical results for excitation, masking and loudness as well as overall masked loudness and detectability numbers. The graphs are of course important when it comes to interpretation.

Some examples have been given illustrating the kind of results that can be expected. The author finds it very illuminating to be able to view errors in the context of listening – and in a graphical form that reflects audibility.

With a bias towards identifying the performance limit required by audio components, perceptual phenomena have been reviewed that tend to modify criteria of detection based on simultaneous masking and energy-detection models.

The models were modified as a result of these considerations to include a detection mode for evaluating the probability of detection for multipart signals in the region just below predicted audibility.

13. ACKNOWLEDGEMENTS

The author would like to thank Brain Moore, Roy Patterson and Brian Glasberg for many helpful, illuminating and stimulating conversations. Brian Moore is also gratefully acknowledged for his review of this work. The author thanks Dolby (UK) Ltd. for permission to publish results of measurements made on their AC2 data-reduction system.

14. REFERENCES

1 Berger, E.H., 'Re-examination of the low-frequency normal threshold of hearing in free and diffuse sound fields', J. Acous. Soc. Amer., Vol 70, No.6, pp 1635-1645, (Dec 1981)

2 Buus, S. et al. 'Decision rules in detection of simple and complex tones', J. Acous. Soc. Amer., 80, 1646-1657 (1986)

3 Buus, S. et al. 'Tuning curves at high-frequencies and their relation to the absolute threshold curve', in 'Auditory Frequency Selectivity', eds. Moore, B.J.C. and Patterson, R.D., Plenum Press (1986)

4 Buus, S. 'Release from masking caused by envelope fluctuations', J. Acous. Soc. Amer., 78, 1958-1965, (1985)

5 Cacace, A.T. and Margolis, R.H., 'On the loudness of complex stimuli and its relationship to cochlear excitation', J. Acous. Soc. Amer., 78, 1568-1573. (1985)

6 Carterrette, E.P. and Friedman, M.C. 'Handbook of Perception', *Ac. Press* vol II, 'Psychophysical judgment and measurement' (1978)

7 Carterrette, E.P. and Friedman, M.C. 'Handbook of Perception', *Ac. Press* vol IV, 'Hearing' (1978)

8 Deutsch, D., 'The psychology of music', Ac. Press 1982

9 Egan, J.P. and Hake, H.W., 'On the masking pattern of a simple auditory stimulus', J. Acous. Soc. Amer., 22, 622-630 (1950)

10 L J Fielder, 'Human auditory capabilities and their consequences in digital audio converter design', *Proc AES 7th, 'Audio in digital times'*, p45, (1989)

11 Fletcher, H. Auditory patterns. Rev. Mod. Phys. 12, 47-65 (1940)

12 Gerzon, M. and Craven, P.G., 'Optimal noise-shaping and dither of digital signals', AES 87th, preprint no. 2822 (1989)

13 Greenwood, D.D., 'Critical bandwidth and the frequency coordinates of the basilar membrane', J. Acous. Soc. Amer., 33, 1344-1356 (1961)

14 Greenwood, D. D. A cochlear frequency-position function for several species - 29 years later. J. Acoust. Soc. Am. 87, 2592-2605 (1990)

15 Groen, J J., 'Super and subliminal binaural beats', *Acta Oto-Lar*, vol 57, p224

16 Hall, J.W. et al. 'Comodulation masking release: evidence for multiple cues', J. Acous. Soc. Amer., 84, 1669 – 1675 (1989)

17 Hall, J.W. 'Experiments on Comodulation Masking Release', in 'Auditory processing of complex sounds', Eds Yost, W.A. and Watson, C.S., Erlbaum and Assoc. (1987)

18 Houtsma, A.J.M. et al, 'Auditory Demonstrations', CD issued by IPO, NIU and Acous. Soc. Amer., Philips 1126-061, tracks 75 – 79 and 1.

19 ISO 226: 1987, 'Acoustics – Normal equal-loudness level contours'

20 ISO 131-1979 (E), 'Acoustics – Expression of physical and subjective magnitudes of sound or noise in air'

21 ISO 532: 1975E, 'Acoustics – Method for calculating loudness level'

22 Long, G.R., 'The microstructure of quiet and masked thresholds', Hearing Res., 15, 73-87 (1984)

23 Miller, G.A., 'Sensitivity to changes in the intensity of white noise and its relation to masking and loudness', J. Acous. Soc. Amer., Vol 19, pp 609-619 (1947)

24 Moore, B.C.J., Peters, R.W., Glasberg, B.R., 'Auditory filter shapes at low centre frequencies', J. Acous. Soc. Amer., Vol 88, pp 132-140, (1990)

25 Moore, B.C.J. 'An introduction to the psychology of hearing', *Ac. Press*

26 Moore, B.J.C., Glasberg, B.R., 'Suggested formulae for calculating auditory-filter bandwidths and excitation patterns', J. Acous. Soc. Amer., 74, 1003-1014 (1983)

27 Moore, B.J.C., Glasberg, B.R., 'The role of frequency selectivity in the perception of loudness, pitch and time', in 'Frequency Selectivity in hearing', ed Moore, B.J.C., Ac. Press (1986)
28 Moore, B.J.C., private communication
29 Patterson, R.D. and Moore, B.J.C., 'Auditory filters and excitation patterns as representations of frequency resolution', in 'Frequency Selectivity in hearing', ed Moore, B.J.C., Ac. Press (1986)
30 Patterson, R.D., 'Auditory filter shapes derived with noise stimuli', J. Acous. Soc. Amer., 59, 640-654. (1976)
31 Robinson, D.W. and Dason, R.S. in ISO131-1959
32 Schafer, T.H. and Gales, R.S., 'Auditory masking of multiple tones by random noise', J. Acous. Soc. Amer., 21, 392-398.
33 Scharf, B., 'Critical bands and the loudness of complex tones near threshold', *JASA* vol 31, p365, (1959)
34 Scharf, B. and Fishken, D., 'Binaural summation of loudness', J. Exp. Psychology, vol 86, 374-379 (1970)
35 Shailer, M.J., Moore, B.C.J., Glasberg, B.R., Watson, N., Harris, S., 'Auditory filter shapes at 8 and 10kHz', J. Acous. Soc. Amer., Vol 88, pp 141-148, (1990)
36 Soderquist, D.R. Lindsey, J.W., 'Physiological noise as a masker of low frequencies: and the cardiac cycle', J. Acous. Soc. Amer., 52, 1216-1220 (1972)
37 Stevens, S.S., 'The measurement of loudness', J. Acous. Soc. Amer., Vol 27, pp 815-829 (1955)
38 Stevens, S.S., 'Procedure for calculating loudness: Mark VI', J. Acous. Soc. Amer., Vol 33, pp 1577-1585 (1961)
39 Stevens, S.S., 'Perceived level of noise by Mark VII and decibels', J. Acous. Soc. Amer., Vol 51, pp 575-601 (1972)
40 Stevens, S.S., 'Loudness functions under inhibition', Perception and Psychophysics, Vol 2, pp 459-465
41 Stuart, J.R. 'High Quality Digital Audio', *Proc. Inst. Acoust.* **12**, part 8 (1990)
42 Stuart, J.R. 'Predicting the audibility, detectability and loudness of errors in audio systems', preprint for AES 91st convention. (1991)
43 Stuart, J.R. 'Estimating the significance of errors in audio systems', preprint for AES 91st convention. (1991)
44 Terhardt, E., 'Music perception and sensory information acquisition: relationships and low-level analogies', Music Perception vol 8 no 3, pp 217-239, Spring 1991
45 Tobias, J.V., 'Foundations of modern auditory theory', *AP* (1970)
46 Wannamaker, R.A., 'Psychoacoustically optimum noise-shaping', AES 89th, Preprint no. 2965 (1990)
47 Warren, R.M., 'Auditory perception, a new synthesis', Pergamon (1982)
48 Zwicker, E., 'Spontaneous oto-acoustic emissions, threshold in quiet and just-noticeable amplitude-modulation at low levels', in Auditory Frequency Selectivity, Plenum Press (1986), NATO ASI A-119, pp 49-59.
49 Zwicker, E., Flottorp, G. and Stevens, S.S. 'Critical bandwidth in loudness summation', J. Acous. Soc. Amer., Vol 29, pp 548-557 (1957)

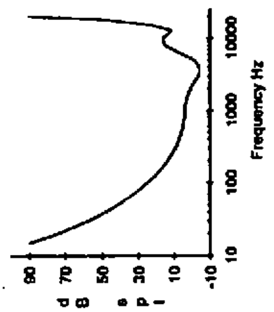

Fig 1. Minimum audible field, frontal incidence.

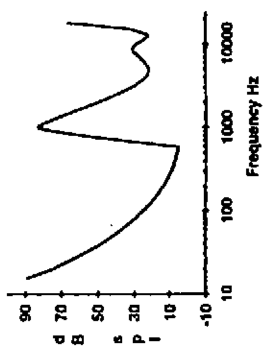

Fig 2. Masked threshold for 90dB spl at 1kHz

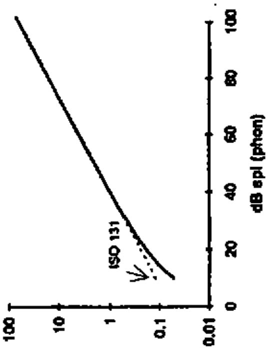

Fig 3. Loudness function for 1kHz tone

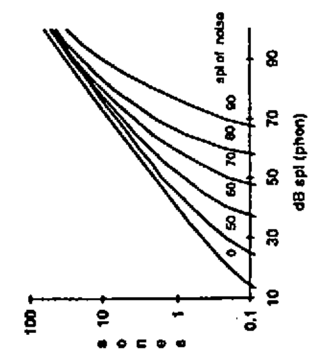

Fig 4. Loudness of 1kHz tone masked by white noise

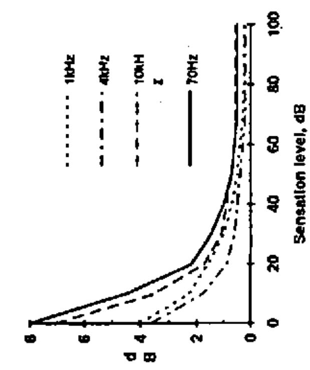

Fig 5. Unmasked DL for intensity

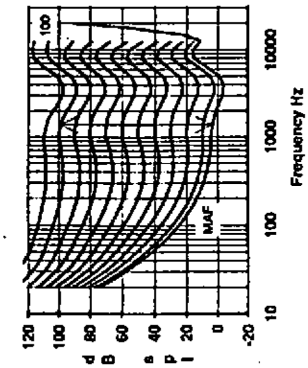

Fig 6. Equal-loudness contours

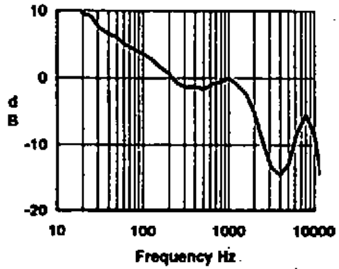

Fig 7. Difference for headphone presentation

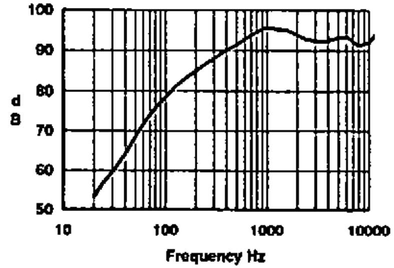

Fig 8. Dynamic range at 100 phon

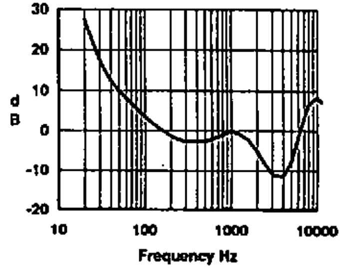

Fig 9. Form of correction

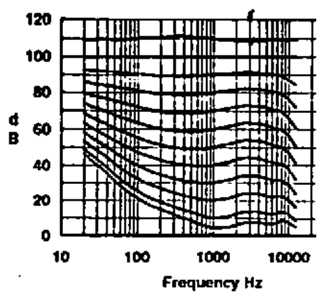

Fig 10. Equal loudness curves for normalised cochlear input

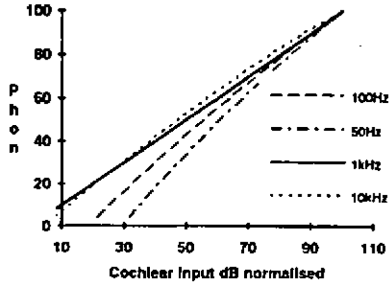

Fig 11. Growth of loudness for Lc

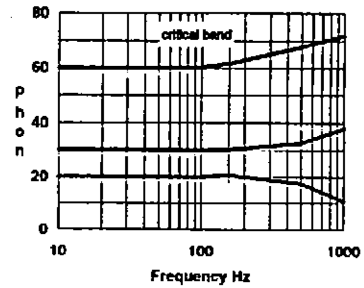

Fig 12. Variation of loudness with bandwidth at three levels

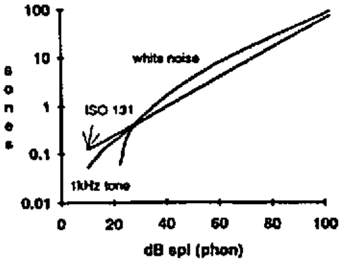

Fig 13. Growth of loudness for white noise

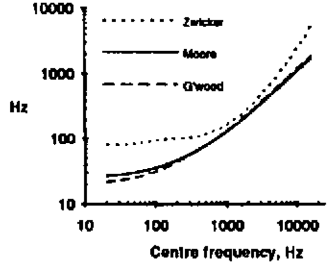

Fig 14. Estimates of width of the auditory filter

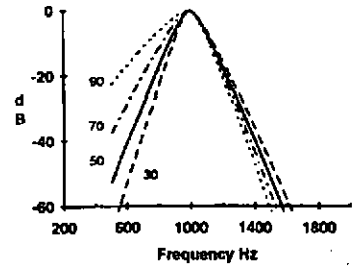

Fig 15. Auditory filter response at different levels of excitation

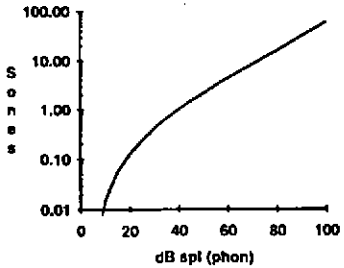

Fig 17. Loudness result from model at 1kHz

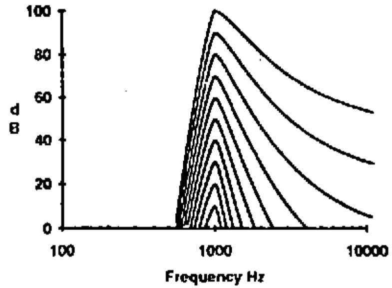

Fig 18. Excitation patterns calculated for 1kHz

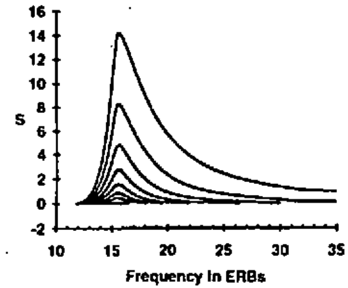

Fig 19. Specific Loudness patterns for 1kHz

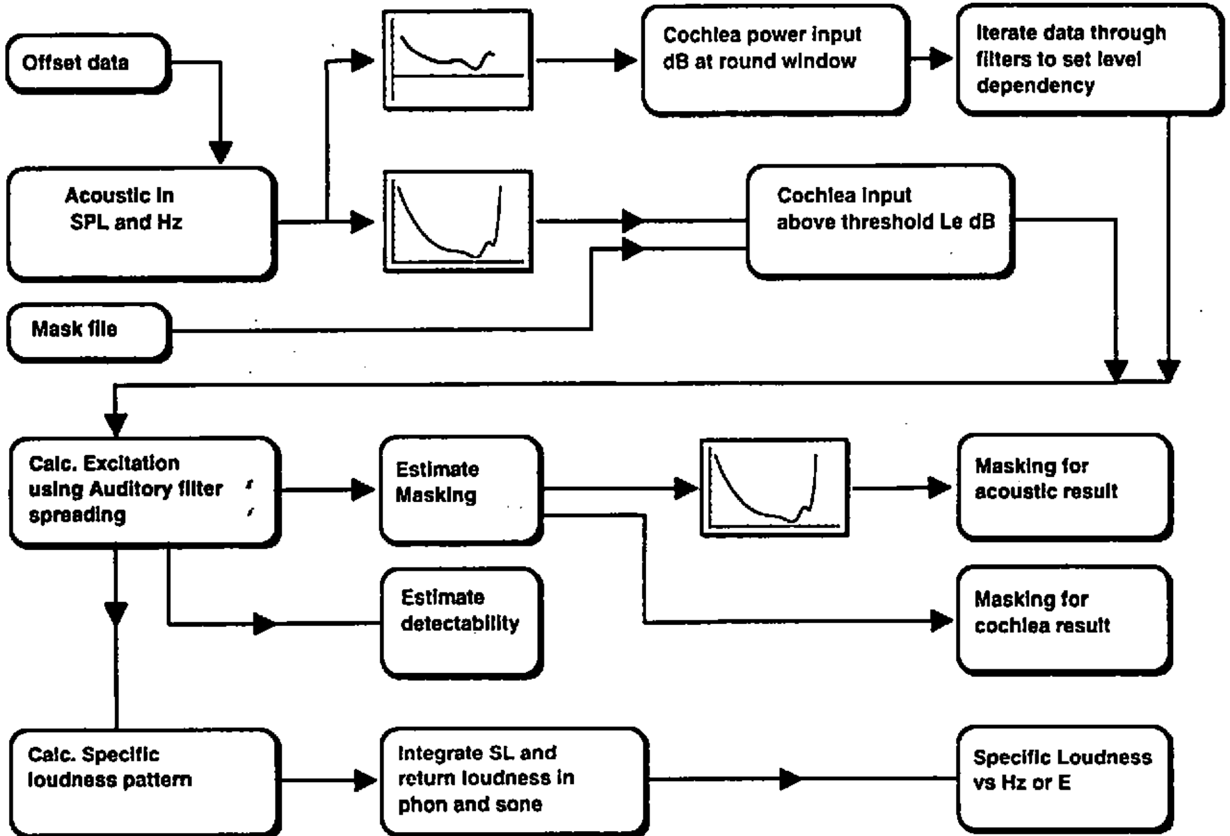
Offset data
Acoustic in SPL and Hz
Mask file
Cochlea power input dB at round window
Iterate data through filters to set level dependency
Cochlea input above threshold Le dB
Calc. Excitation using Auditory filter spreading
Estimate Masking
Masking for acoustic result
Estimate detectability
Masking for cochlea result
Calc. Specific loudness pattern
Integrate SL and return loudness in phon and sone
Specific Loudness vs Hz or E

Fig 17.

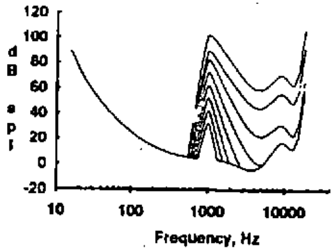

Fig 20. Masked thresholds for 1kHz; 40dB to 110dB spl

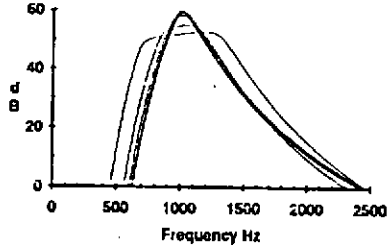

Fig 21. Excitation patterns for complexes of varying width centred on 1kHz

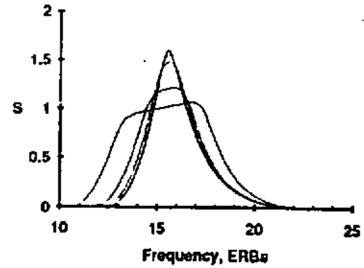

Fig 22. Specific Loudness patterns corresponding to Fig 21.

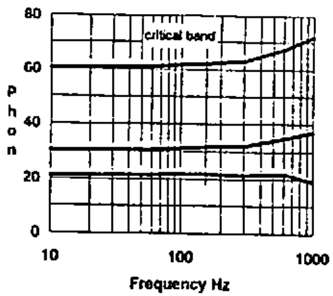

Fig 23. Predicted loudness variation with bandwidth

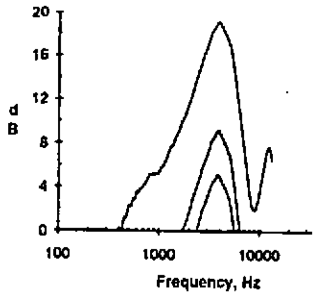

Fig 24. Excitation patterns for noise at replay levels ref. 122, 112 and 108dB

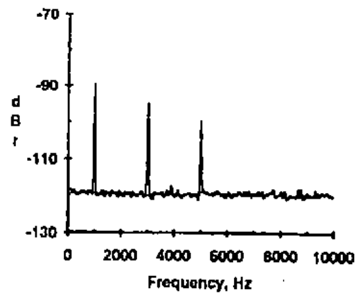

Fig 25. FFT analysis of D/A distortion

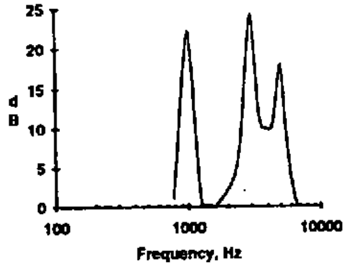

Fig 26. Excitation pattern for the distorted signal of Fig 25

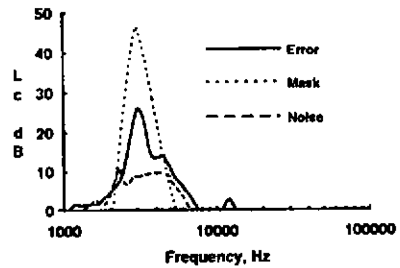

Fig 27. Excitation patterns for noise and signal error, compared with masking

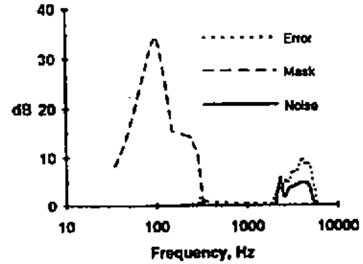

Fig 28. Excitation pattern of noise, signal-induced error and mask.

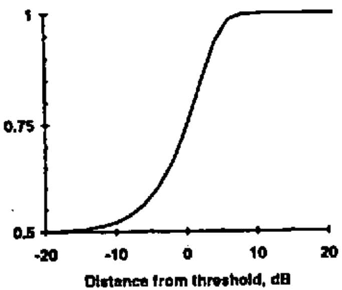

Fig 29. Psychometric function. Probability of detection vs distance from threshold

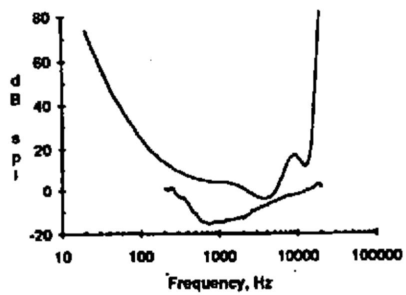

Fig 30. Excitation pattern of system noise at observed 'just audible'

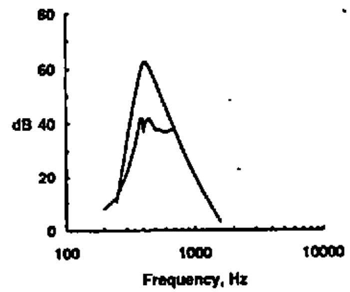

Fig 31. Difference in masking pattern for narrow noise and tone at 400Hz

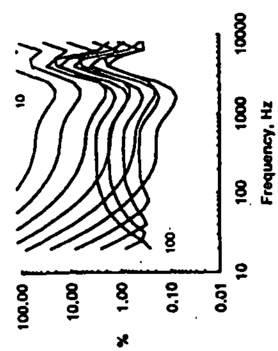

Fig 32. Calculated audibility thresholds for 2nd harmonic of Hz, between 10 and 100dB spl

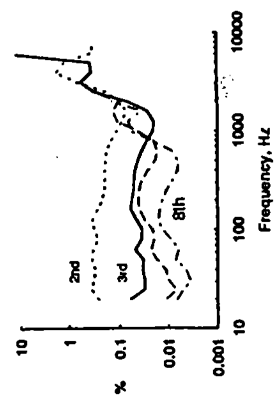

Fig 33. Calculated audibility thresholds for 2nd, 3rd, 6th and 8th. Curves plot minima for each harmonic

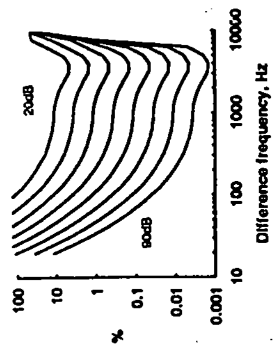

Fig 34. Calculated threshold of audibility of first-order difference distortion; lower tone at 10kHz

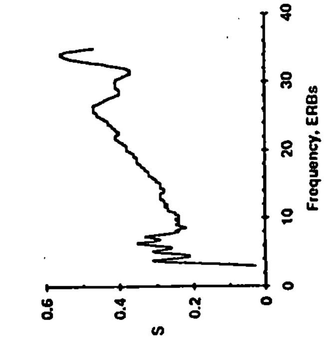

Fig 35. Loudness pattern for noise-shaped dither