PSYCHOACOUSTIC MODELS AND NON-LINEAR HUMAN HEARING

David J M Robinson, Malcolm O J Hawksford

Centre for Audio Research and Engineering
Department of Electronic Systems Engineering
The University of Essex
Wivenhoe Park
Colchester CO4 3SQ
United Kingdom
PHONE: +44 (0)1206 872929 FAX: +44 (0)1206 872900
e-mail: DavidR@europe.com, mjh@essex.ac.uk

Abstract - Non-linearity in the human ear can cause audible distortion not present in the original signal. Such distortion is generated within the ear by inter-modulation of a spectral complex, itself containing possible masked components. When psychoacoustic codecs remove these supposedly masked components, the in-ear-generated distortion is also removed, and so our listening experience is modified. In this paper, the in-ear distortion is quantified and a method suggested for predicting the in-ear distortion arising from an audio signal. The potential performance gains due to incorporating this knowledge into an audio codec are assessed.

0 INTRODUCTION

Perceptual audio codecs aim to discard signal components that are inaudible to human listeners. Typical codecs (e.g. [1]) calculate theoretical masking thresholds from quasi-linear models of the human auditory system. It is assumed that any signal components below the masking threshold may be disregarded, without causing any audible degradation of the signal. However, the ear is a non-linear device, and linear models of masking do not account fully for its behaviour. In particular, signal components that are predicted to be inaudible by a linear analysis are found to be very audible to a real, non-linear, human ear.

This concept first came to the authors' attention through the following example. The signal illustrated in Figure 1 is intended to demonstrate spectral masking. An 800 Hz tone and a series of tones, ascending from 400 Hz to 1600 Hz, are presented simultaneously. If the amplitude of the stepped tones is smaller than that of the 800 Hz tone (e.g. 20-40 dB down), then in the region around 800 Hz, they will be inaudible, as they are masked by the louder tone. However, as the inaudible stepped tones pass above 800 Hz, listeners often perceive a second series of tones, descending in frequency, below 800 Hz. These are illustrated by the shaded blocks in Figure 1. They are not present in the actual signal, but are generated by distortion within the human auditory system.

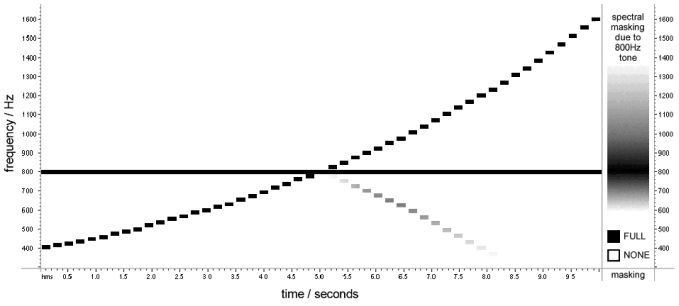

Figure 1 – Masking of stepped tones due to 800Hz tone, and resulting cubic difference tones.

As the ascending stepped tones are supposedly masked at this point, this raises an interesting question: If an audio codec removes inaudible sounds, what will be the effect if it removes these masked tones? Surely, the (audible) descending distortion tones will also be removed, thus changing what a human listener hears. This is precisely what a good audio codec should **not** do. The audible effect, especially for more complex audio signals, may be slight. However, transparent audio coding claims to make no audible change to the signal whatsoever, so this effect merits investigation.

This paper will focus on the most audible distortion component generated by the human ear: the cubic distortion tone (CDT). Possible methods of determining the amplitude and frequency of this internal distortion tone will be discussed, and an equation that accurately predicts these properties will be presented. The CDTs generated by two tones will be examined for the case where one tone is below the masking threshold predicted by a psychoacoustic model, and the audibility of the resulting CDT will be determined. The true nature of the masking threshold will be discussed, and the extent to which the CDT can mask other spectral components will be examined. Finally, the relevance of these theoretical calculations to real world applications will be assessed. The study commences by examining the properties of the cubic distortion tone.

1 THE CUBIC DISTORTION TONE

The frequency of the cubic distortion tone, arising from two primary frequency components, f_1 and f_2 ($f_1<f_2$) is given by

$$f_{CDT} = 2f_1 - f_2 \tag{1}$$

The cubic distortion tone (CDT) is so called because a difference tone at this frequency is generated by a 3rd order polynomial transfer function. An early hypothesis [2] suggested that the bones in the middle ear were responsible for such a transfer function, thus giving rise to the distortion component given by equation (1). However, it is now widely believed that the cubic distortion tone is generated within the cochlea, by the action of the outer hair cells [3]. These hair cells are part of the cochlea amplifier – a mechanism whereby the basilar membrane motion due to incoming sound waves is varied by an active process, which is beyond the scope of this paper. Further details may be found in [4] and [5], but here it suffices to understand that this gain control function of the ear generates, as a by-product, the cubic distortion tone.

Though calculating the frequency of the CDT is trivial, determining the amplitude or perceived loudness of this tone is a more difficult task. There are three possible methods of gathering this data from human listeners, which will now be discussed in tern.

1.1 Distortion-product otoacoustic emissions

An otoacoustic emission is a sound generated by the ear, which can be detected by objective rather than subjective means. In our present study, the otoacoustic emission is due to two external tones yielding a distortion product within the ear, hence the name.

In 1979 [6] it was found that the cubic distortion tone can be detected by a probe microphone inserted into the ear canal of a listener who is presented with two appropriate primary tones. The fact that the cochlea-generated tone propagates back through the auditory system, into the ear canal, allows the amplitude and phase of the CDT to be recorded, without relying on subjective feedback from the listener.

Figure 2 shows a diagram of the apparatus that may be used. The two primary tones, f_1 and f_2 are generated by two separate loudspeakers. Two loudspeakers are used to prevent any distortions that may be created by the speaker itself, if two tones were generated by a single device. The two signals are fed via rubber tubes into an earpiece containing a miniature microphone. The signals first mix acoustically in the ear canal, and the levels of the primaries are calibrated from the microphone in-situ.

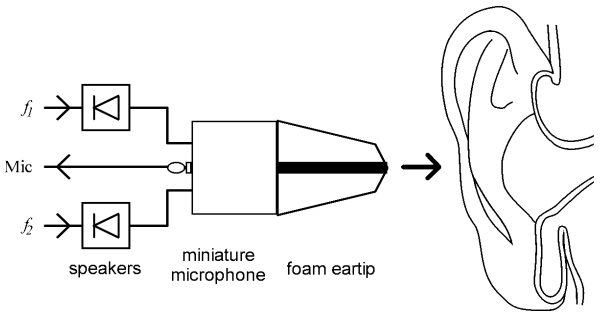

Figure 2 – Apparatus used for the measurement of DPOAEs.

By varying the amplitude and frequency of the two primaries, it is theoretically possible to map the complete CDT response of the human auditory system (e.g. [7]). However, comparing the reported subjective level of the CDT with the measured DPOAE level reveals a large, frequency dependent difference.

The problem lies in the transmission of the CDT from the site of origin within the cochlea, back through the auditory system via the middle ear, into the ear canal. It has been suggested [8] that this reverse path accounts for a 12 dB loss for frequencies around 1-1.5 kHz, and that the loss increases at around 12 dB/octave either side of this frequency region. Unfortunately it is not possible to measure the transfer function of this reverse path in any direct manner, so it can only be inferred by comparing measured and subjective data.

Thus the DPOAE fails to yield an objective, absolute measure of the amplitude of the CDT within the human cochlea. As calibration of the DPOAE relies on subjective data, it seems sensible to turn to that subjective (psychoacoustic) data as an indication of the amplitude of the CDT. Two main psychoacoustic methods have been used to determine the CDT level caused by given stimulus conditions, as follows.

1.2 Loudness matching

In this method, a listener is instructed to match the loudness of the CDT with that of a probe tone of the same frequency presented externally, but non-simultaneously. This subjective judgement is made more reliable by pulsing the primary tones (and hence the CDT), and the probe tone alternately, as shown in Figure 3. Thus, when the level of the internal CDT and the external tone are matched, the listener will hear a continuous tone, whereas if the probe level is too high or too low, then the pulsing will be clearly audible. Data from such an experiment [9] will be referred to later in this paper.

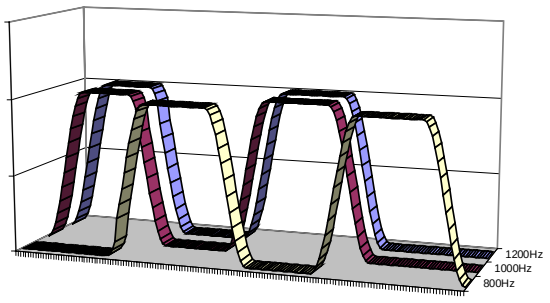

Figure 3 – Measurement of CDT level by pulsing primary tones and probe tone alternately.

1.3 Cancellation Tone

The second subjective method of determining the level of the CDT is to attempt to cancel the difference tone using an external tone. In addition to the two primary tones, the listener is presented with a third tone - the amplitude and phase of which are completely under their control. A highly trained listener can adjust these two parameters until the internal CDT is completely cancelled out by the external tone. At this point, the amplitude of the external tone is assumed to match that of the internal CDT. One advantage of this method is that the phase of the internal tone can also be calculated, as being 180º out of phase from the external tone.

Many experimenters have employed this method, e.g. [10], [11] and [9] again. Several key features are:

1. The CDT level determined via the cancellation tone method can be used to predict the masking due to the CDT to within 2 dB [12].
2. The phase prediction via this method is an “equivalent external phase” and its relationship to the actual internal phase of the CDT is not known.
3. A complex formula has been produced to calculate the level of CDT for any pair of primary tones [13]. Such a formulaic prediction is vital if this phenomenon is to be usefully incorporated into a masking model.
4. Discrepancies exist between the CDT level as measured by this method, and that measured by the Loudness matching method. The cancellation method can produce CDT predictions up to 15 dB higher than the loudness-equivalent method.

To explain this final point briefly, the primary tone f_1 is thought to suppress the cancellation tone, such that the required cancellation tone level is larger than the perceived CDT. A long discussion of this phenomenon is given in [14].

Thus there is a wide range of data available that quantitatively describes this phenomena. Until the reverse transfer function from the cochlea to the ear canal (in the presence of auditory stimulation) has been determined, the DPOAE data, though numerous, and objective, are not suitable for the present study. This leaves the two subjective measures. The first is believed to correlate well with what we perceive, the second predicts the masking due to the combination tone well. The difference between the two can be up to 15 dB. Due to the larger amount of data available to the authors from the second type of study, the cancellation method is chosen to provide the reference CDT levels throughout the rest of this paper, with the proviso that the real level may be slightly lower.

2 MODELLING THE CDT

As mentioned previously, there exists a formula [13] for calculating the level of the $2f_1$-f_2 cubic distortion tone L_{CDT} for any given f_1, f_2, L_1, and L_2 where L_1, and L_2 are the levels, in dB, of f_1 and f_2 respectively. However, this formula is rather complex, and includes several logarithms. To improve the computational efficiency of the formula, and to gain a clearer insight into how the CDT varies with the various parameters, the authors developed their own formulae, which are presented here. The level dependent data used to tune this formula were the same as those presented in [13]. The frequency dependent data were taken from [10], cancellation-tone results only.

The level, in external dB SPL equivalent, of the cubic distortion tone is given by

$$L_{CDT} = \frac{L_2}{2} - (0.07 - 0.00055L_2).\left(L_1 - L_2 - \frac{400}{L_2}\right)^2 - \Delta z - (0.19z_1^2 - 3.5z_1 + 22)\Delta z^{3/2} + 19.6 \quad (2)$$

Where

$$\Delta z = z_2 - z_1 \quad (3)$$

And z_n represents frequency f_n in the bark domain, thus:

$$z_n = 13\tan^{-1}\left(0.76\frac{f_n}{\text{kHz}}\right) + 3.5\tan^{-1}\left(\frac{f_n}{7.5\text{kHz}}\right) \quad (4)$$

This equation is taken from [15], and was used for the following graphs, for consistency with [13]. A more accurate (and more computationally efficient) formula can be found in [16] which has the advantage of being invertable (i.e. yields *f* from *z* as well as *z* from *f*). The formulae include frequencies in the bark domain because most psychoacoustic audio codecs process the frequency information in the bark domain when considering the masked threshold.

Equation (2) matches the measurements from human subjects for stimulus levels from 30-90 dB, and for frequencies of 1 kHz or above. This amplitude range corresponds to that which human response data was available, however, the equation behaves well outside this range, and gives realistic values (though for level which would destroy human hearing, the predicted L_{CDT} is doubtful!). Any calculated L_{CDT} below the threshold of hearing will be inaudible. Also any L_{CDT} masked by the primary tone f_1 may be inaudible, though beats between the CDT and f_1 may themselves be audible. The just audible L_{CDT} , derived from the minima of [17], is given by

$$L_{MIN} = \frac{L_1}{2} - 15 \quad (5)$$

Below 1 kHz the equation still follows the same trend as human subjects, but in this region the human response varies dramatically, especially for lower frequencies. If a more accurate prediction of human perception is required, the frequency dependent term in equation (2) may be replaced, thus:

$$L_{CDT} = \frac{L_2}{2} - (0.07 - 0.00055L_2).\left(L_1 - L_2 - \frac{400}{L_2}\right)^2 - \Delta z - \left(14 - \frac{1}{77}z_1^3\right)\Delta z^{2.5\sin\left(\left(\frac{\pi f_1}{2}\right)^{\frac{3}{4}}\right)} + 19.6 \quad (6)$$

The one limit to both equations (2)+(6) is that for $\Delta z < 0.45$ no CDT is audible, as it merges into the lower primary tone. This is not indicated in L_{CDT} as calculated, and must be checked separately via equation (3).

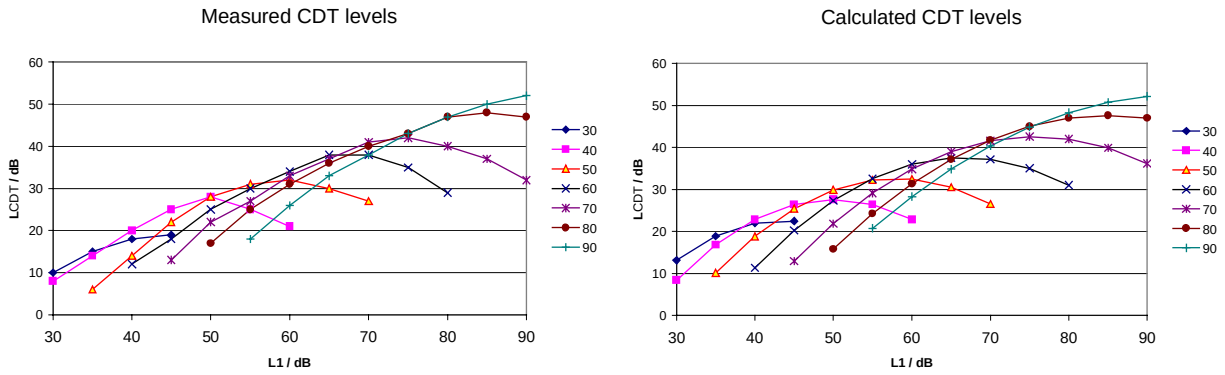

Figure 4 – Variation in L_{CDT} with L_1 and L_2 primary tone levels.

3-(a) – levels measured via cancellation method from human subjects [17];
3-(b) – levels calculated using equation (2).

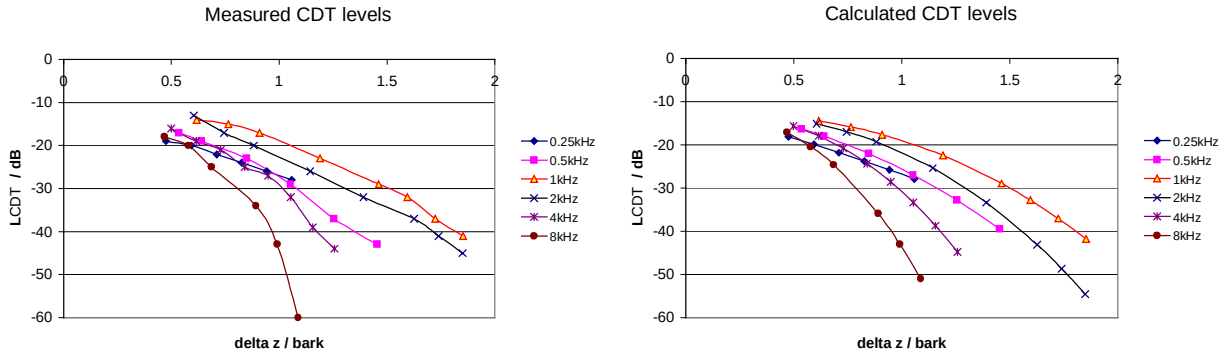

Figure 5 – Variation in L_{CDT} with f_1 and Δz.

4-(a) – levels measured via cancellation method from human subjects [10];
4-(b) – levels calculated using equations (2)-(6)

Figure 4 and Figure 5 show a comparison of the CDT level as measured from human subjects with the CDT level as predicted by equations (2)-(6). Thus the formulae are shown to be excellent predictors of the cancellation-tone measured cubic distortion tone level over a wide variety of stimulus conditions.

3 PSYCHOACOUSTIC CODECS and the CUBIC DISTORTION TONE

The task of a psychoacoustic-based codec is to reduce the amount of data required to represent an audio signal, whilst minimising the audible difference caused by this data reduction, by exploiting the properties of the human auditory system.

A typical psychoacoustic codec will calculate the theoretical masking threshold of the incoming audio signal on an instant by instant basis. Any frequency components below this masked threshold are assumed to be inaudible. Thus a signal to mask ratio can be derived by comparing the masked threshold with the actual signal level at each frequency. Then, depending on the number of bits available to code the signal, the codec can determine which frequency bands are most audible, and require accurate coding; and

which contain no audible information, and can be filled up to the masking threshold with quantisation noise, or ignored.

There are two situations where knowledge of the cubic distortion tone may improve the accuracy of this masking calculation. In the first, the codec may incorrectly remove a supposedly *inaudible* frequency component that creates an *audible* CDT within the auditory system. This mistake can be prevented by calculating the CDT level due to a dominant frequency and the closest masked spectral component, and retain the masked component if the CDT is audible. Secondly, if the CDT is large, it may itself create masking, and so yield a higher masking threshold than traditional masking threshold measures. Here, knowing the presence of the CDT may save some bits, or free some bits to encode an audible part of the signal spectrum. Each situation will be considered turn.

3.1 Masked primary tone

Consider a single 1 kHz tone @ 85 dB. This is an uninteresting (and unchallenging) signal to code, however it serves as a good example of how, even in a simple situation, a codec may remove an audible frequency component. Figure 6 shows the masking threshold of the 1 kHz tone as predicted by two psychoacoustic models: The classic Johnston model [18] and the MPEG-1 Psychoacoustic model I [1]-D. Though these are two of the simplest psychoacoustic models, the methods employed in these codecs are widely used. The lower masking thresholds predicted by the Johnston model are due to that models level of masking reduction for a pure tone – the MPEG-1 model reduces the masking prediction by around 5 dB, the Johnston model by around 25 dB relative to the masking produced by a similar amplitude noise.

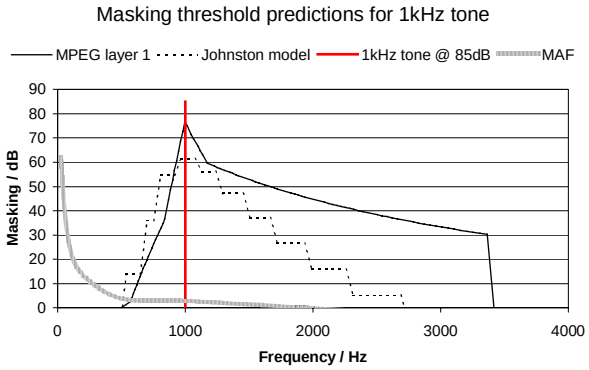

Figure 6 – Masking threshold predictions for 1 kHz tone.

A masked tone may lie in the frequency region above or below the masker, as long as it falls under the masking threshold curve. First, consider a masked tone of 1.1 kHz, i.e. one that is higher in frequency than the masker. If the level is set at the threshold of masking predicted by the MPEG-1 model (see Figure 7), then the resulting CDT is also below the predicted masking threshold. So our MPEG-1 model predicts that the tone at 1.1 kHz has no audible effect, either due to itself, or due to the resulting CDT. However, equation (5) suggests that the CDT *will* be audible at this level, and human listeners confirm this.The Johnston model predicts the masking threshold at 1.1 kHz to be 56 dB, and this matches the f_2 level at which the 900 Hz CDT is just audible. In this instance it would seem calculating the CDT merely confirms the masking prediction of the Johnston model, but shows that the MPEG-1 model is incorrect.

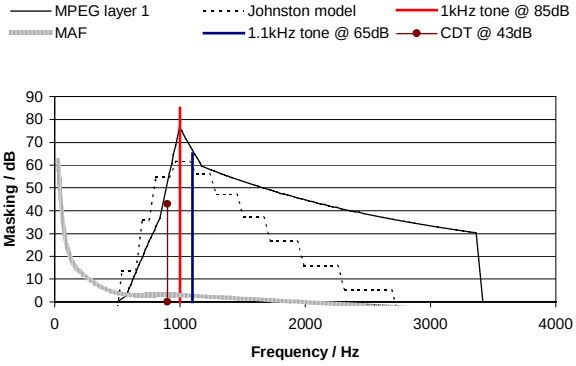

Figure 7 – Cubic Distortion Tone at masked threshold for $L_1 > L_2$

At a slightly lower f_2 frequency of 1.08 kHz @ 60 dB, the CDT is just audible (at 36 dB[1]) whereas both models predict that the CDT and the f_2 primary tone are masked. Thus both psychoacoustic models are in error. However, the spectral/intensity region over which this occurs is only 4 dB high and 100 Hz wide.

Now, consider the condition where the "masked" tone is at a lower frequency than the masker. Taking a 1 kHz tone @ 85 dB, a tone is added that the psychoacoustic models predict to be masked – a 900 Hz tone @ 52 dB. The resulting difference tone, 800 Hz @ 26 dB, also lies under the predicted masking curve, as illustrated in Figure 8. However, it is known from the formulae outlined in section 2, and from actual experimental data, that this CDT is audible.

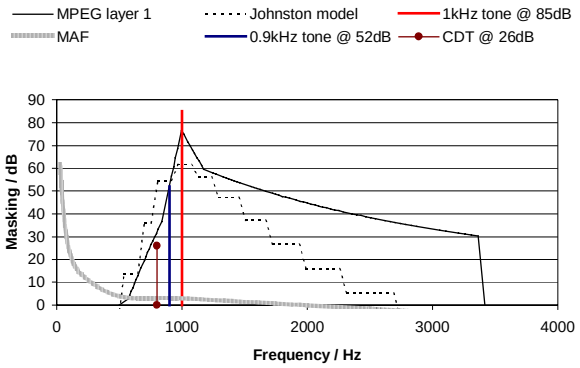

Figure 8 – Cubic Distortion Tone at masked threshold for $L_1<L_2$

These examples prove that there are possible 2-tone combinations where the quieter tone, though the psychoacoustic models predict that it is inaudible, does make an audible contribution to the sounds in the form of a cubic distortion tone at $2f_1$-f_2. Figure 9 shows the regions over which such a (theoretically masked) second frequency component will yield an audible CDT. In effect, the shaded area under the masking curve indicates the region over which the non-linearity of the ear will unmask sounds. At 8 kHz (Figure 9-b) this region is smaller, but still present for the MPEG-1 psychoacoustic model.

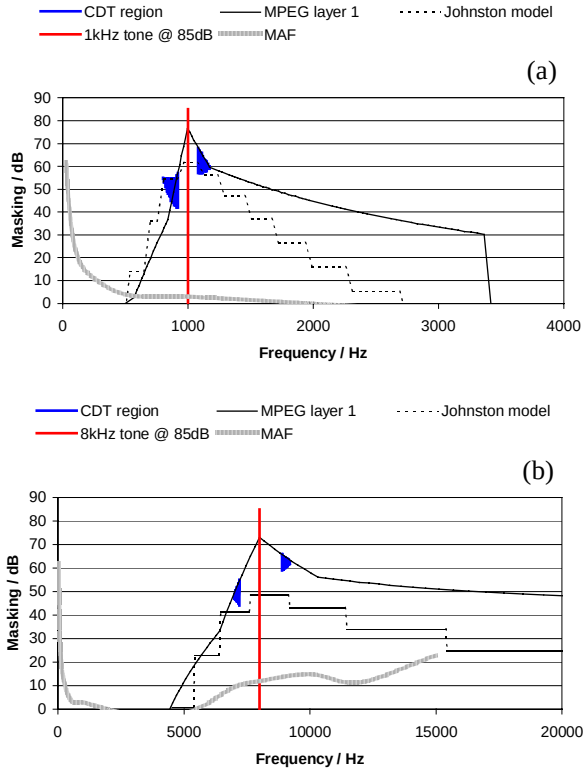

Figure 9 – Region where CDT unmasks f_2.

(a) for 1 kHz tone; (b) for 8 kHz tone.

Real World Applications

Though it has been shown that the CDT generated by non-linear properties of the human ear may cause unmasking in certain theoretical conditions, there are a number of issues to consider with respect to using this knowledge in a real-word audio codec.

Firstly, the CDT is generated by two tones. There is also a similar effect generated by two bands of noise, but it is at a much lower level [3]. Thus, the CDT is only relevant for highly tonal signals. Such signals are of a relatively low complexity, and in many instances, require less bits to code transparently than a more noise-like signal. If a fixed bit-rate codec finds that there are bits to spare after encoding the most prominent spectral peaks, it may allocate some bits to "just masked" spectral peaks. As our unmaksed f_2 is always within 15 dB of the masking threshold, it is likely that the codec may allocate some

[1] Δz=0.496 – equation (3) – if f_1 and f_2 were any closer, the CDT would be indistinguishable – see section 2

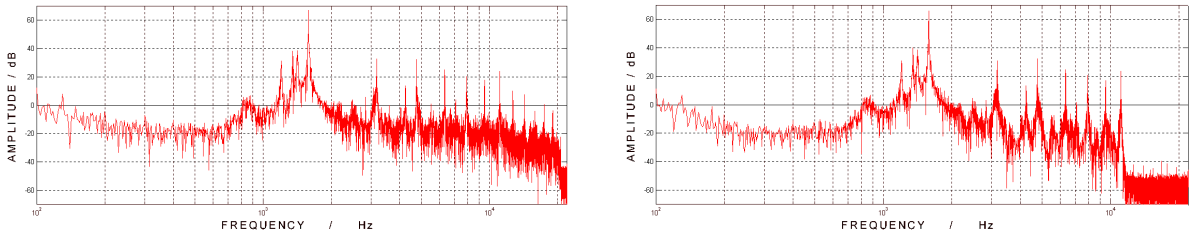

Figure 10 – Spectrum of recorder note: (a) original, (b) coded mp3 @ 96kbps

bits to it, even though it is "inaudible", since there are bits to spare and a tonal component near threshold represents a sensible allocation of those bits.

There may be very few highly tonal signals where spectral peaks are close enough to generate a CDT. Figure 10 shows the spectrum of a recorder note, taken from a commercial CD, and also the spectrum of a poor quality coded version of it. Note the spectral peaks just below the fundamental tone, caused by reverberation of the previous notes. These are closely spaced, and may cause a CDT, though the harmonic structure of a single note (without the echo of previous notes) does not have such closely spaced frequencies. In this example, the three largest spectral peaks below the fundamental are all *just above* the masking threshold (as predicted by the Johnston and MPEG-1 models) so it is not surprising that the MPEG-1 layer 3 codec retains them. The authors are unaware of any recordings containing spectral peaks such as these that fall *just below* the predicted masking threshold, but are unmasked by CDT, though they may exist. An automated search for such situations can only by achieved via incorporating CDT detection into a psychoacoustic codec. This task has not been attempted, and is hampered by the fact that the tonal/noise-like discrimination in the two codecs discussed herein does not correctly identify the f_2 components of section 3 as tones, but incorrectly classes them as noise-like signals. Without any automatic system for detecting signals that may benefit from CDT additions to the masking threshold calculation, all that can be stated is that it seems likely that the CDT phenomenon will only be relevant for a very small percentage of audio signals.

Secondly, the temporal response of the distortion tone has not been studied here, but as with all auditory phenomena, the steady state response can only yield an approximate indication of the instantaneous response to a sound.

Thirdly, it should be noted that third order distortion is not confined to the ear. Even high-quality transducers are non-linear devices, and can add a considerable amount of distortion, especially at high amplitude levels. In our tests, the authors found that the CDT produced by sending both primary tones through one loudspeaker could often be greater than the CDT due to the human auditory system. Especially at levels in excess of 80 dB, when L_2 was 10-30 dB lower than L_1, considerable amounts of CDT were audible, well outside the range of audibility predicted by our formulae. It is possible that non-linear equipment, in addition to the non-linear human ear, may unmask certain spectral components, and account for differences that we hear between original and coded audio extracts.

Finally, it may have occurred to the reader that the effect of the CDT on the masking threshold may be more simply modelled by lowering the masking threshold slope to match that implied by the CDT (i.e. the lower boundary of the shaded area in Figure 8). This raises the question: what exactly does the predicted masked threshold measure, and what is the definition of the true masking threshold?

The masking threshold predicted by most audio codecs matches the internal *excitation* due to the masker, shown in Figure 11. [4] provides a full description of the auditory process that gives rise to this excitation pattern, but in simplified terms, the filter-bank within the human auditory system has a sharp cut off

above the target frequency, but a shallower cut off below it, which tends to –40 dB rather than –infinity. Thus lower frequencies leak into the higher frequency bands, and the excitation due to any spectral component will extend to higher frequencies, causing the well-known upward spread of masking. For noise-like signals, this excitation pattern matches the masking threshold, but for tone-like signals, as has been shown, there is a discrepancy between the known excitation pattern, and the known threshold of masking.

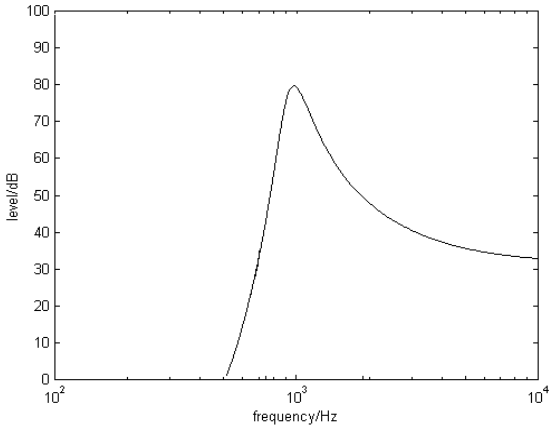

Figure 11 – Excitation within human auditory system due to 1kHz tone.

Another problem in calculating the masking threshold is one of definition. Is a tone masked when the tone itself is inaudible, or when all effects due to the tone are inaudible? There is a difference between these two thresholds, since the CDT (and beats between the masker and the possibly masked tone) is audible even when the masked tone is not. These two thresholds are determined by different experimental conditions: The first by instructing the listener to concentrate on the possibly masked tone; The second by instructing the listener to listen for any difference in the sound produced by the presence of the possibly masked tone. Surely the second type of test is relevant to psychoacoustic codec design, since the aim is to make no audible difference to the signal. However, codecs are often designed using data from the first type of test.

Effect of Beats

Figure 12 shows the human masking thresholds measured using a test of the second type [19], overlaid on the threshold predictions of our codecs, and the region of audible CDTs. It is evident that the codecs are inaccurate by up to 20 dB, and also that the CDT does not account for the whole discrepancy. It must be stressed that, as explained in section 1.3, our calculated CDT may be up to 15dB different from the internal perceived CDT level, and this may account for some of the discrepancy (though this is unlikely, as the discrepancy is largely in the opposing direction – see [14]). However, [19] suggests that the lower threshold is due to beating between the second (masked tone) and the 1 kHz tone, and also to beating between the CDT and the 1 kHz tone. These two sets of beats fall at the same frequency, since

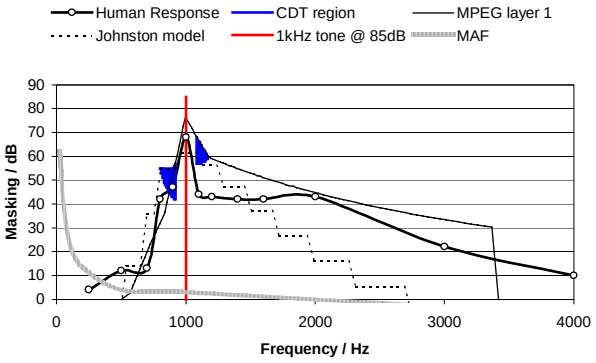

Figure 12 – Masking threshold of human subject.

$$| f_1 - f_2 | = | f_{CDT} - f_1 | \qquad (7)$$

Hence they re-enforce each other, lowering the threshold to the measured values. Methods and equations to predict such interactions are beyond the scope of this paper, but it is suggested that, rather than attempting to fit curves to steady state data, a complete auditory model may be more effective. By modelling the processes within the auditory system that give rise to measurable masking, a more accurate threshold of masking can be calculated than by extrapolating from simple steady-state tone or noise

masking measurements. If this is the distant future of transparent audio coding at ever lower bit-rates, then adapting the masking curves that are built in to existing audio codecs to more closely match the measured data can only be a short-term solution. However, an accurate auditory processing model will be prohibitively computationally burdensome in comparison to existing codecs, and the quality/bit-rate gains may, or may not be significant.

3.2 Masking due to CDT

If two tones create a third (distortion) tone in the human auditory system, this tone will also have its own region of masking. Any spectral components falling below this CDT masking threshold will be inaudible, and hence may be ignored.

An audible CDT will be generated if the two primary spectral components have the same amplitude. However, much of the masking due to this extra distortion tone will coincide with the masking due to f_1. For 1 kHz and 1.2 kHz tones at 85 dB, this will alter the lower masking threshold slope by 2 dB for the Johnston model, and 3 dB for the MPEG-1 psychoacoustic model.

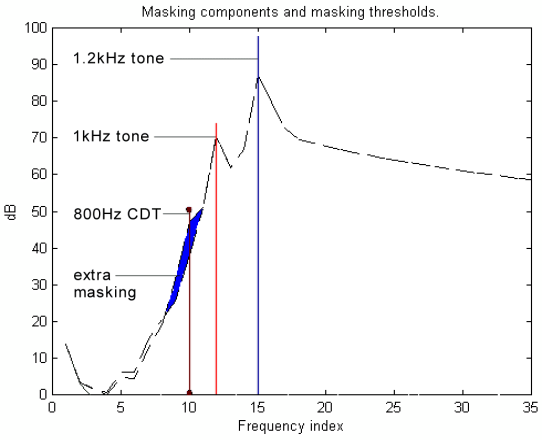

Figure 13 – Extra region of masking (shaded) provided by the CDT.

A more significant effect occurs if the amplitude of f_2 is larger than that of f_1. Consider a 1 kHz tone at 70 dB, and a 1.2 kHz tone at 90 dB. Equations (1)-(5) indicate that these two tones will give a distortion tone of 42 dB at 800 Hz. The MPEG-1 psychoacoustic model is used to predict the masking due to the two primaries, and also the masking due to the two primaries *plus* the CDT. The difference between the two masking threshold curves indicates that the CDT increases the masking threshold around 800 Hz by 15 dB (see Figure 13). As the cancellation-tone measured CDT level accounts for the masking due to the CDT to within 2 dB (see section 1.3), this is a significant result.

Effect of Beats

In the previous section, it was noted that beats also have a role to play in the unmasking of spectral components. However, the beats themselves cannot be used to mask other spectral components. Whereas the CDT causes an excitation within the auditory system at the frequency associated with it, there is no such excitation at the beat frequency, hence no masking will occur[2]. We perceive the CDT by detecting the resulting frequency component in the same manner as we do any external frequency. However, beats are detected by the amplitude modulation that occurs within the auditory filter mid-way between the two primary frequencies, which is generally not tuned to the actual beat frequency. For example, a 1 kHz and 1.1 kHz tone will beat at 100 Hz, but this beating will be detected by the auditory filtered centred at 1.05 kHz; the auditory filter at 100 Hz will not be excited, hence there will be no resultant masking[3].

[2] The amplitude modulation of the beats may hinder the detection of another simultaneous amplitude modulation, but this is beyond the cope of this paper.

[3] If the two frequencies are sufficiently separated, such that no auditory filter detects both, then we fail to perceive any beats.

Real World Applications

In the section 3.2, the rarity of audio signals that may be more accurately processed by taking the CDT into account was discussed. These comments are also true here. However, as the region over which the CDT affects the masking threshold is larger for this second phenomenon, it should find slightly wider use.

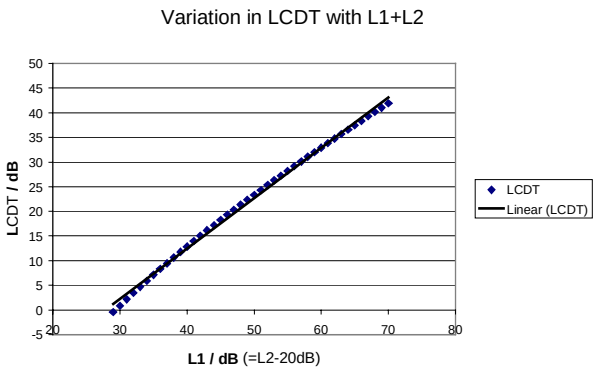

Figure 14 – Showing how L_{CDT} changes linearly as L_1 and L_2 are varied together.

One important feature is that the CDT tone depends on the levels of both f_1 and f_2 in a complex manner, as shown by equation (2). For the CDT to provide a useful masked area in which to hide quantisation distortion, the level of the CDT tone must scale with the level of f_1 and f_2 in a roughly linear manner. Otherwise, decreasing the gain of the replay system may cause the CDT level to decrease more rapidly than the quantisation noise that it is hiding, hence unmasking it.

Figure 14 shows that, as the gain is reduced (i.e. L_1 and L_2 are reduced by equal amounts), the level of the CDT is reduced correspondingly. Thus any quantisation distortion masked by the CDT will not be unmasked as the replay level is altered. Hence, the extra masking produced by the CDT is shown to be a useful region in which to hide quantisation noise.

4 Conclusion

In this paper, it as been shown that distortion tones generated by non-linearities within the human auditory system may affect the masking thresholds of pure tones. The cubic distortion tone (CDT) was identified as the most audible distortion product, and formulae were presented to calculate its frequency and amplitude.

The masking threshold predictions of two auditory models were shown to be misleading. Over a small frequency range, tones that were up to 15 dB below the predicted masking threshold were found to generate audible CDTs. It was suggested that a true definition of a masked tone is that its presence makes no audible difference to the signal; Thus the CDT can be said to unmask tones by up to 15 dB, though the level of CDTs is dependent on measurement method. The presence of beats may unmask tones by even larger amounts.

It has also been shown that the presence of the CDT can raise masked thresholds around the CDT frequency by up to 15 dB. This masking is independent of replay level, so may be utilised by an audio codec to conceal quantisation distortion.

The overall effect of the cubic distortion tone on masking thresholds is found to be small. Both effects discussed here are most prominent for tonal signals, and so may find a limited application in audio coding. It is suggested that, where perfectly transparent coding is required at the minimum possible bitrate, a model of the non-linear processing within the human auditory system may be used to accurately predict masking thresholds. Such an approach will be computationally burdensome, and may or may not yield an improved quality/bit-rate ratio. In the interim, prediction of the CDT using the methods outlined in this paper may be applied to existing audio codecs.

REFERENCES

[1] ISO/IEC. "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s –". Part 3: Audio. ISO/IEC 11172-3 International Standard, 1993.

[2] J. E. Helmholtz, "Die Lehre von den Tonempfindungen, als physiologische Grundlage für die Theorie der Musik," *Dritte Ausgabe* (von Friedrich Vieweg und Sohn, Braunschweig, 1870).

[3] R. Probst, B. L. Lonsbury-Martin, and G. K. Martin, "A review of otoacoustic emissions," *J. Acoust. Soc. Am.*, Vol. 89, pp. 2027-2067 (1991 May).

[4] D. J. M. Robinson and M. J. Hawksford, "Time-Domain Auditory Model for the Assessment of High-quality Coded Audio," presented at the 107th Convention of the Audio Engineering Society, New York, (1999 Sept.), preprint 5071.

[5] G. K. Yates, "Cochlear Structure and Function" in *Hearing*, B. C. J. Moore (Ed), (Academic Press, San Diego, 1995).

[6] D. T. Kemp, "Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea," *Arch. Otorhinolaryngol*, vol. 224, pp. 37-45 (1979).

[7] G. R. Popelka, P. A. Osterhammel, L. H. Nielsen, and A. N. Rasmussen, "Growth of distortion product otoacoustic emission with primary-tone level in humans," *Hear. Res.*, vol. 71, pp. 12-22 (1993).

[8] D. T. Kemp, "Towards a model for the origin of Cochlea echoes," *Hear. Res.*, vol. 2, pp. 533-548 (1980).

[9] G. F. Smoorenburg, "Combination tones and their origin," *J. Acoust. Soc. Am.*, vol. 52, pp. 615-632 (1972).

[10] J. L. Goldstein, "Auditory Nonlinearity," *J. Acoust. Soc. Am.*, vol. 41, pp. 676-689 (1967).

[11] J. L. Hall, "Auditory distortion products f2-f1 and 2f1-f2," *J. Acoust. Soc. Am.*, vol. 51, pp. 1863-1871 (1972).

[12] E. Zwicker, and H. Fastl, "Cubic difference sounds measured by threshold- and compensation-method," *Acustica*, vol. 29, pp. 336-343 (1973).

[13] E. Zwicker, "Formulae for calculating the psychoacoustical excitation level of aural difference tones measured by the cancellation method," *J. Acoust. Soc. Am.*, vol. 69, pp. 1410-1413 (1981 May).

[14] C. Giguère, G. F. Smoorenburg, and H. Kunov, "The generation of psychoacoustic combination tones in relation to two-tone suppression effects in a computational model," *J. Acoust. Soc. Am.*, vol. 105, pp. 2821-2830 (1997 Nov.).

[15] E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523-1525, (1980 Nov.).

[16] H. Traunmüller, "Analytical expressions for the tonotopic sensory scale," *J. Acoust. Soc. Am.*, vol. 88, pp. 97-100 (1990 July).

[17] E. Zwicker, "Different behavior of quadratic and cubic difference tones," *Hear. Res.*, vol. 1, pp. 283-292 (1979).

[18] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 314-323 (1988 Feb.).

[19] B. C. J. Moore, J. I. Alcántara, and T. Dau, “Masking patterns for sinusoidal and narrow-band noise maskers,” *J. Acoust. Soc. Am.*, vol. 104, pp. 1023-1038 (1998 Aug.).