Time-domain auditory model for the assessment of high-quality coded audio

David J M Robinson & Malcolm J Hawksford

Centre for Audio Research and Engineering
Department of Electronic Systems Engineering
University of Essex
Wivenhoe Park
Colchester CO4 3SQ
UK

Email: d.j.m.robinson@essex.ac.uk

A time-domain auditory model is described. The spectral analysis function of the inner ear is simulated by a non-linear filter bank, while the temporal response is simulated by complex filtering of the envelope. The spectral and temporal masking properties of the model are calibrated with known data. The effectiveness of this model in predicting the perceived quality of coded audio streams is examined.

0 Introduction

In this paper, we describe an auditory model for assessing the perceived quality of coded audio signals. This model simulates the functionality of the physiology found within the human ear. Modelling the processes within the ear, rather than simulating the effects that arise from these processes, may yield a more accurate prediction of human perception.

0.1 Current methods of assessing perceived audio quality

State-of-the-art audio codecs reduce the amount of data needed to represent an audio signal by discarding components that may be inaudible to human listeners. This process is highly non-linear, and traditional performance metrics, such as frequency response, or signal to noise ratio, cannot quantify the perceived audio quality of the resulting signal.

Typically, the quality is assessed via a subjective test, where the opinions of a number of human listeners are canvassed (e.g. [1]). This process is both expensive and time consuming, and an alternative is sort. Ideally, we would like an objective measurement that accurately reflects the perceived sound quality of the system under test, effectively an "electronic ear" and "electronic brain" which can yield the same results as human subjects in a "listening" test.

Two measurement algorithms for objectively assessing the quality of audio codecs have been enshrined in international standards: the PSQM algorithm for the assessment of speech codecs (see ITU-T P.861, [2]); and the PEAQ algorithm for assessing high-quality wide-band audio codecs (see ITU-R BS.1387, [3]). (This latter standard draws on previous models such as [4] and [5]). A complete measurement system incorporating these algorithms, which can yield an objective measurement of perceived sound quality under many circumstances, is described in [6].

0.2 Beyond current standards

At the present time, the perceptual models incorporated into state-of-the-art audio codecs are less advanced than those within the PEAQ measurement algorithm. That is to say, the representation of the human auditory system within the measuring device is more accurate than that within the codec itself. While this is the case, the algorithm will yield a correct indication of perceived audio quality. However, as the perceptual models incorporated into audio codecs become more advanced, existing measurement algorithms may fail. Thus, there is a need to develop more advanced algorithms, incorporating more accurate perceptual models, for future use.

A perceptual model aims to simulate human perception. The task of the perceptual model within any audio assessment algorithm is to simulate the human auditory system. There are many approaches to this task. They range from modelling the coarse *effects* of the auditory system, to modelling the actual *processing* that occurs at a neural level. The former gives a poor approximation to human perception, while the latter is computationally burdensome, and yields such vast quantities of data that any further processing (e.g. the prediction of perceived sound quality) is very complex.

The first aim of this paper is to suggest a model that simulates the *processes* present within the human auditory system, but on a macro-, rather than micro-scale. The *effects* of the auditory system (e.g. spectral and temporal masking) which are so important in any measurement algorithm, are found to arise "naturally" from the simulation of these processes. The second aim of this paper is to discover if such a model is appropriate for the quality-assessment of coded audio.

1 The perceptual model

As our aim is to simulate the processing carried out within the human auditory system, we will commence by examining the actual processing found within the human ear.

1.1 The Human Auditory System

Figure 1 shows the main components of the human auditory system. The upper illustrations represent the physiology, while the lower graphs indicate the functionality of each section.

[The hair cell illustration is adapted from [7]. All frequency domain plots show amplitude in dB against log frequency. All time domain plots are linear on both scales.]

The function of each section is as follows:

- The **pinna** directionally filters incoming sound. The resulting spectral coloration of the incoming sound (called the Head Related Transfer function, or HRTF) enables human listeners to localise the sound source in 3-dimensions. Measurements of human HRTFs are given in [8].
- The **ear canal** filters the sound, attenuating low and high frequencies, giving a resonance at around 5 kHz. The **timpanic membrane** (ear drum), **malleus** and **incus** transmit the sound pressure wave into the cochlea.
- The fluid-filled **cochlea** is a coil within the ear, partially protected by bone. It contains the **basilar membrane**, and **hair cells**, responsible for the transduction of the sound pressure wave into neural signals.
- The **basilar membrane** (BM) semi-partitions the **cochlea**, and acts as a spectrum analyser, spatially decomposing the signal into frequency components. Each point on the BM resonates at a different frequency, and the spacing of resonant frequencies along the BM is nearly logarithmic. The effective frequency selectivity is governed by the width of the filter characteristic at each point.
- The **outer hair cells** are distributed along the length of the BM**.** They react to feedback from the brainstem, altering their length to change the resonant properties of the BM. This causes the frequency response of the BM to be amplitude dependent.
- The **inner hair cells** fire when the BM moves upwards, so transducing the sound wave at each point into a signal on the auditory nerve. In this way the signal is effectively half wave rectified. Each cell needs a certain time to recover between firings, so the average response during a steady tone is lower than that at its onset. Thus, the inner hair cells act as an automatic gain control. The firing of any individual cell is pseudo-random, modulated by the movement of the BM. However, in combination, signals from large groups of cells can give an accurate indication as to the motion of the BM.

The net result so far is to take an audio signal, which has a relatively wide-bandwidth, and large dynamic range, and to encode it for transmission along nerves which each offer a much narrower bandwidth, and limited dynamic range.

The function of each individual stage of the subsequent neural processing is less well understood (see Figure 1). The Cochlea Nucleus is thought to sharpen features of the (now highly compressed) signal. The superior Olivary Complex is responsible for lateralisation of sound sources. Little is known of the following stages of neural processing, other than their existence, and the fact that they give rise to our human “understanding” of the sounds around us as speech, music, and noise.

A critical factor is that any information lost due to the transduction process within the cochlea is not available to the brain – the cochlea is effectively a lossy coder. The vast majority of what we *cannot* hear is attributable to this transduction process. Predicting the signal present at this point should give a good indication of what we can and cannot hear.

1.2 The Structure of the model

The model presented here is based upon the processing present within the human auditory system, as described in section 1.1. The structure of the auditory model is shown in Figure 2. Each individual component is described in the following sections.

1.2.1 Pre-filtering

The filtering of the pinna and ear canal is simulated by an FIR filter, derived from measurements made using a KEMAR dummy head. An arbitrary angle of incidence is chosen, in this case 30 ^{o}. The KEMAR measurements were used because they were readily available for this research. Measurements from human subjects could be used as a more accurate and realistic alternative ([8]).

1.2.2 Basilar membrane filtering

A bank of amplitude dependent filters simulates the response of the BM. Each filter is an FIR implementation of the gammachirp, described in [9], and simulates the response of the BM at a given point.

The formula for the time domain (impulse) response of the gammachirp filter is

$$g_c(t) = at^{n-1} \exp(-2\pi b \,\mathrm{ERB}(f_r)t) \cos(2\pi f_r t + c\ln(t) + \phi) \qquad (1)$$

Where ERB is the equivalent rectangular bandwidth of the filter, given by

$$\mathrm{ERB}(f_r) = 24.7 + 0.108 f_r \qquad (2)$$

In [9], data from four separate studies on the shape of the human auditory filter is used to calculate values for *a, b, c* and *n*. The chosen values for *b* and *n* are 1.14 and 4 respectively. The term *a* determines the overall amplitude of the filter. In the model, *a* is set individually for each filter, such that the centre frequency is not attenuated. The term *c* ln(*t*) causes the amplitude dependency, where *c* is proportional to the amplitude of the signal in each band (see section 1.2.2.2 for more details).

1.2.2.1 Spacing of the filters

The filters are spaced linearly on the Bark frequency scale, or critical-band rate scale. This scale correlates with the spacing of resonant frequencies along the BM. The critical band number *z* (in Bark) is related to the linear frequency *f*, thus

$$z = [26.81 f / (1960 + f)] - 0.53 \tag{3}$$

with the correction that for calculated z>20.1, the actual Bark, *z'* is given by

$$z' = z + 0.22(z - 20.1)\,. \tag{4}$$

This definition of the Bark scale is taken from [10].

It is debatable as to what spacing (and hence number) of filters are needed to match the frequency resolution of the human ear. Humans can differentiate between about 620 frequencies, equally spaced in the bark domain [11]. However, it does not follow that there are 620 differentiable points along the BM. At lower frequencies, the *firing rate* of the inner hair cells will represent the frequency, irrespective of the resolution of the BM. Hence we do not need to simulate 620 points on the BM via a bank of 620 gammachirp filters. Instead, we choose a spacing such that all frequencies can be transduced without significant inter-filter gaps.

Choosing a spacing of ½ bark causes all frequencies between 200 Hz – 16 kHz to be within 3 dB of the resonant peak of a filter (See Figure 3(a)). Decreasing the filter spacing to ¼ bark causes *all* audible frequencies to be within 1 dB of the resonant peak of a filter (See Figure 3(b)). At this spacing 96 filters are needed to cover the full audible range.

During the testing and calibration of the model, only a single gammachirp filter was used. For each task, this filter was centred on the particular frequency of interest.

1.2.2.2 Amplitude dependence

Figure 4 illustrates the mechanism used to simulate the amplitude dependent nature of the auditory filters. The output of each gammachirp filter is processed to provide the amplitude information necessary to tune the response of that individual filter. This feedback causes the Q of each filter to be proportional to the amplitude of the signal passing through it. This corresponds to the tuning mechanism mediated by the outer hair cells, and the mechanical properties of the BM itself.

The process is as follows:

1. Rectification
2. Peak detect and hold
3. Low pass filtering

The “peak detect and hold” stage yields the envelope of the signal. It is necessary to low pass filter this envelope, because changing the response of the filter in direct response to the envelope of the signal causes the system to be unstable. Empirically, a time constant of 1 ms was found to be sufficient to stabilise the system. It would be desirable to match this time constant to that of the actual process within the auditory system, but this data is unknown at the present time.

The value of *c* in equation (1) is calculated from the low-pass filtered envelope, *env(t)*, thus:

$$c(t) = 3.29 + 0.059 * 20\log_{10}(env(t) + ina) \tag{5}$$

The values in this equation were obtained by fitting the filter shapes generated by equation (1), to those measured using human subjects in [12], for various values of *c*. *ina* is present to prevent the calculation of log(0) during a silent input signal. The value of *ina* is below the minimum audible threshold, and has a negligible effect on supra-absolute-threshold calculations.

1.2.3 Hair cell transduction

At each point along the basilar membrane, its movement is transduced by a number of hair cells – see Figure 1 (c). In the model, this will be simulated by appropriate processing of the output of each gammachirp filter. All subsequent processing will take place in parallel, in as many bands as there are gammachirp filters (see section 1.2.1.1).

As noted, each individual hair cell yields little information about the incoming sound wave – it is only by examining the firing rates of hundreds of hair cells that a complete picture may be constructed. The hair cell model detailed in [13-15] simulates the random firing of individual cells, and the probability of each cell firing. The matter is complicated by the existence of at least three categories of cells, each operating over a different dynamic range, and each present at every point along the BM. An accurate functional simulation would require hundreds of simulated hair cells at the output of each gammachirp filter. The information from all these hair cells must then be re-combined into a single detection probability. This is a computationally burdensome task, and a simpler model of hair cell action is sought.

1.2.3.1 Adaptation

It is suggested in [16] that a simpler model may be just as effective. The output of each gammachirp filter is half wave rectified, then low pass filtered at 1 kHz (see Figure 5). This partially simulates the response of the inner hair cells, but without taking account of their increased sensitivity to the onset of sounds. This change in sensitivity can be viewed as a type of adaptation. The next stage is to simulate this adaptation by a cascade of 5 feedback loops, each with a different time constant. The adaptation of the auditory system is not only due to the response of the inner hair cells – feedback to the outer hair cells also accounts for some of the adaptation found in our hearing. The overall measured adaptation of the human auditory system is well matched by the adaptation model shown in Figure 5 (from [17]). The time constants used are 5, 50, 129, 253, and 500 ms.

1.2.3.2 Internal Noise

The random firing of the inner hair cells, combined with the blood flow within the ear, gives rise to an internal noise that limits our absolute hearing threshold. The model as it stands contains no such internal noise. To account for this, an absolute threshold is fixed into the model at this point, such that signals with a level below the threshold of hearing are replaced by the threshold value. This internal value is calculated from the MAF figures in [18], processed through the stages of the model. This is only an approximation to real behaviour, and a more realistic mechanism would be needed to accurately model our response to signals near the absolute threshold of hearing.

1.3 Perceiving a difference

So far, we have simulated the passage of sound through the ear canal and cochlea – see Figure 1(a-c). The output of the auditory model is a signal analogous to that transmitted along the auditory nerve. If two signals are processed independently by the auditory model, the difference between the two resulting outputs will be related to the perceived difference between the two signals. This concept is illustrated in Figure 6. The next task is to determine the processing that is necessary to yield an accurate indication of perceived difference.

The outputs of the model are *n* time varying signals, where *n* is the number of gammachirp filters, as discussed in section 1.2.2.1. Hence forth, *n* will be referred to as the number of bands. Calculating the perceived difference will involve comparing the signals in each band, and determining at what level these differences would become audible to a human listener. In the following sections, the value of this internal threshold level will be set, such that the "perception" of the model matches that of a real human listener.

[It is important to note the difference between the *internal threshold level*, set in the model, and the *threshold condition of a psychoacoustic test*. The former is a single number, which will determine when the model will flag an "audible" difference. The latter is the measured level, usually in dB, at which a real human listener is just able to perceive some sound in a particular situation (e.g., in noise, in the presence of another similar sound, etc.). The latter is different depending on the task we are discussing. For example, the threshold condition is 0 dB SPL for detecting a 2 kHz tone in silence, but 80 dB SPL for detecting a 2 kHz tone in the presence of 65 dB SL noise. However, if we compare two signals which are on the threshold of being audibly different from each other (e.g. the 65 dB noise in isolation, with the 65 dB noise + 80 dB tone), then the perceived difference calculated by the model should be the internal threshold level.]

The first step is to determine how the perceived difference should be calculated. If calculated incorrectly (i.e. in a manner that does not reflect the workings of the auditory system), then it will be impossible to set a single internal threshold value across a variety of tests.

1.3.1 Calculating the perceived difference by simple subtraction

One possible method of calculating the perceived difference is to simply take the difference between the two sets of outputs from the auditory model (one set for each signal under test). This will yield a time-varying calculated perceived difference (CPD) signal for each auditory band.

This attractively simple method of determining the CPD fails to match human perception. This fact can be demonstrated by using the model to simulate some tests with have previously been carried out by human listeners, and comparing the results.

1.3.1.1 Testing the validity of this approach

The threshold of perceiving a difference between two signals is known for a wide variety of possible signals, from various psychoacoustic experiments using human subjects. If the difference between two signals is greater than this threshold value, then a human listener will perceive a difference. If it is less, then the difference, though present, will be imperceptible.

The simplest differentiation experiment is that of discriminating between the level of two tones. A human listener can detect a 1 dB level difference between two tones of the same frequency, but any smaller change in level is undetectable. If, in Figure 6, *signal A* is a

60 dB sine wave, and *signal B* is a 61 dB sine wave, then the calculated perceived difference (CPD) represents a just detectable difference, and allows us to set the internal threshold CPD. If the CPD fails to reach this value during any subsequent test, then the difference can be said to be imperceptible, whilst if the CPD exceeds this value, the model has “perceived” a difference.

As our model claims to represent part of the human auditory system, it might be expected that this threshold CPD value would be obtained at the threshold condition of *any* difference detection task. However, this is not the case.

Simulating the tone masking noise experiment detailed in [19], *signal A* is the tone in isolation, and *signal B* is the tone plus noise at the threshold (just audible) level. The resulting CPD peaks at twice that obtained in the previous experiment. However, the mean CPD over 1 second in both experiments is almost identical. This indicates that some form of temporal averaging is needed.

Simulating the temporal masking experiment detailed in [20], *signal A* is a burst of noise, and *signal B* is the same burst of noise followed by a short tone at the threshold level. The resulting CPD peaks at a level similar to the first experiment, but the mean CPD over 1 second is much lower than the previous two, because the difference is isolated to within a few milliseconds.

In a fourth test condition, *signal A* and *signal B* are two random noise sequences. These are perceptually identical, but have different waveforms. The resulting CPD is much greater (by an order of magnitude) than the CPD at threshold in any other experiment.

Thus, we see that the CPD calculated by simple subtraction is not a good indicator of human perception. Modelling the auditory periphery alone is insufficient to account for what humans can and cannot hear, and a further stage of processing is needed.

1.3.2 Calculating the perceived difference by integration

In our auditory model, we have modelled the functionality of the actual physiology found within the human ear. Modelling the subsequent cognitive process in this way is beyond the scope of this research (though others have taken this path; see [3]). However, the discrepancies described in section 1.3.1.1 can be dealt with without a complex cognitive model.

Rather than looking at the CPD on a sample by sample basis, and declaring an audible difference whenever the CPD peaks above a threshold value, the CPD can be summed over time. Then, a threshold *area* can be defined, and whenever this area is exceeded within a certain time interval, a difference will be perceived. This will make the CPD threshold consistent between the first two experiments detailed in section 1.3.1.1.

If this mechanism alone were to compensate for the misleadingly high CPD resulting from two perceptually identical noise signals, the time constant would need to be in excess of 200 ms. This would preclude the detection of any temporal masking effects, and a time constant of 20 ms is found to be more realistic. Therefore, a second mechanism is needed to account for our inability to differentiate one noise signal from another.

1.3.3 Adjusting the perceived difference according to the variance of the signal

The discrepancy in the CPD figure obtained from two sequences of random noise is due to the signal in any given band varying dramatically from moment to moment. The listener would need a perfect auditory memory to record each moment, and compare it with the corresponding moment in the second noise signal. This is exactly what the model is doing, hence it predicts perceived differences, but this is beyond the capabilities of the human auditory system. In effect, the more complex the signal, and the more the signal is varying in any given band, the less sensitive we will be to any differences.

To simulate this, the CPD is scaled by the inverse of the variance of the signal in each band. The actual process is shown in Figure 7, as follows:

1. The variance of the output of the auditory model is calculated for *each signal* over a 20 ms period
2. The lower of the two values is chosen
3. Brief peaks in the variance are suppressed
4. The variance signal is low pass filtered with a time constant of 1.4 ms
5. This signal is summed over 100 ms
6. The CPD is scaled by 1/(1+30*the resulting figure)

The new CPD should indicate whether a human listener would perceive any difference between the two signals, by reference to a single known CPD threshold value, which represents human threshold in any task. Also, larger CPD values should indicate that a human listener would perceive a greater difference between the two signals. We will test the first hypothesis in the following section.

2 Validation of the model

2.1 Psychoacoustic tests

The following series of psychoacoustic tests were simulated via the model.

Experiment	source	threshold condition	Fig.	CPD	below	above
Tone masking noise (simultaneous)	[19]	1 k tone @ 81 dB 80 Hz wide 1k noise @ 85dB	8	25.2	18.5	31.5
Noise masking tone (post-masking)	[20]	200 ms, 80 dB white noise 50 ms delay, 31 dB 2 k tone	9	26.3	20.1	33.8
Level differentiation	[16]	1 dB level difference	10	25.7	5.1	78.0
Random Noise		--	11		20.5	28.2

Table 1 – details of psychoacoustic tests simulated via the model at threshold

Experiment	name of the psychoacoustic test simulated
Source	reference to the results of that experiment on human subjects
Threshold Cond.	actual threshold simulated via model
Fig.	Figure number showing time varying CPD value for threshold condition
CPD	peak calculated perceived difference at threshold condition in target band
Below	peak CPD obtained at 3 dB below threshold (for masking expts.)
Above	peak CPD obtained at 3 dB above threshold (for masking expts.)

In the level differentiation test, the **below** threshold condition consisted of a level difference of 0.2 dB, whilst the **above** threshold condition consisted of a level difference of 3 dB.

In the random noise test, the **below** threshold condition consisted of two sequences of perceptually identical random noise, with non-identical waveforms. For the **above** threshold condition, the level of one signal was raised by 3 dB.

Examining the results in Table 1, we see that for each of the simulated tests, the “CPD at threshold” is in close agreement. This shows that, in determining a “just detectable” difference, the CPD correlates well with human perception. The CPD between the two noise signals is also in accordance with human perception.

Thus, the model is shown to accurately predict human perception in a range of psychoacoustic tests.

2.2 Codec assessment

The motivation behind developing a model of human perception is to create an objective method of assessing audio codecs. It has been shown that the model can predict the threshold level in a range of psychoacoustic listening tests, but can it perform equally well with real world audio signals? In particular;

1. Can the model accurately determine if the difference between two complex audio signals is perceptible?
2. Can the model quantify *how* perceptible this difference will be?

In assessing an audio codec, the first issue is analogous to determining the transparent bit-rate – the level of data-reduction which, for a given coding scheme, produces an audio signal that is perceptually identical to the original. The second issue relates to determining how “bad” a codec sounds when its performance is worse than “transparent”.

2.2.1 MPEG-1 layer-2 codec assessment

To test how accurately the model addresses these issues, we will use the model to assess the quality of some audio streams generated by an MPEG-1 layer-2 codec, operating at a variety of bit-rates. It is known from previous subjective tests that the performance of this codec is “good” (1 on the diff-grade scale [21]) when operating at 160 kbps, and near “transparent” (0 on the diff-grade scale) at 192 kbps. Below 160 kbps, the performance of this codec becomes audibly worse, whilst above 192 kbps the codec is perceptually transparent.

The audio sample is taken from LINN CD AKD 028, Track 4, 0:27-0:29, a piece of vocal jazz. This extract was chosen because it was quite difficult to code. A more complete “listening” test, whether with real human subjects, or a perceptual model, would include many diverse extracts. However, for our purposes, one (difficult to code) example will be sufficient.

The extract was coded at a variety of bit rates from 112 kbps to 256 kbps. The MPEG streams were decoded, then time aligned with the original by cross-correlation.

Each coded stream was analysed by the model (Figure 7), and compared with the original. The result is a time varying CPD output from each band (see Figure 12). Table two lists the largest CPD (across all bands) during each extract, and also the number of bands in which the CPD exceeded the "perceptible" threshold value.

Bit-rate	Peak CPD (band)	Number of bands in which peak CPD>25	Can a human hear any difference?
256	13.8 (14)	0	None
192	16.0 (9)	0	None
160	26.6 (19)	2	Slight
128	46.6 (19)	9	Much
112	63.0 (19)	11	More

Table 2 – MPEG 1 layer 2 assessment test

Bit-rate the kilo-bit per second at which the extract was coded
Peak CPD the highest CPD during the extract in any band
(Band) the band number in which the above occurred
No. bands CPD>25 how many bands a difference was perceived in
Hear any difference in a real listening test, how great a difference can a human subject perceive between the particular bit-rate, and the original extract.

The results in Table 2 indicate that, in this particular test, the model predicted human perception very well. At the two bit-rates where a human listener perceives no difference between the original and coded signals, the model does likewise. At 160 kbs, where the human perception is of a signal that *is* perceivably different from the original, but that difference is not annoying, the model predicts that a difference is just perceptible (a peak CPD of 26.6 compared to a threshold of 25). At the lower bit-rates, the model predicts that the difference between the original and coded signals will be well above threshold. The human perception is that the difference is very audible, hence again the model accurately predicts human perception.

2.2.2 Other codec assessments

Similar tests were performed to assess the performance of other codecs via the model, and to compare the results given by the model, with those given by human listeners.

MPEG-1 layer-3 coded audio was successfully assessed by the model in accordance with human perception.

The model incorrectly assessed the Microsoft audio codec, which maintains a better frequency response, at the expense of poorer temporal resolution. The model predicted an audible difference where human listeners could hear none. Also, where there was an audible difference, the severity of the difference was greatly overestimated by the model in comparison to the human perception of the difference. This affected comparisons with the MPEG codecs, which were themselves correctly assessed. The model indicated that MPEG-1 layer-2 compression at 128 kbps sounded better than the Microsoft audio codec at 64 kbps, but human listeners perceived the reverse. Though both are audibly far from transparent, the temporal smearing of the Microsoft audio codec was preferred by all listeners compared to the frequency “drop outs” associated with the MPEG-1 layer-2 codec.

The problem seems to lie in the models emphasis on the onset of sounds. This feature is found within the human auditory system, and to the same extent. However, it seems that some later process must suppress these onsets in certain situations, to account for our tolerance of the temporal-smearing distortion encountered in the final example. This will be the subject of further research.

3 Conclusion

An auditory model has been described that simulates the processes found within the human auditory system. The output of this model is analysed to detect perceptible differences between two audio signals. The model was found to correctly predict human perception in a range of psychoacoustic tests. The perception of a range of coded audio extracts was also correctly predicted by the model. Finally, the model was shown to be over-sensitive to temporal errors in the input signal, which are inaudible to human listeners due to pre-masking.

References

[1] G. A. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, “Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs,” *J. Audio Eng. Soc.*, vol. 46, pp. 164-177 (1998 Mar.).

[2] ITU-T Rec. P.861, “Objective Quality Measurement of telephone-band (300-3400 Hz) speech codecs,” International Telecommunication Union, Geneva, Switzerland (1996).

[3] ITU-R Rec. BS.1387, “Method for Objective Measurements of Perceived Audio Quality (PEAQ),” International Telecommunication Union, Geneva, Switzerland (1998).

[4] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne, “PERCEVAL: Perceptual Evaluation of the quality of Audio Signals,” *J. Audio Eng. Soc.*, vol. 40, pp. 21-31 (1992 Jan.).

[5] C. Colomes, M. Lever, J. B. Rault, and Y. F. Dehery, “A Perceptual Model Applied to Audio Bit-Rate Reduction,” *J. Audio Eng. Soc.*, vol. 43, pp. 233-240 (1995 Apr.).

[6] M. Keyhl, C. Schmidmer, and H. Wachter, “A combined measurement tool for the objective, perceptual based evaluation of compressed speech and audio signals,” presented at the 106th Convention of the Audio Engineering Society, preprint 4931.

[7] G. K. Yates, “Cochlea Structure and Function,” in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California, 1995), pp. 41-74

[8] H. Moller, M. F. Sorensen, D. Hammershoi, and C. B. Jension, “Head-Related Transfer-functions of Human-Subjects,” *J. Acoust. Soc. Am.*, vol 43, no. 5, pp.300-321 (1995).

[9] T. Irino and D Patterson, “A time-domain, level-dependent auditory filter: The gammachrip,” *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 412-419 (1997 Jan.).

[10] H. Traunmüller, “Analytical expressions for the tonotopic sensory scale,” *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 97-100 (1990 July).

[11] B. C . J. Moore, *An Introduction of the Psychology of Hearing,* 4th ed. Academic Press, New York, 1997).

[12] B. C. J. Moore, “Frequency Analysis and Masking,” in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California, 1995), pp. 161-205.

[13] R. Meddis, “Simulation of mechanical to neural transduction in the auditory receptor,” *J. Acoust. Soc. Am.*, vol. 79, no. 3, pp. 702-711 (1986 Mar.).

[14] R. Meddis, “Simulation of auditory-neural transduction: Further studies,” *J. Acoust. Soc. Am.*, vol. 83, no. 3, pp. 1056-1063 (1988 Mar.).

[15] R. Meddis, M. J. Hewitt, and T. M. Shackleton, “Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1813-1816 (1990 Apr.).

[16] T. Dau, D. Püschel, A. Kohlrausch, “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure,” *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615-3622, (1996).

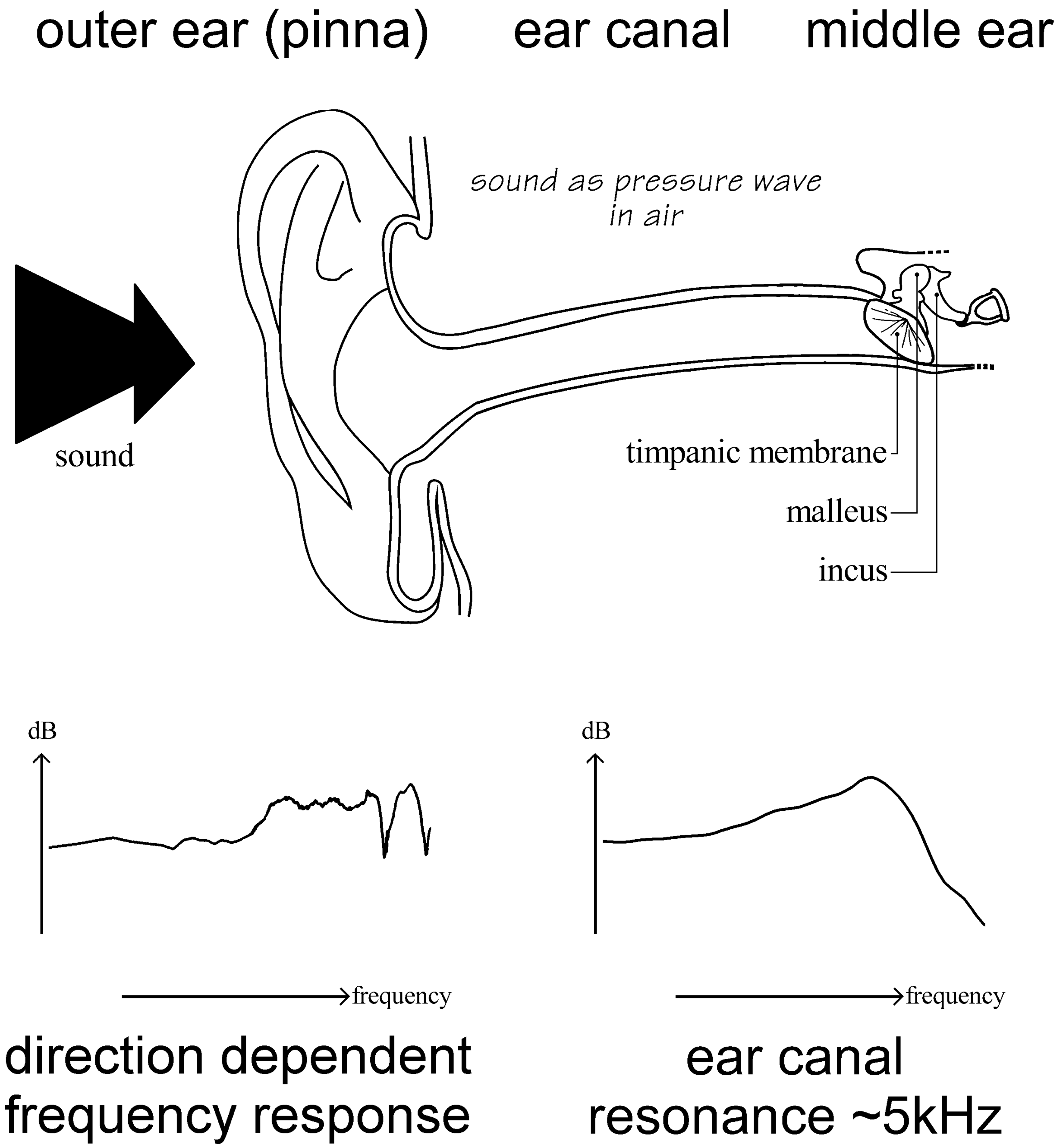
[17] D. Püschel, “Prinzipien der zeitlichen Analyse beim Hören,” Ph.D. thesis, University of Göttingen (1988).

[18] D. W. Robinson and R. S. Dadson, “A re-determination of the equal-loudness relations for pure tones,” *British Journal of Applied Physics*, vol. 7, no. 5, pp. 166-177 (1956 May).

[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, no. 2.1, pp. 1023-1038 (1998 Aug.).

[20] E. Zwicker, “Dependence of post-masking on masker duration and its relation to temporal effects in loudness,” *J. Acoust. Soc. Am.*, vol. 75, no. 1, pp. 219-223 (1984 Jan.).

[21] K. Brandenburg and M Bosi, “Overview of MPEG Audio: Current and Future Standards for Low-Bit-Rate Audio Coding,” *J. Audio Eng. Soc.*, vol. 45, no. 1/2, pp. 4-21 (1997 Jan./Feb.).



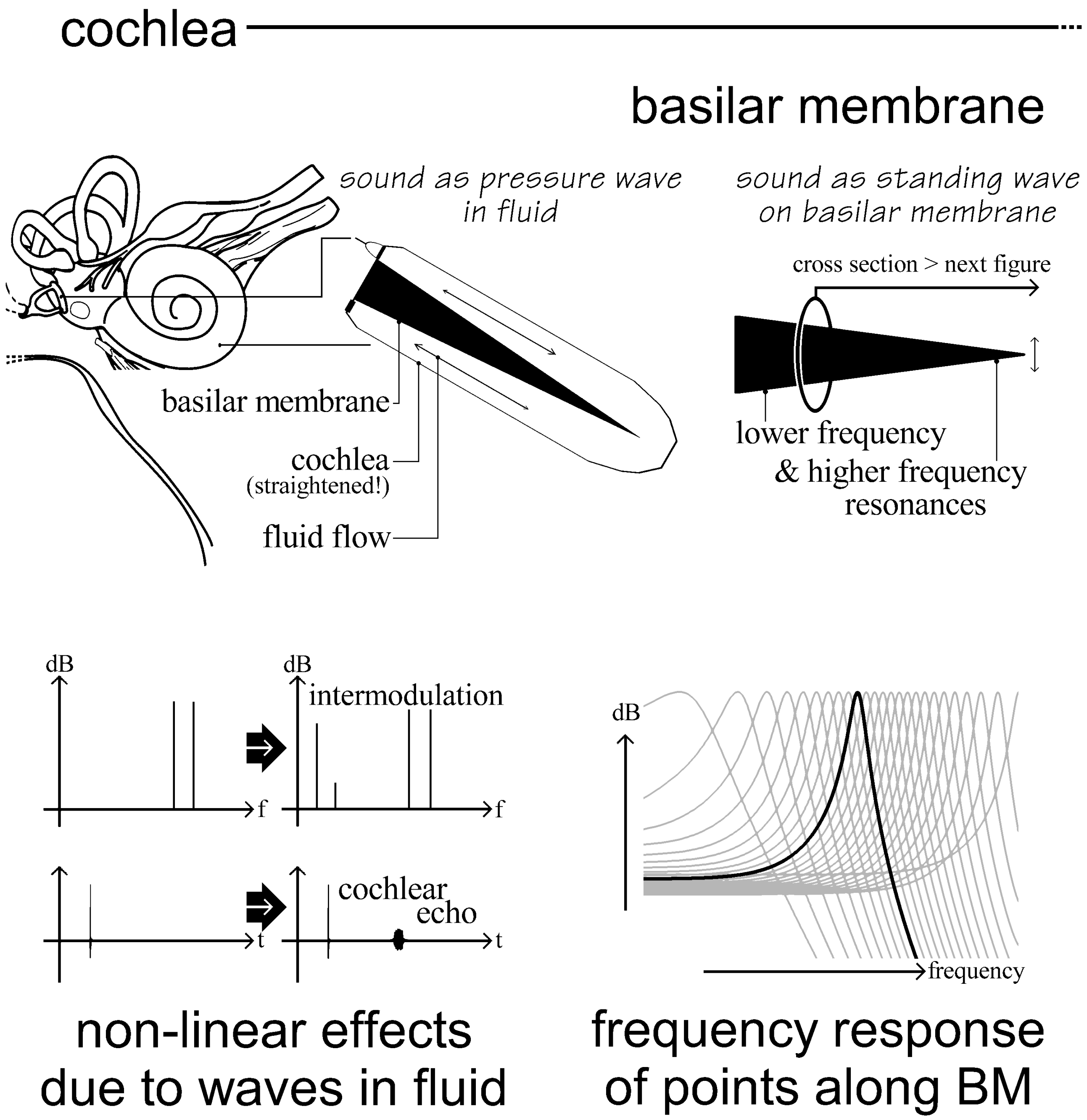
outer ear (pinna)
ear canal
middle ear
sound as pressure wave in air
sound
timpanic membrane
malleus
incus
dB
frequency
dB
frequency
direction dependent frequency response
ear canal resonance ~5kHz

Figure 1 (a) Signal path through the human auditory system

- from free field to middle ear

Upper half: physiology. Lower half: functionality

Sound waves incident from different angular positions are spectrally shaped by the pinna in a direction dependent manner. The ear canal further filters the waveform, before it passes through two small bones, and on to the cochlea.



cochlea

basilar membrane

non-linear effects due to waves in fluid

frequency response of points along BM

Figure 1 (b) Signal path through the human auditory system

- within the cochlea: the basilar membrane (BM)

Upper half: physiology. Lower half: functionality

Sound waves enter the cochlea and set the fluid within in motion. The cochlea is partially partitioned by the BM, different points of which resonate at different frequencies. Thus the BM acts as a spectrum analyser.

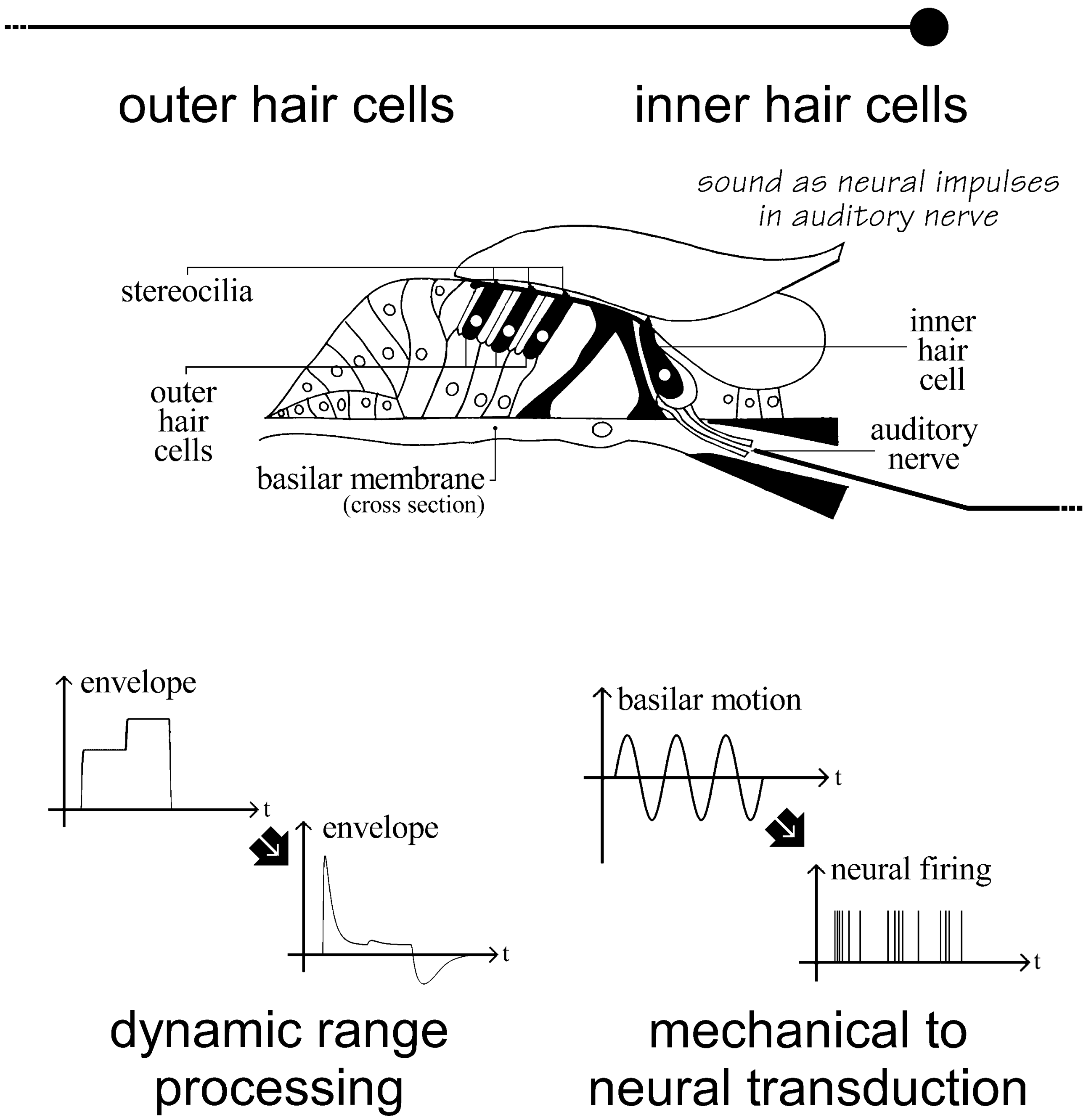


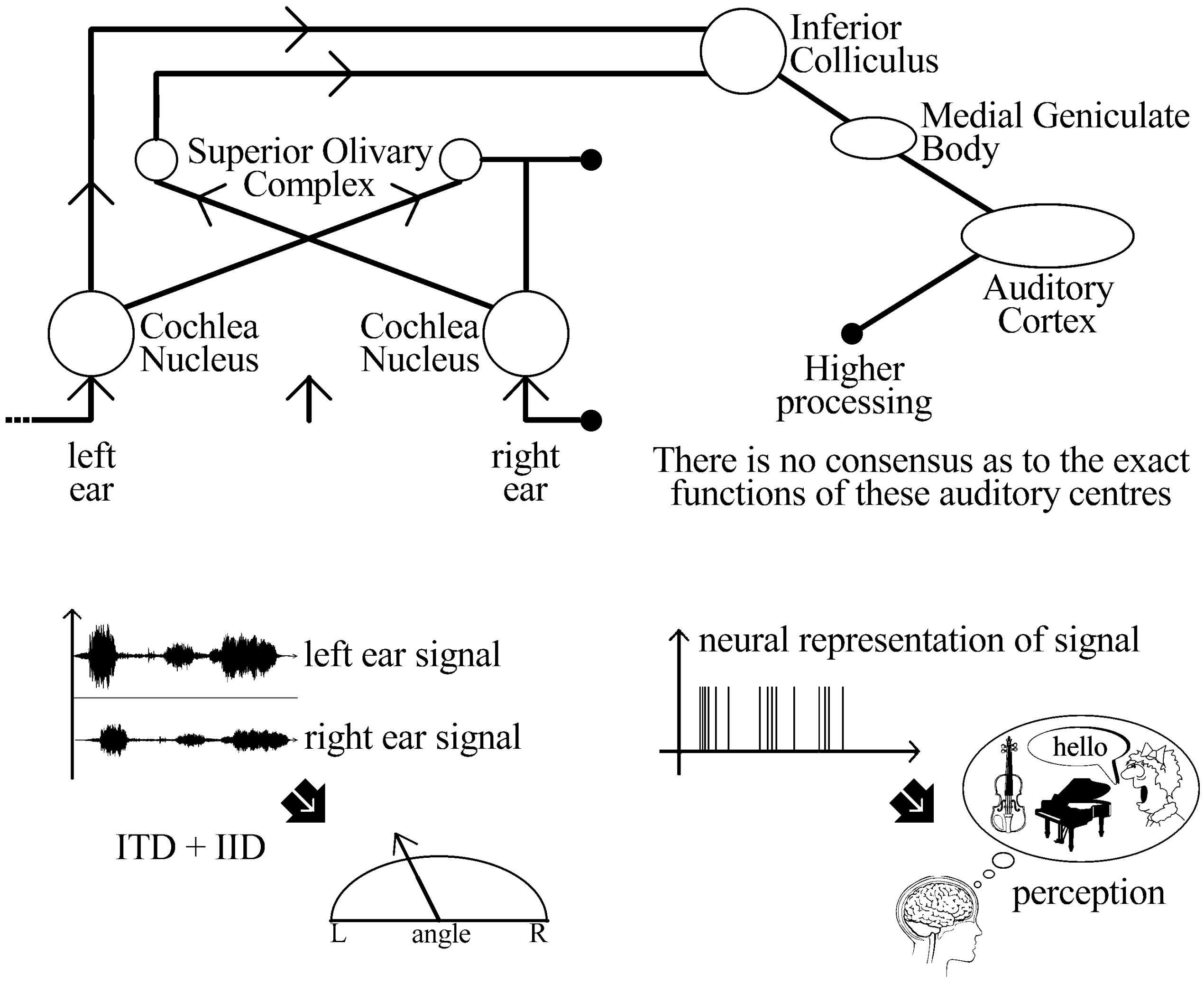
Figure 1 (c) Signal path through the human auditory system

- within the cochlea: the hair cells

Upper half: physiology. Lower half: functionality

The motion of the BM causes the firing of the inner hair cells that are distributed its length. The outer hair cells act to tune the resonant properties of the BM due to signals fed back from the brain. The signals from the inner hair cells pass along the auditory nerve.

neural signal processing



There is no consensus as to the exact functions of these auditory centres

interaural difference gives source angle

attention, analysis and perception

Figure 1 (d) Signal path through the human auditory system

- neural signal processing

Upper half: physiology. Lower half: functionality

The cochlea nucleus acts to sharpen the features of the incoming sound, while the superior olivary complex is responsible for our perception of sound location. The function of other neural centres higher up the human auditory system is debated, but they lead to our perception and understanding of the audio signal as speech, music, noise, or any other event.

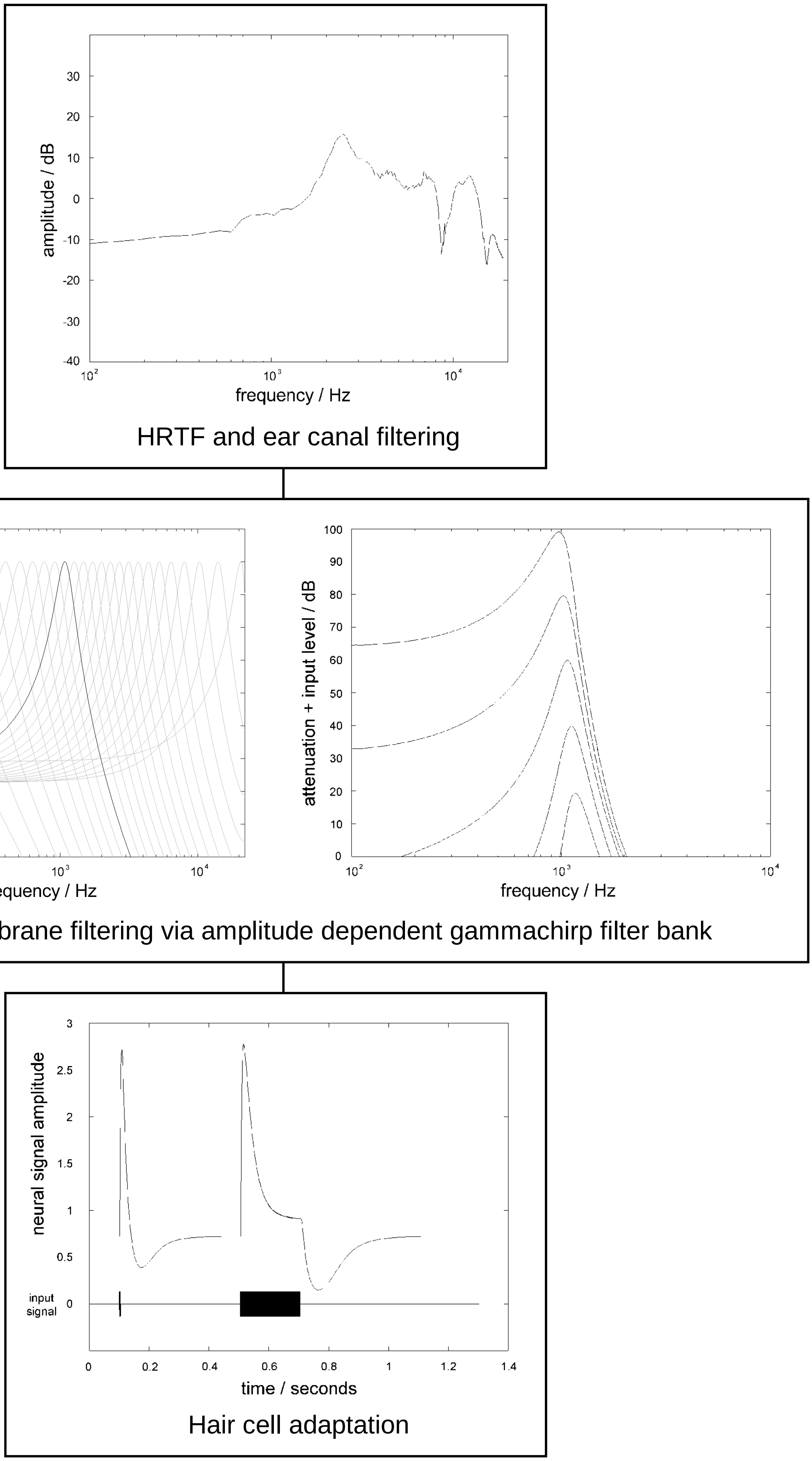


Figure 2 – Structure of the auditory model. See section 1.2 for details.

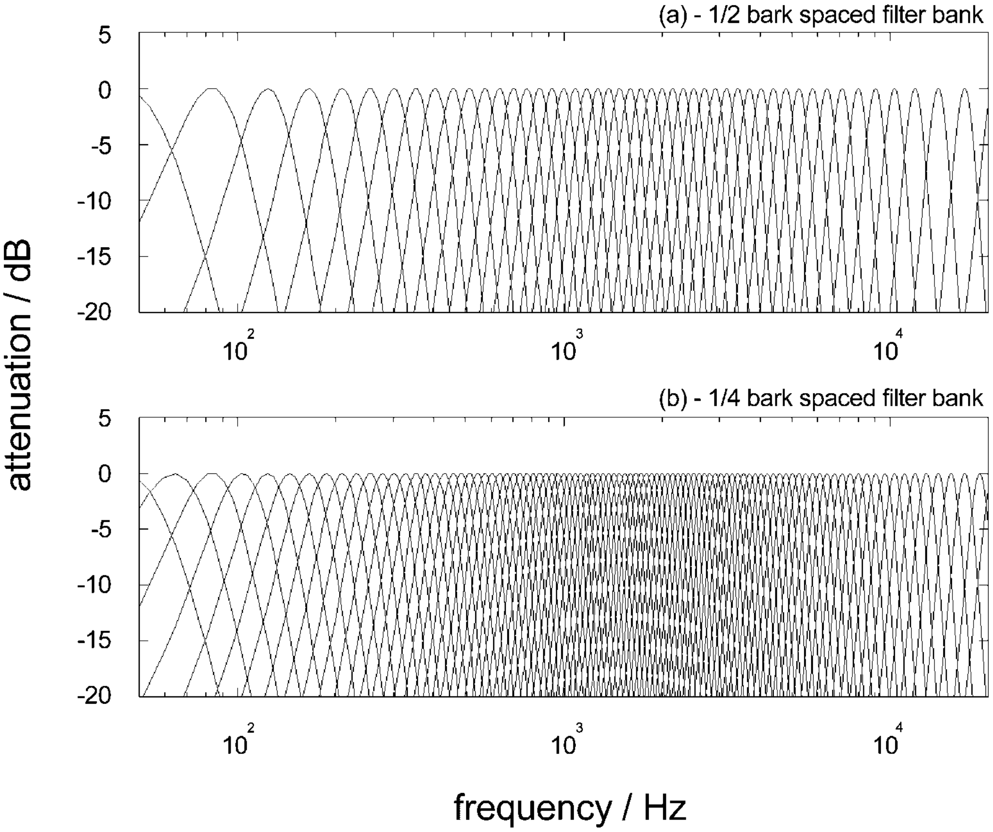


Figure 3 – bank of gammachirp filters, illustrating the coverage given to the audible range of frequencies by various filter spacings.

– filters spaced at ½ bark; (b) – filters spaced at ¼ bark

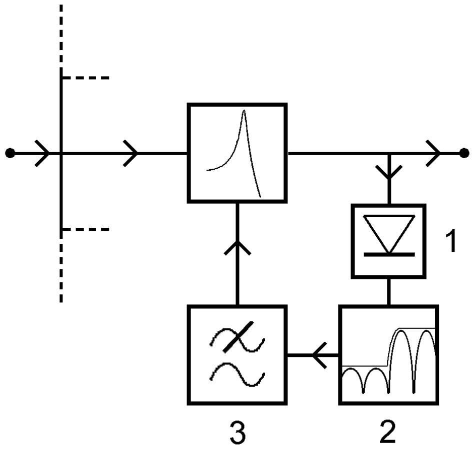


Figure 4 – amplitude dependence processing

The output of the gammachirp filter is rectified (1), peak detected (2), low pass filtered (3) *t*=1ms, and the result is fed back to modify the shape of the filter.

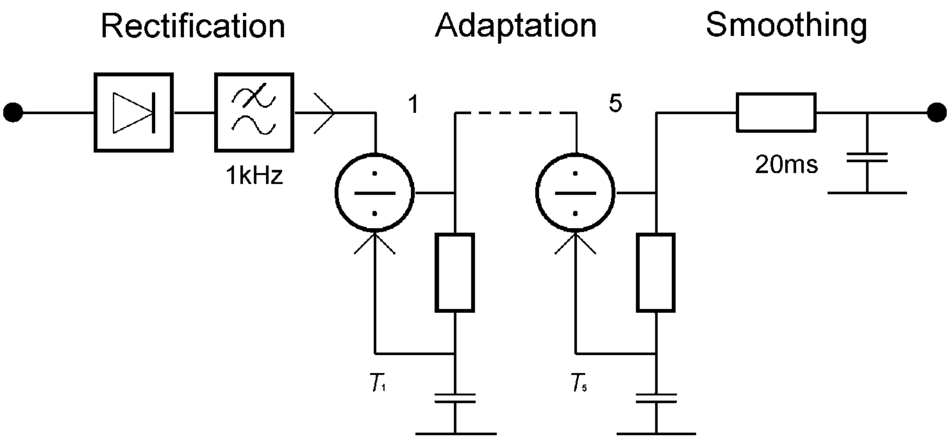


Figure 5 – simulation of hair cell transduction

The output of the gammachirp filter is rectified and low-pass filtered. The signal passes through 5 attenuators, and is then integrated with a time constant of 20ms.

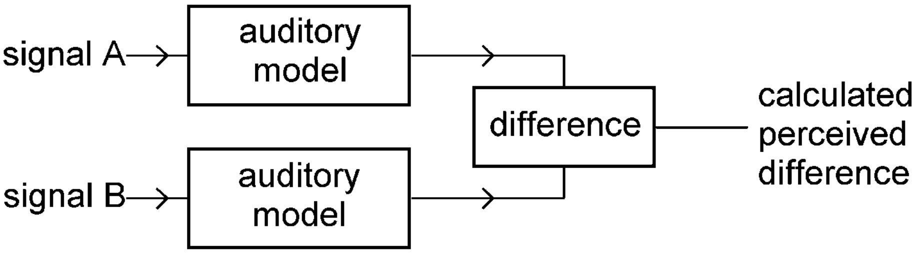


Figure 6 – general method for calculating the perceived difference between two audio signals

In this generalised process, “difference” an represent any method of calculating a difference between the two sets of time varying, frequency band signals resulting from the auditory models.

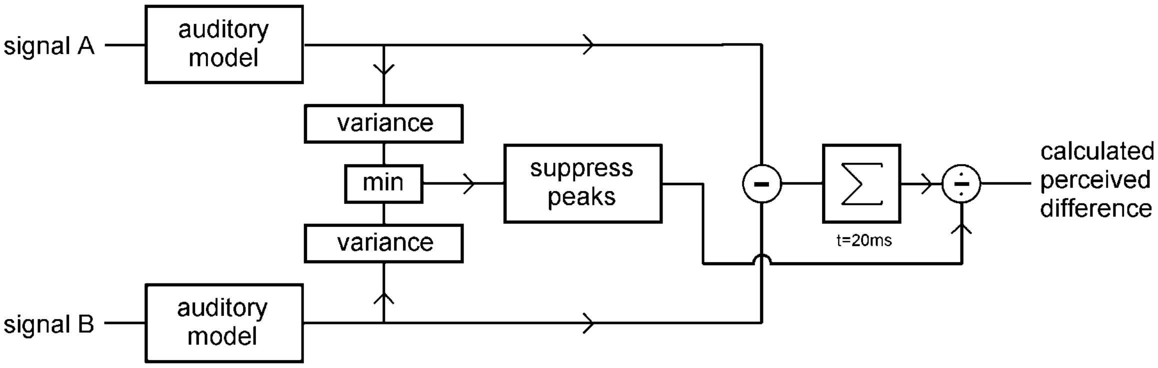


Figure 7 – actual method for calculating the perceived difference between two audio signals

The difference signal is summed over a 20ms interval, and weighted by the inverse of the variance. See section 1.3.2 for a full explanation.

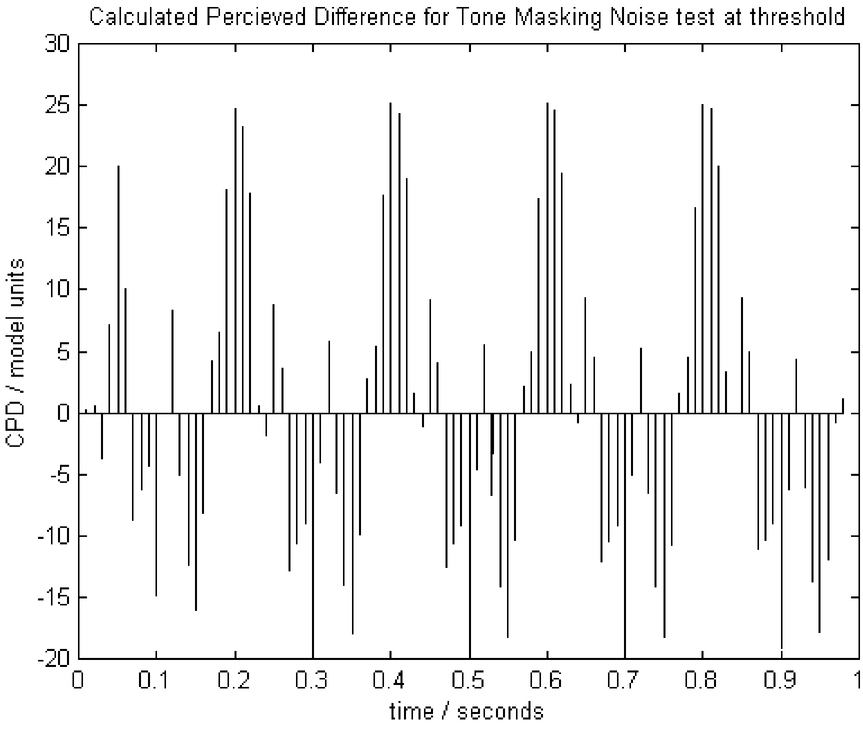


Figure 8 – CPD at threshold for Tone masking Noise test

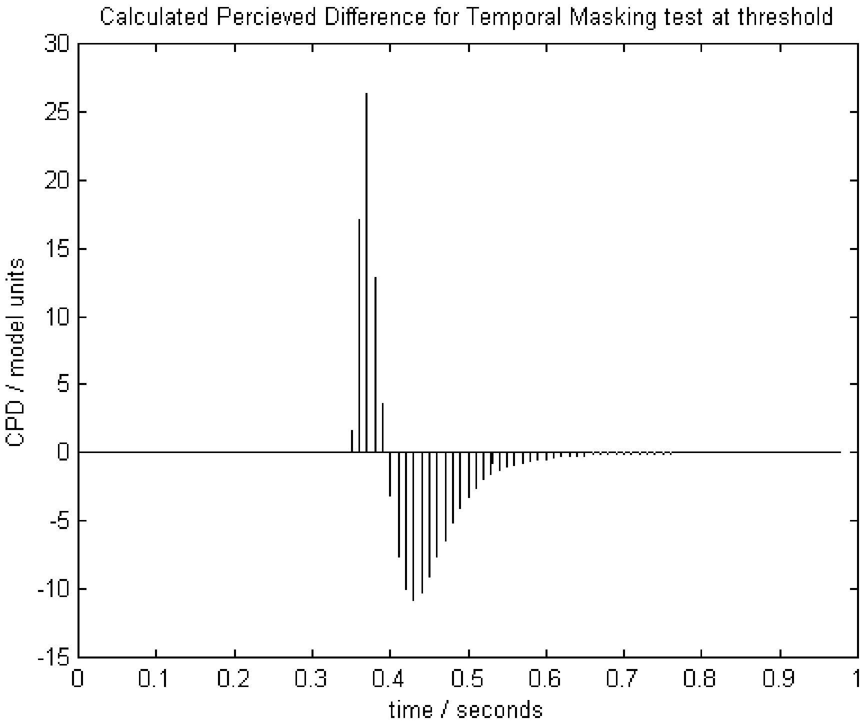


Figure 9 – CPD at threshold for Temporal Masking test

The large value at 0.36 seconds is the tone detected, following the burst of noise (0.1-0.3 seconds – no difference detected)

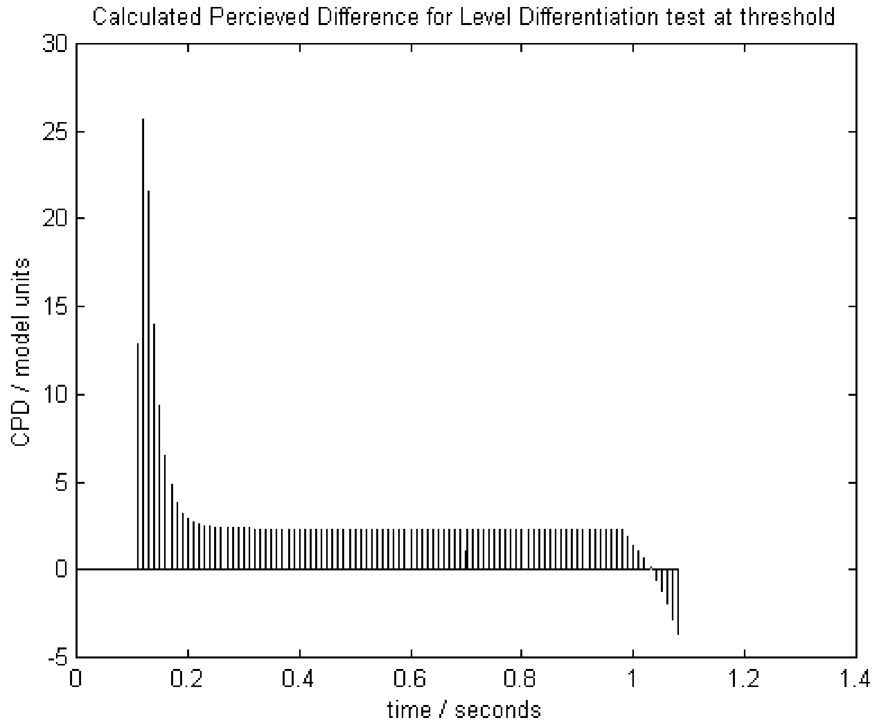


Figure 10 – CPD at threshold for Level Differentiation test

The largest cue to the level of the tone is at its onset (0.1 seconds), hence this is where the change in level yields the greatest perceived difference.

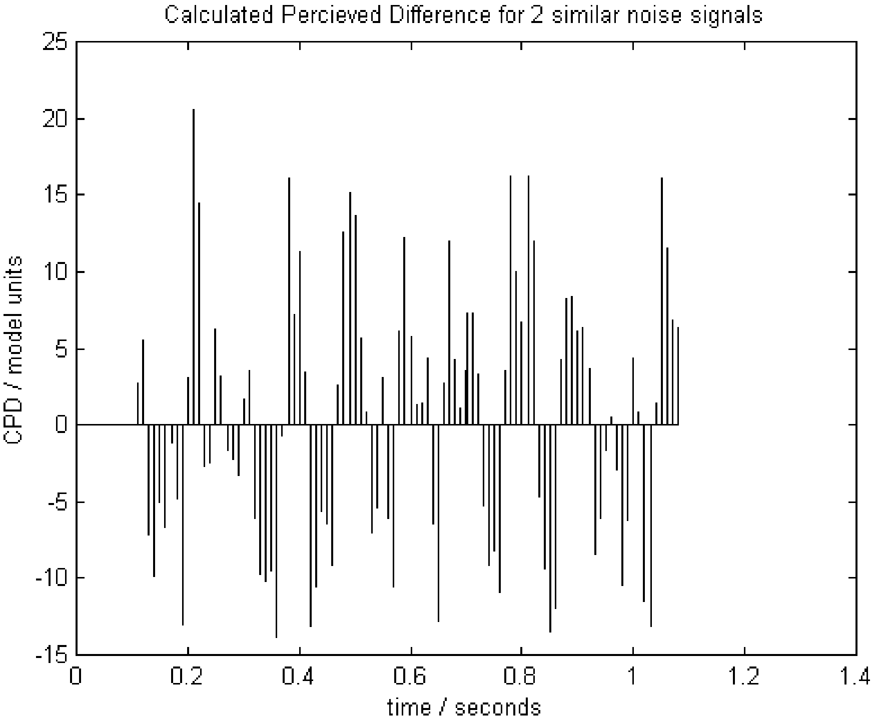


Figure 11 – CPD for 2 perceptually identical random noise signals

Note how, though the random differences cause a random difference signal, this signal never reaches the audible difference threshold of 25, discovered from the three previous tests at threshold.

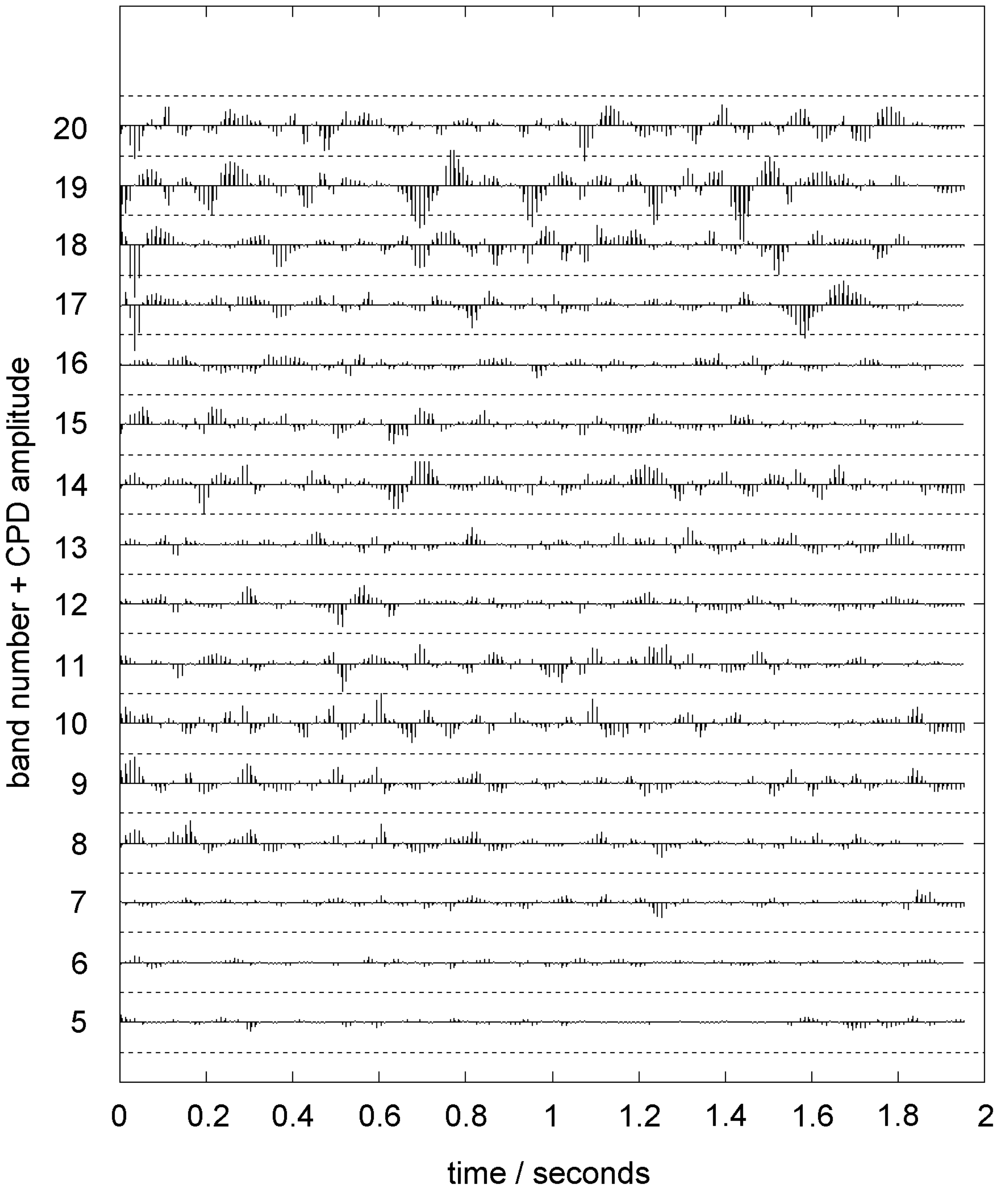


Figure 12 – Calculated perceived difference for an audio signal coded using MPEG-1 layer-2

Only 15 of the 25 bands are shown. The dashed lines show the threshold of audibility in each band (where CPD > 25). See section 2.2.1 for full results and analysis.