TIME-DOMAIN AUDITORY MODEL FOR THE ASSESSMENT OF HIGH-QUALITY CODED AUDIO

David J M Robinson & Malcolm J Hawksford [*]

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 human ear, as well as its characteristics, to yield an accurate prediction of human perception.

1 Background

State-of-the-art audio codecs reduce the amount of data required to represent an audio signal by discarding components that may be inaudible to human listeners. Traditional performance metrics, such as the signal to noise ratio, frequency response etc., cannot quantify the perceived audio quality of the resulting signal, so human listeners are required to grade the audio quality in a subjective test (for an example, see [1]). This is an expensive and time-consuming task, and an alternative is sought. We wish to replace the subjective test with an objective, automated process that can yield the same results. To do this, we require a simulated listener that "hears" the same level of detail as a human, but no more. In this paper we discuss an auditory perceptual model that provides the "ears" for such a listener. Such a model may form the core of a measurement algorithm that replaces the entire subjective test procedure.

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], also [4-6]). 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, hence the assessment of the former, by the latter, will yield a correct indication of perceived audio quality. However, as audio codecs improve, existing measurement algorithms may fail, and there is a need to develop more accurate perceptual measurement models, for future use.

We hypothesise that a perceptual model based closely on the actual processes found within the human ear may yield the most accurate performance. In this paper we will examine the processes involved in human hearing, develop a model to simulate these processes, and determine if the model is appropriate for the quality-assessment of coded audio.

[*] Centre for Audio Research and Engineering, Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Email: djmrob@essex.ac.uk

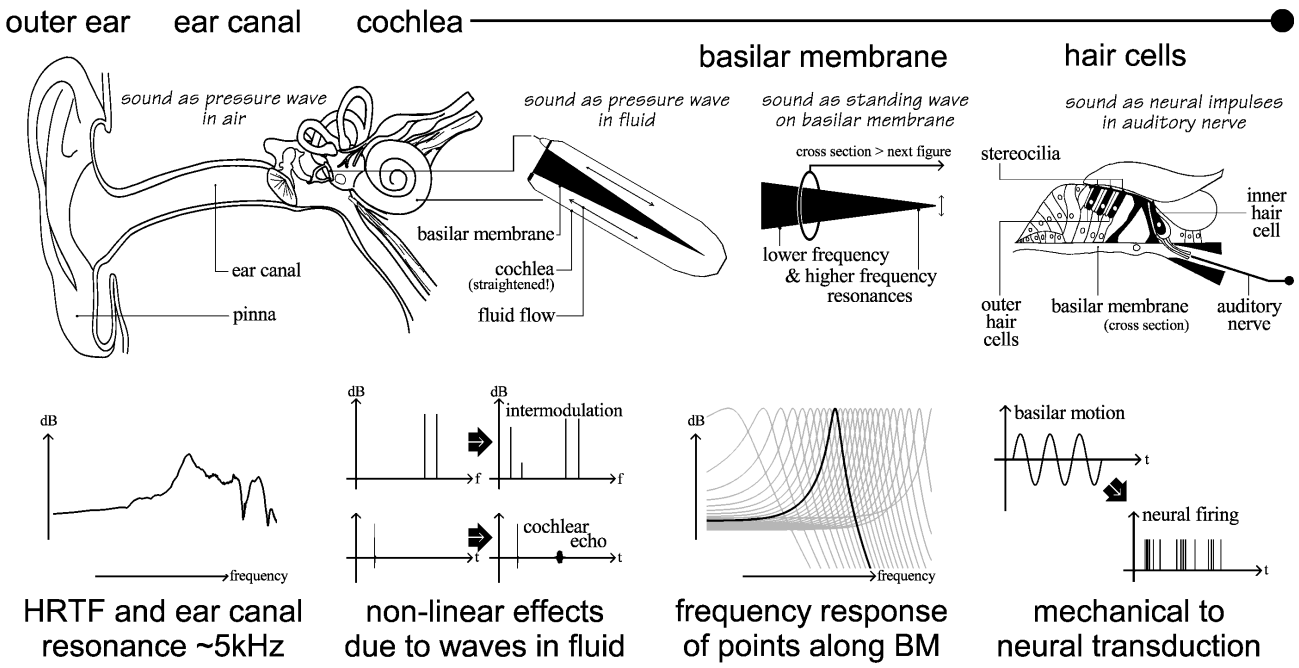

Figure 1. Structure and function of the human auditory system.

All frequency domain plots show amplitude in dB against log frequency. All time domain plots are linear on both scales.

2 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. (Hair cell from [7]).

The **pinna** directionally filters incoming sound (see [8]). The **ear canal** filters the sound, giving a resonance at around 5 kHz. The sound is transmitted through small bones in the middle ear into the **cochlea**. The fluid-filled **cochlea** is a coil within the ear, partially protected by bone. 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 frequency selectivity is governed by the width of the filter characteristic at each point. The **outer hair cells**, distributed along the length of the BM, react to feedback from the brainstem to change the resonant properties of the BM. The **inner hair cells** on the BM fire when the BM moves upwards, so transducing the sound wave at each point into a signal on the auditory nerve. This effectively half wave rectifies the signal.

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 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. 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.

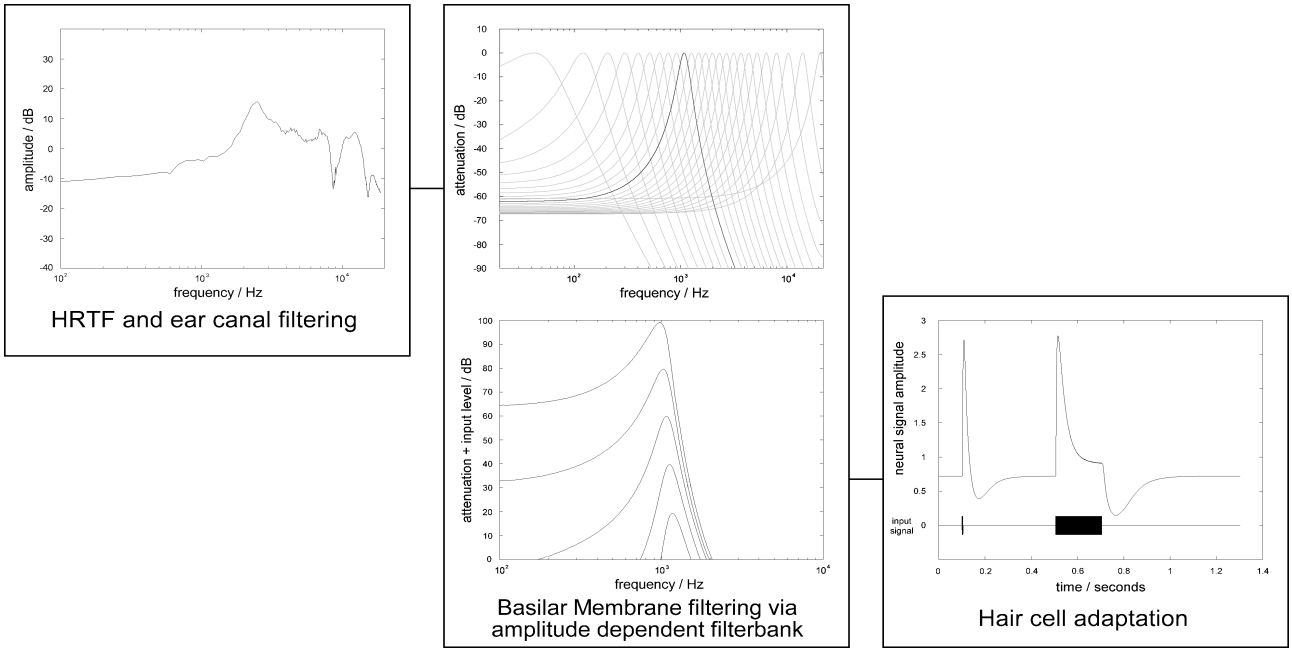

Figure 2. Outline of the structure of the auditory perceptual model.

3 The Structure of the auditory model

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

3.1 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 realistic alternative.

3.2 Basilar membrane filtering

A bank of amplitude dependent filters simulates the response of the Basilar Membrane. Each filter is an FIR implementation of the gammachirp, described in [9], and simulates the response of the BM at a given point. The filters are linearly spaced on the Bark frequency scale [10], which itself accurately describes the spacing of resonant frequencies along the BM.

A key feature of this model is the amplitude dependent nature of the filter bank. As the basilar membrane changes its resonant properties in response to the amplitude of the incoming signal, so the shape of each gammachirp filter is dependent on the signal amplitude at the output of the filter. The envelope of the signal is derived by rectification, peak detection, and low pass filtering, as illustrated in Figure 3. This envelope is then used to adjust the filter coefficients.

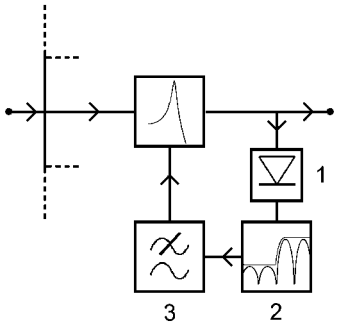

Figure 3. Amplitude dependence circuit.

3.3 Hair cell transduction

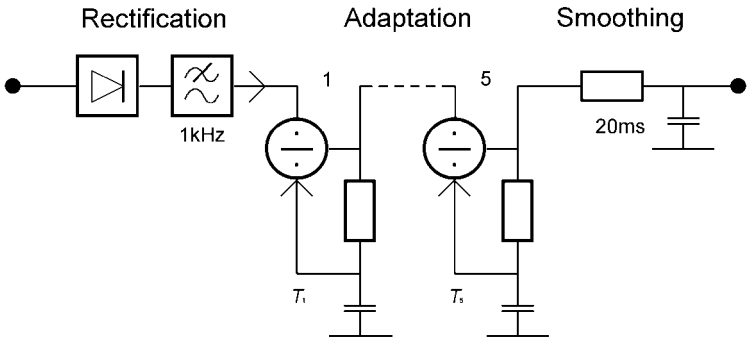

Figure 4. Hair cell transduction circuit.

At each point along the basilar membrane, its movement is transduced by a large number of hair cells. The firing of individual hair cells is pseudo-random, and only by combining signals from large numbers of hair cells can the auditory system sense the motion of the BM. It would be a computationally burdensome task to simulate the response of each individual hair cell [13-15], and then combine many thousands of responses, so a simpler solution is sought, as suggested in [16].

The first stage is to half wave rectify the output of each gammachirp filter, and low pass filter the result at 1 kHz (see Figure 4). Thus we simulate the half wave response of the inner hair cells. The increased sensitivity of the inner hair cells to the onset of sounds can be viewed as a type of adaptation, which is compounded by feedback to the outer hair cells. We simulate this adaptation by a cascade of 5 feedback loops, followed by a smoothing circuit, as shown in Figure 4 (from [17]). Finally, we limit the minimum value to be equivalent to the absolute threshold of hearing, taken from [18]. This takes account of the internal noise, due to the random firing of the inner hair cells and the blood flow within the ear.

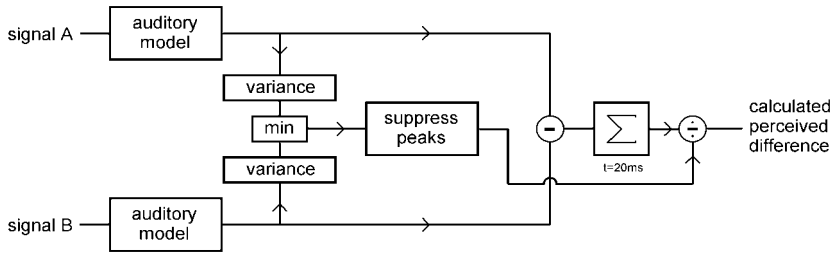

Figure 5. Circuit for calculating the perceived difference between two audio signals.

4 Perceiving a difference

So far, we have simulated the passage of sound through the ear canal and cochlea; thus the output of the auditory model is a signal analogous to that transmitted along the auditory nerve. If two audio signals are processed independently by the auditory model, the difference between the two resulting sets of outputs will be *related* to the perceived difference between the two signals. However, determining the exact nature of this relationship is a non-trivial task, which cannot be tackled by an examination of the underlying physiology, since the physiology involved in this process is not yet understood. Instead, we shall use the results from previous tests upon human subjects to infer the processing that may be present. From a comparison of the calculated nerve signals, with the perception of a real human listener, we deduce that the subsequent auditory processing includes three critical features, simulated thus:

a) further temporal smearing, simulated by a summation over time.
b) greater sensitivity to distortion in tone-like sounds than in noise-like sounds, simulated by weighting the results by the variance of the nerve signal.
c) an internal perceptual threshold, simulated by defining a calculated perceived difference (CPD) value below which no difference is perceived.

The circuit shown in Figure 5 carries out this processing. Tuning the circuit to match actual human perception yields a CPD threshold of 25 model units.

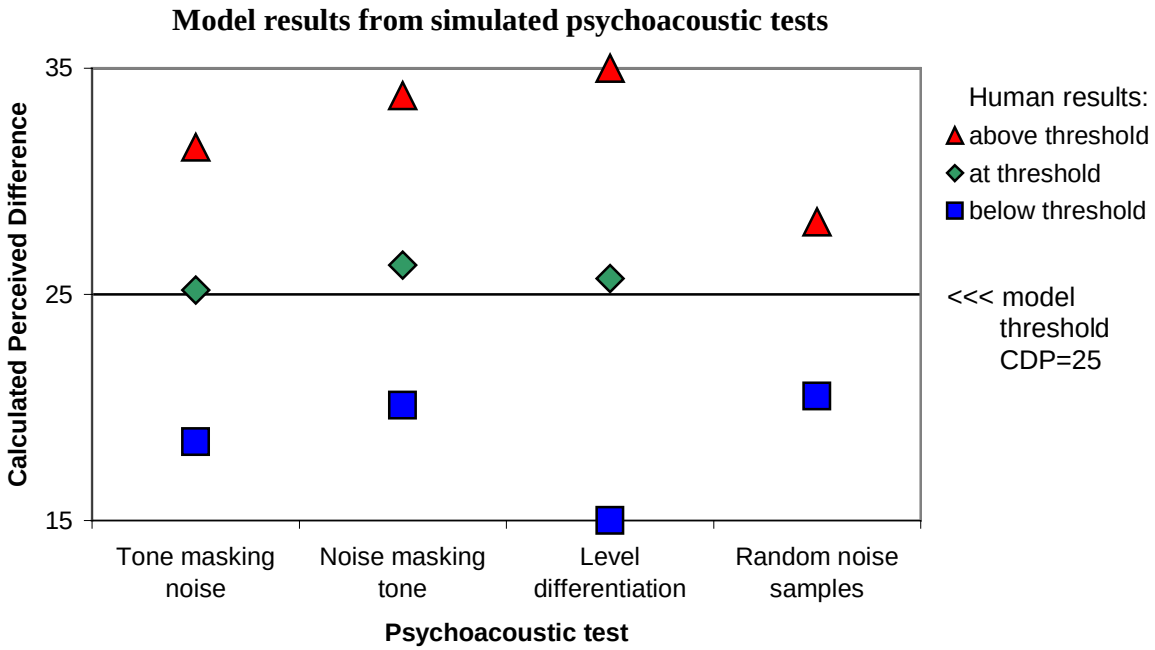

Figure 6. The model correctly predicts the audible and inaudible conditions in a variety of psychoacoustic listening tests.

5 Validation of the model

5.1 Psychoacoustic tests

The following series of psychoacoustic tests were simulated via the model: Tone masking noise (simultaneous) [19]; Noise masking tone (temporal post-masking) [20]; Level differentiation [16]; and Random noise. The references give the threshold values at which it is just possible for a human listener to detect whichever difference is present. The results in Figure 6 show that in all cases the model correctly predicted the threshold at which human listeners could just perceive a difference, and also correctly identified above and below threshold (i.e. threshold +/- 3 dB) conditions.

5.2 Codec assessment

To test its accuracy in predicting human perception, the model was used for the quality assessment of coded audio. An audio sample taken from LINN CD AKD 028, Track 4, 0:27-0:29 (a piece of vocal jazz) was encoded at a variety of bit-rates (ranging from 112-256 kbps) using MPEG-1 layer 2 and layer 3 codecs. The results of existing subjective tests were used as a benchmark of the perceived audio quality of these codecs, and the model compared the original audio sample with each of the coded versions to give a calculated perceived difference, as explained above. The model results matched human perception in two important aspects: Firstly, the model correctly predicted the transparent bit-rate for each codec; and secondly, the model placed the performance of both codecs at all bit-rates in an identical ranking to the human listeners.

As a further test, the model was used to test the Microsoft audio codec. Informal listening tests were used to assess the human perception of this codec. Though the model once again correctly predicted the transparent bit-rate, the ranking of this codec along side the MPEG codecs by the model did not match human perception. Specifically, the model predicted 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 upon the onset of sounds.

It seems that some higher auditory process must attenuate 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.

6 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] D. J. M. Robinson and R. G. Greenwood, "A Binaural simulation which renders out of head localisation with low cost digital signal processing of Head Related Transfer Functions and pseudo reverberation," presented at the 104th Convention of the Audio Engineering Society, preprint 4723.

[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.).