Perceptual model for assessment of coded audio

David J M Robinson

A thesis submitted for the degree of Doctor of Philosophy

Department of Electronic Systems Engineering

University of Essex

March 2002

Summary

In this thesis, a time-domain binaural auditory model is developed for the assessment of high quality coded audio signals.

Traditional objective measurements of sound quality yield misleading results when applied to psychoacoustic based codecs. Whilst human subjective opinion must be the ultimate arbiter of perceived sound quality, an objective measurement is sought that accurately predicts human perception.

The auditory model described in this thesis is designed to match the processing within the human auditory system as closely as possible. The outer and middle ear transforms are simulated by linear filters; the basilar membrane response is simulated by a bank of amplitude dependent gammachirp filters, and the compressive action of the inner hair cells is simulated by an adaptation circuit. The whole model is calibrated using known human performance data.

This monophonic model is used as the basis for a model of binaural hearing. Signals from the inner hair cell simulation are compared with those from the opposing ear, and an internal binaural image is formed. Changes in this image are shown to reflect audible changes in the binaural properties of the sound field. The binaural model predicts binaural masking and localisation data using a single detection criterion, which suggests that both phenomena may be due to the same internal process.

Finally, the model is used for audio quality assessment. The monophonic model correctly predicts human perception for some coded audio extracts, but not all. A specific problem in the hair cell simulation is identified, and solutions to this problem are suggested. The binaural model accurately predicts human perception of spatial qualities. In particular, the model can detect changes to the stereo soundstage within real music stimuli.

Table of Contents

Abbreviations

Abbreviation	Description
AAC	Advanced Audio Coding
ABX	Blind testing methodology
AC-3	Audio Coding 3 – perceptual codec from Dolby Laboratories
ADC	Analogue to Digital Converter
ADPCM	Adaptive Differential Pulse Code Modulation
AM	Arithmetic Mean
B+K	Brüel and Kjær
BBC	British Broadcasting Corporation
BM	Basilar Membrane
BMLD	Binaural Masking Level Difference
BT	British Telecommunications
CB	Critical Band
CD	Compact Disc
CD-ROM	Compact Disc Read Only Memory
CDT	Cubic Distortion Tone
CPD	Calculated Perceived Difference
DAC	Digital to Analogue Converter
dB	Decibels
DPOAE	Distortion Product Otoacoustic Emission
DVD	Digital Versatile Disc
EBU	European Broadcasting Union
EEVM	Element by Element Vector Multiplication

ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Frequency Modulation
fs	Sampling Frequency
GM	Geometric Mean
GSM	Groupe Spécial Mobile or Global System for Mobile Communications
HAS	Human Auditory System
HRTF	Head Related Transfer Function
ID3	Identification Tag for mp3 audio files
IID	Interaural Intensity Difference
IIR	Infinite Impulse response
ILD	Interaural Level Difference
ina	Inaudible level
ISO	International Organisation for Standardisation
ITD	Interaural time difference (or Delay)
ITU	International Telecommunication Union
kB	kilo Bytes
kbps	kilo bits per second
KEMAR	Knowles Electronic Manikin for Acoustic Research
LAME	Lame Ain't an Mp3 Encoder
LCDT	Level of the Cubic Distortion Tone
LFD	British Audio equipment manufacturer
LPAC	Lossless Predictive Audio Compression
LSO	Lateral Superior Olive
MAF	Minimum Audible Field
MATLAB	Matrix Laboratory – mathematical programming environment
MB	Mega Bytes
MDCT	Modified Discrete Cosine Transform
MHz	Mega Hertz
MLS	Maximum Length Sequence
mp2	MPEG layer II audio coding scheme
mp3	MPEG layer III audio coding scheme
MPC	MusePack – independent audio codec, formerly known as MPEG*plus*

MPEG	Moving Pictures Expert Group
ms	maximum SPL
MSO	Medial Superior Olive
MUSHRA	Multi-Stimulus test with Hidden Reference and Anchor
ODG	Objective Difference Grade
PCM	Pulse Code Modulation
PEAQ	Perceptual Evaluation of Audio Quality
RGAD	Replay Gain Adjustment
RMS	Root Mean Square
SDG	Subjective Difference Grade
SFM	Spectral Flatness Measure
SL	Spectrum Level
SMPTE	Society of Motion Picture and Television Engineers
SMR	Signal to Mask Ratio
SNR	Signal to Noise Ratio
SOC	Superior Olivary Complex
SPL	Sound Pressure Level
THD	Total Harmonic Distortion
THD+N	Total Harmonis Distortion plus Noise
TV	Television
UK	United Kingdom

1

Introduction

For over a century, scientists and engineers have refined the art of sound reproduction. Throughout this period, the goal has been to increase the fidelity of the reproduced sound to that of the original event. To this end, each component within the audio reproduction chain has been extensively perfected. Subjectively, the ideal audio device should have no audible character of its own, and the audio signal should pass through the device without the addition of colouration, distortion, or noise.

The success or otherwise in approaching this ideal can be judged by listening to the resulting sound. However, a rigorous engineering approach requires that the performance of any audio device must be quantifiable in objective terms. A variety of objective measurements have been developed, and these quantify the deviation of the device from the stated ideal. Many state-of-the-art audio devices are so close to perfection that most measurable performance deviations are below the threshold of audibility, whilst any perceived audible character of the device is due to some effect which is not yet quantifiable. However, where measured performance deviates significantly from the ideal, such measurable anomalies are audible, and the quantitative measurement correlates well with human perception.

This correlation between objective and subjective performance is broken by audio codecs. An audio codec aims to reduce the amount of data that is required to store or transmit an audio signal. High quality psychoacoustic-based audio codecs achieve this goal by discarding information that is inaudible to a human listener. This contradicts the historical notion of an ideal audio device, which should not alter the audio signal in any manner. Instead, the ideal audio

codec should change an audio signal as much as possible (by discarding data) without changing the perception of the audio signal by a human listener.

It is inappropriate to apply existing objective performance measures to audio codecs, as these measurements merely confirm that the codec has significantly altered the audio signal. Rather than measuring *all* the changes introduced by the device, it would be more appropriate to quantify only the *audible* changes, whilst ignoring any changes that are inaudible to a human listener. This may be termed an objective perceptual measurement, and several such measurements have recently been defined.

The present work focuses on some of the areas of human auditory perception that are neglected within current perceptual measurement algorithms. These areas are temporal masking, spatial masking, and binaural hearing. By necessity, auditory phenomena that are adequately defined in existing models are also discussed. Rather than duplicating previous work, a new approach is tested in this area, whereby the mechanisms within the human auditory system are simulated directly. This allows the appropriateness of this approach for the assessment of coded audio to be investigated.

In Chapter 2, the principles of psychoacoustic-based audio codecs are introduced, and the MPEG-1 series of codecs is described. It is shown that traditional objective audio quality measurements are inappropriate for the quality assessment of coded audio. It is also demonstrated that certain other techniques that are often mistakenly used to predict perceived audio quality will yield misleading results. The definitive arbiter of perceived audio quality must be a human listener, but subjective opinions are notoriously unreliable. Stringent subjective test procedures are described that yield reliable, repeatable qualitative judgements. The expense of implementing these test procedures demonstrates the need for an objective measure that can match the perceptions of a human listener.

In Chapter 3, the mechanisms within the human auditory system are described in detail. The path of an audio signal is followed through the human ear, from air-born pressure wave to internal neural signal. Particular attention is paid to the processes within the human auditory system that give rise to the measurable performance limits of human hearing.

In Chapter 4, existing models of human hearing are described and compared. It is shown that such models can simulate human hearing in both psychoacoustic tests, where the limits of human hearing are probed, and in the assessment of coded audio signals. The important features of those models that give accurate performance are identified for use in the present work. The areas in which existing models do not match human perception are demonstrated, and possible solutions are proposed.

In Chapter 5, a time-domain monophonic auditory model is developed. This model is based upon the processing present within the human auditory system. Some of the features are drawn from existing psychoacoustic models, whilst others are unique to the present model. The model is calibrated to match human performance in a wide range of psychoacoustic tests.

In Chapter 6, a psychoacoustic experiment to investigate spatial masking is described. This experiment investigates the audibility of one sound in the presence of another, where the two sounds are spatially separate. Existing psychoacoustic data describes this phenomenon for headphone presented stimuli, and the present experiment extends this knowledge to free-field stimuli, delivered over loudspeakers. This data is required to calibrate a binaural model capable of auditioning free-field sounds.

In Chapter 7, a time-domain binaural auditory model is developed. Existing binaural models are reviewed, and the most successful features of these models are incorporated into the present model. Novel features are included to facilitate the processing and comparison of an arbitrary time-domain input stimulus. The two binaural processing tasks (localisation and detection) are successfully accounted for by a single detection criterion. This model is calibrated to match human perception using the data from the spatial masking experiment.

In Chapter 8, the model is used to assess the quality of coded audio signals. The monophonic model is shown to operate correctly with transient signals, where many other models fail. However, problems are revealed during passages of real music, where the sensitivity of the model is too low. This is traced to an error within the difference detector. A previous version of the difference detector is shown to predict human performance of real music to a reasonable degree, but is not able to handle transient signals. A solution to this problem is proposed, which involves incorporating knowledge gained during the development of the binaural model into the monophonic model. Finally, the binaural model is used to detect changes in the spatial

characteristics of a stereo audio signal. The correlation between the prediction of the binaural model and the perception of a human listener is excellent.

In Chapter 9, the present work is reviewed. The success of the monophonic and binaural models is critically assessed, and possible future additions to the combined model are discussed.

2

Background

2.1 Overview

In this chapter, psychoacoustic-based audio codecs are introduced. The motivation for the development of these devices is discussed, and several codecs are examined in detail. It is shown that conventional objective measurements of audio quality are inappropriate for the assessment of coded audio. A commonly employed alternative is the subjective test, the procedure of which is described herein. It is shown that these tests are expensive and time consuming, and an accurate objective alternative is sought.

2.2 Audio coding

An audio codec is a device that reduces the amount of data required to represent an audio signal. In this section, the uses and operations of audio codecs are discussed.

2.2.1 Why reduce the data rate?

The compact disc is now so much a part of everyday life that its technological properties are taken for granted. Indeed, the 750 MB of audio data contained upon a typical CD seems small compared to the capacity of current storage devices. Moore's law [Moore, 1965] predicts that computational processing power will double every 18 months. Data storage capacity is increasing at a similar rate. The capacity of the humble CD will seem minuscule compared to next year's hard disk drives and future optical disc formats.

As the storage capacity of a CD is dwarfed, it is easy to forget that the data requirements of CD quality digital audio are immense compared to textual media. For example, 30 seconds of CD

quality digital audio requires the same storage space as the complete works of Shakespeare[1]. Though the cost of digital storage falls year on year, the data rate of CD quality audio is still too high for certain applications. Two pertinent examples are discussed below.

Firstly, audio broadcasters wish to transmit CD quality radio services. However, the radio spectrum is very crowded, and the proliferation of devices such as mobile phones has made radio bandwidth an expensive commodity. If CD quality audio were transmitted on existing analogue FM frequencies, then the frequency range from 88 MHz to 108 MHz would accommodate just 12 radio stations. However, analogue transmissions must continue during the transition to digital broadcasting, so additional bandwidth has been allocated for the digital services[2]. The bandwidth allocated for the five BBC national radio stations is 1.54 MHz. After channel coding, this yields a broadcast data rate of 1.2 Mbps. The data rate of CD is 1.4 Mbps. Thus, a single CD-quality audio service requires more bandwidth than is available for five radio stations.

Secondly, computer networks, especially home connections, have failed to increase in capacity in accordance with Moore's law. The most common internet connection at home in the UK is currently the 56k modem. Data transfer rates of approximately 3-4 KB per second (32000 bits per second) are typical. Thus, for every one second download time, the user can transfer 0.0227 seconds of CD quality audio. Real-time delivery of audio in this manner is impossible. Distributing albums of music over the internet for off-line listening is similarly impractical, since a 3 minute pop song requires over two hours download time.

The data rate of CD quality digital audio is too high for both these applications. The data rate must be reduced in order to make either application practical. In addition, there are other applications where the data rate of CD quality audio is not prohibitive, but reducing this data rate would provide economic or functional benefits. For these reasons, it is desirable to reduce the

[1] The complete works of Shakespeare in ASCII Plain text format [Farrow, WEB] occupy 5219KB, or 44153344 bits. 30 seconds of CD quality audio occupy 44100*16*2*30=42336000 bits. Thus this edition of the complete works of Shakespeare requires the same binary storage as 31.3 seconds of CD quality digital audio.

[2] In the United Kingdom, 12.5 MHz of Band III spectrum from 217.5 - 230 MHz has been allocated to digital audio broadcasting. This will accommodate seven data channels. The BBC has been allocated one of these channels for its national services [Bower, 1998].

data rate of the audio signal, *without* compromising the audio quality. However, without sophisticated audio codecs, the data rate and audio quality are inextricably linked.

2.2.2 Data reduction by quality reduction

The simplest method of reducing bitrate[3] is to reduce the audio quality. Three bitrate reduction strategies are listed below, together with the quality implications for each strategy.

1. Reduce the sampling rate. This will reduce the frequency range (bandwidth) of the audio signal.
2. Reduce the bit-depth. This will increase the noise floor of the audio signal.
3. Convert a stereo (2-channel) signal to a mono (1-channel). This will remove all spatial information from the audio signal.

Table 2.1 lists some common audio formats. These illustrate various combinations of the above strategies.

name	samples / second	PCM bits / sample	channels	frequency range / Hz	SNR / dB	PCM bit rate / kbps
DVD	96000	24	6	48 kHz	144	13824
DAT	48000	16	2	24 kHz	96	1536
CD	44100	16	2	22 kHz	96	1411
“FM”	32000	12	2	16 kHz	72	768
“FM”	32000	12	1	16 kHz	72	384
“PC”	22050	8	1	11 kHz	48	176
Phone	8000	8	1	3.4 kHz	48	64

Table 2.1: Linear PCM Bitrates

The lowest bitrate in Table 2.1 is still too high to transmit in real time over a 56k modem. The stereo “FM” parameters define a digital channel with comparable quality to existing analogue

[3] Throughout this discussion, the data rate of an audio signal will be referred to as the “bitrate”. The bitrate is specified in bits per second (bps), kilobits per second (kbps), or Megabits per second (Mbps). The “k” and “M” prefixes are used to represent 10^3 and 10^6 respectively (SI units) rather than 2^{10} and 2^{20} (commonly used in PC specifications - see [IEC 60027-2, 2000] for clarification of this issue).

FM broadcasts. This quality is acceptable to most consumers, but quality reductions below this level are perceived and disliked by many listeners.

To reduce the bitrate further, a more sophisticated approach is required.

2.2.3 Lossless and lossy audio codecs

There are two distinct types of audio codec: *lossless* and *lossy*. A lossless codec will return an exact copy of the original digital audio signal following the encode and decode process. A similar approach is often used within the computer world to reduce the size of documents or program files, without changing the data. Algorithms suitable for data include "Zip" [PKWARE, WEB] and "Sit" [Aladdin Systems, WEB]. Algorithms suitable for audio include "LPAC" [Liebchen, WEB], "Meridian Lossless Packing" (MLP) [Gerzon *et al*, 1999], and "Monkey's Audio" [Ashland, WEB]. Both types of algorithm exploit redundancies within the data. For example, the waveforms of musical signals are often repetitive in nature. Storing the difference between each cycle of the waveform, rather than the waveform itself, often requires fewer bits. In a lossless codec, the difference between the predicted values and the actual waveform is also stored, so that the waveform can be reconstructed exactly.

A lossless audio codec by definition cannot reduce the audio quality. However, lossless audio codecs rarely reduce the bitrate to below 50% of the original value. Also, the exact bitrate reduction is highly signal dependent, so the bitrate of the audio data cannot be guaranteed to match that of the transmission channel. A burst of white noise (which is random and hence difficult to predict or compress) may cause the encoded bitrate to match or exceed that of the original signal.

To reduce the bitrate still further, *lossy* audio codecs discard audio data. This means that the decoded waveform is not an exact copy of the original. However, unlike the measures described in 2.2.2, lossy audio codecs aim to discard data in a manner that is inaudible, or at least not objectionable to a human listener. This is possible due to the complex nature of human hearing. This topic will be discussed in depth in Chapter 3, but here it is sufficient to note that the presence of one sound can prevent a human listener from hearing a second (quieter) sound. This phenomenon is illustrated in Figure 2.1 [Rimell, 1996].

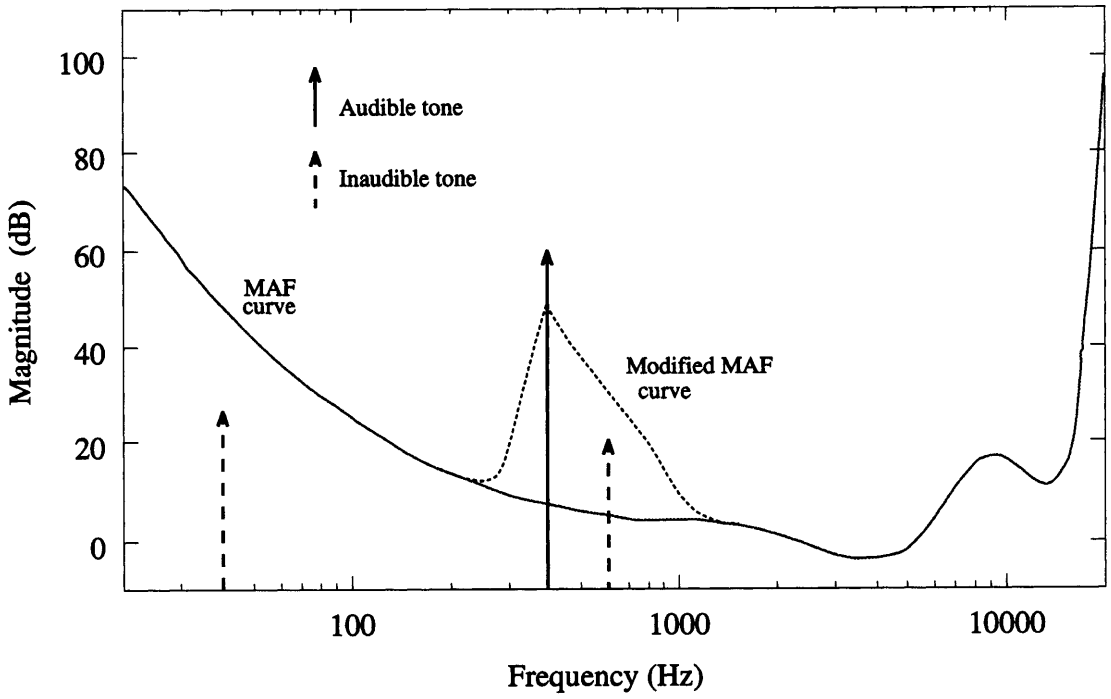


Figure 2.1: Spectral masking

The Minimum Audible Field (MAF) curve represents the threshold of audibility at a given frequency. Thus, the 40 Hz tone (shown by the dashed arrow on the left Figure 2.1) is inaudible, because it lies below the MAF curve.

The presence of an audible tone raises the threshold in the spectral region around the tone, and any additional sound falling below the modified MAF curve will be inaudible. For example, the dashed arrow in the centre of Figure 2.1 represents a tone of 600 Hz at 20 dB SPL. This tone would be audible in isolation, but is rendered inaudible (or masked) by the 400 Hz tone at 60 dB. The modified MAF curve is often referred to as the masked threshold.

This concept of masking is used in audio coding. A masked sound can be removed or distorted by the audio codec without changing the *perceived* quality of the audio signal. Lossy codecs which operate in this manner are often referred to as psychoacoustic based codecs, since they require knowledge of the properties of the human auditory system.

By combining this approach with lossless data reduction, the bitrate may be reduced by 90% without significantly reducing the perceived audio quality. The result is that a 128 kbps data

stream, which provides little better than telephone quality without data reduction, can yield near CD quality with data reduction.

Psychoacoustic based codecs are the most recent generation of lossy audio codecs. Two other types or families of lossy audio codec exist, and these are mentioned in passing. The first type aims to discard data without significantly reducing the perceived quality of the audio signal, but does so without sophisticated knowledge of the human auditory system. The oldest such codecs are is the A-law and μ-law coding schemes, where non-linear quantisation steps are used to increase the perceived signal to noise ratio of an 8-bit quantiser.

Another lossy coding mechanism is Adaptive Differential Pulse Code Modulation. In ADPCM, each sample is predicted from the previous samples, and only the difference between the prediction and the actual value is stored. The decoder follows the same predictive rules as the encoder, and adds the stored difference to each predicted sample value. Typically, the input samples are of 8 or 16 bit resolution, and the encoded differences are stored in four bit resolution, giving 50% or 75% data reduction. This codec is lossless, except where the difference between the predicted and actual values cannot be represented in four bits. In practice, this situation is common, but the error is sometimes inaudible, and rarely annoying.

Both the above lossy codecs are designed for use with telephone quality speech signals, though they can be used with some success to code CD quality music signals. There is a further type of lossy codec which is designed for speech coding only. Code excited linear predictive coding employs a code book of excitation signals followed by a linear predictive filter. The output of the code book and filter is compared with the incoming speech signal, and the code book index which gives the best match is transmitted. Typically, a single 10-bit index into the code book can represent 40 incoming samples. This mechanism of lossy coding is used on digital mobile telephone networks, and the code book is designed to represent speech-like sounds. This approach is not suitable for high quality music coding, as anyone who has heard music via a GSM mobile phone can testify. These speech only lossy codecs are not relevant to the present work, and will not be discussed further.

Psychoacoustic based lossy codecs are most relevant to the present work. The general principle of operation, and the details of the popular MPEG-1 family of codecs will now be discussed.

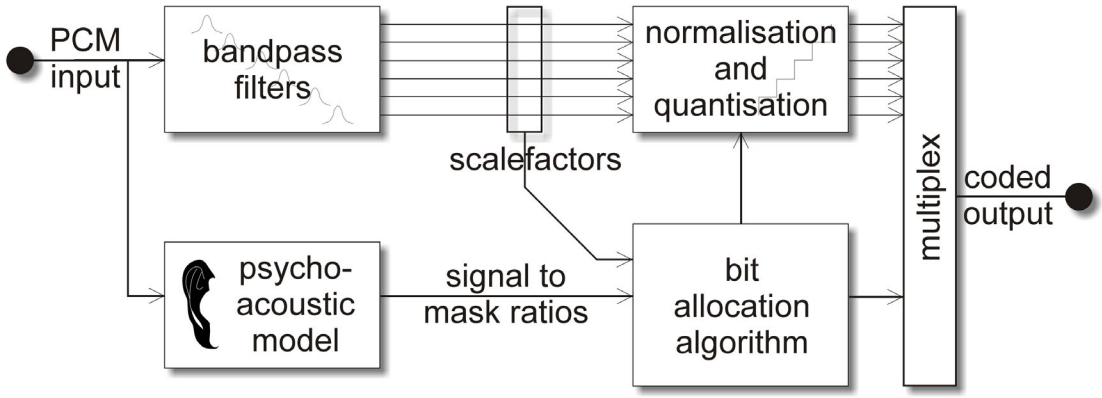


Figure 2.2: General structure of a psychoacoustic codec

2.2.4 General psychoacoustic coding principles

A generalised psychoacoustic codec may operate as shown in Figure 2.2. In the first stage of the encoder, the incoming signal is split into several frequency bands by a bank of bandpass filters. A psychoacoustic model calculates the masked threshold for each frequency band, and this is converted into a Signal to Mask Ratio (SMR) for each band. Spectral components that lie above the masked threshold are judged to be audible, and yield a positive Signal to Mask Ratio. Spectral components that lie below the masked threshold are judged to be inaudible, and yield a negative Signal to Mask Ratio.

The Signal to Mask ratio directs a bit allocation algorithm. The number of bits allocated to each frequency band determines the accuracy of the quantiser, which in turn determines the amount of noise that will be added within each band. The intention is to add noise within masked spectral regions of the audio signal, but not to change or distort audible spectral components.

The amplitude of the signal in each band is normalised to unity *before* quantisation, and the scale factor required to revert the signal to its original level is stored, along with the output of the quantiser. The scale factor and/or quantiser output for a given band may be omitted if the signal within the frequency band lies well below the masked threshold. The resulting bitrate is much less than that of the original audio signal.

The decoder reverses this process by generating the signal in each band from the quantised values, multiplying each signal by the appropriate scale factor, and bandpass filtering the

contents of each band. Finally, outputs of all the frequency bands are summed to yield the final decoded audio signal. Hopefully, the decoded signal will sound almost identical to the original signal.

The accuracy of the psychoacoustic model will effect the perceived sound quality of the coded audio. If the model incorrectly predicts that a spectral component is inaudible, when in reality is it above the masked threshold, then a human listener will perceive the noise added by the codec within this frequency region. However, even if the psychoacoustic model perfectly predicts human perception, the resulting coded audio signal will still contain audible noise if the bitrate is too low. In a constant bitrate compressed audio signal, only a certain number of bits are available per second. If the psychoacoustic model calculates a high Signal to Mask Ratio for many frequency bands, this may instruct the bit allocation model to use more bits than are available. In this case, the bit allocation model must choose the best compromise to minimise the audible coding noise, whilst remaining within the allocated bitrate. Variable bitrate coding overcomes this problem, by allocating the correct number of bits to ensure that the quantisation noise within each frequency band is below the masked threshold. This will reduce the bitrate during quiet or easy to encode passages, whilst increasing the bitrate during loud or complex passages. Variable bitrate encoding is only available within some audio codecs.

There are two sub-types of psychoacoustic codec: *subband* codecs and *transform* codecs. Subband codecs store the waveform present in each frequency band in a sub-sampled, quantised form. Transform codecs perform a time to frequency transformation (e.g. the Fast Fourier Transform) upon the original audio signal, or the signal within each frequency band. The resulting transform coefficients are stored, after quantisation, according to the SMR prediction of the psychoacoustic model. Transform codecs typically offer greater bitrate reduction than subband codecs. This is partly due to the higher frequency resolution offered by the transform, which allows the coding noise to be distributed more accurately according to the masked threshold. The major disadvantage of transform coding is that all current time to frequency transformations process the audio in discrete time domain blocks, and this blocking can cause audible problems. These problems will be discussed in Section 2.2.5.3, with respect to the MPEG-1 layer III codec.

2.2.5 MPEG audio codecs

These general principles of audio coding are seen at work in the MPEG-1 family of audio codecs. The MPEG-1 standard consists of three "layers" of coding, where each layer offers an increase in complexity, delay, and subjective performance with respect to the previous layer. The higher layers build on the technology of the lower layers, and a layer n decoder is required to decode all lower layers. The MPEG-1 standard [ISO/IEC 11172-3, 1993] supports sampling rates of 32 kHz, 44.1 kHz and 48 kHz, and bitrates between 32 kbps (mono) and 448 kbps (Layer I stereo). The MPEG-2 standard [ISO/IEC 13818-3, 1998] contains a backwards compatible multi-channel codec, and extends the range of allowed bitrates and sampling rates[4]. A proprietary extension called MPEG-2.5 [Dietz *et al*, 1997] is in common use for layer III. The sampling rates and bitrates are summarised in the following table.

codec	sampling rates / kHz	allowed bitrates / kbps
MPEG-1	32, 44.1, 48	
layer I		32, 64, 96, 128, 160, 192, 224, 256, 288, 320, 352, 384, 416, 448
layer II		32, 48, 56, 64, 80, 96, 112, 128, 160, 192, 224, 256, 320, 384
layer III		32, 40, 48, 56, 64, 80, 96, 112, 128, 160, 192, 224, 256, 320
MPEG-2	16, 22.05, 24	
layer I		32, 48, 56, 64, 80, 96, 112, 128, 144, 160, 176, 192, 224, 256
layer II		8, 16, 24, 32, 40, 48, 56, 64, 80, 96, 112, 128, 144, 160
layer III		8, 16, 24, 32, 40, 48, 56, 64, 80, 96, 112, 128, 144, 160
MPEG-2.5	8, 11.025, 12	
layer III		8, 16, 24, 32, 40, 48, 56, 64, 80, 96, 112, 128, 144, 160

Table 2.2: Allowed bitrates in the MPEG audio coding standards

[4] The MPEG-2 standard also defines a non-backwards compatible codec known as MPEG-2 AAC (Advanced Audio Coding). This section of the standard was finalised some years after layers I, II, and III. It includes several refinements that improve coding efficiency (most notably temporal noise shaping), but the general coding principles are very similar to MPEG-1 layer III. Further details can be found in the standards document and an excellent description appears in [Bosi *et al*, 1997].

A review of the MPEG standards for audio coding is found in [Brandenburg and Bosi, 1997], and a clear description of layer III and AAC coding is contained in [Brandenburg, 1999]. Parts of the following explanation are drawn from [Hollier, 1996].

2.2.5.1 MPEG-1 layer I audio coding

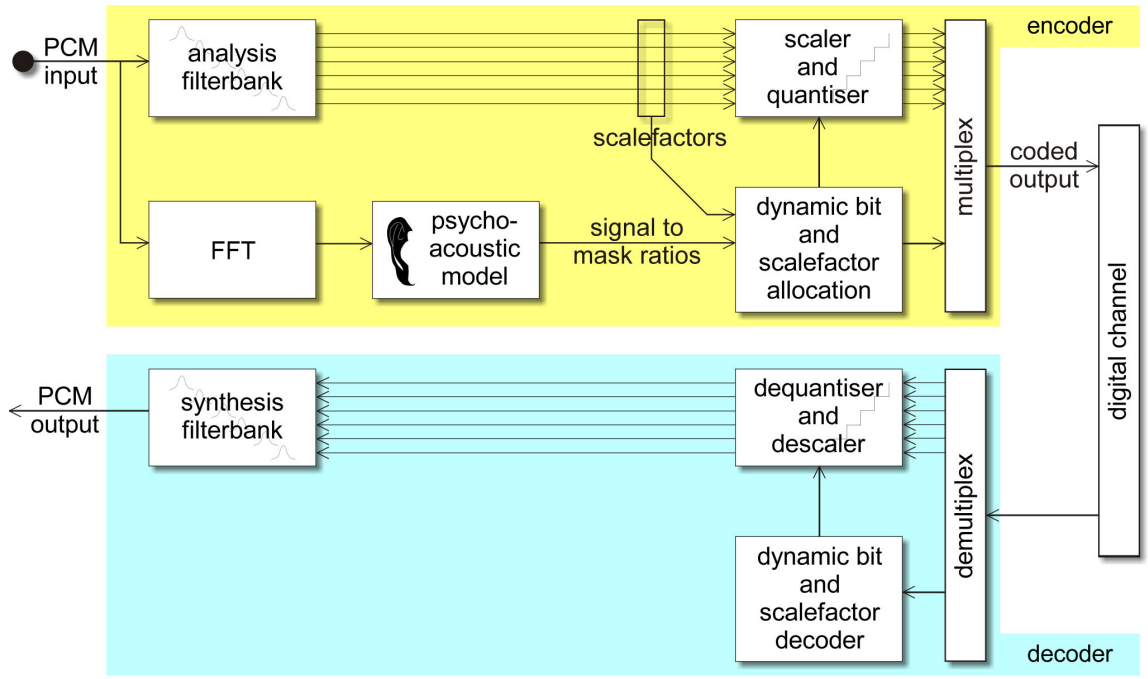


Figure 2.3: Structure of MPEG-1 audio encoder and decoder, Layers I and II

The structure of the MPEG-1 layers I and II encoder is shown in Figure 2.3.

The operation of the layer I encoder is as follows. All references to time and frequency assume 48 kHz sampling.

1. The **analysis filterbank** splits the incoming audio signal into 32 spectral bands. The filters are linearly spaced, each having a bandwidth of 750 Hz.
2. The samples in each band are **critically decimated**, and split into blocks of 12 decimated samples. **Scalefactors** are calculated which normalise the amplitude of the maximum sample in each band to unity.
3. In a parallel process, the signal is **windowed**, and a 512-point **FFT** is performed, to calculate the spectrum of the current audio block.

4. The **psychoacoustic model** calculates the masked threshold from the spectrum of the current block. This is transformed into a Signal to Masker Ratio for each band.
5. The **dynamic bit and scalefactor allocator** selects one of 15 possible quantisers for each band, based upon the available bitrate, the scalefactor, and the masking information. The aim is to meet the bitrate requirements whilst masking the coding noise as much as possible.
6. The **scaler and quantiser** acts as instructed by the allocator, to scale and quantise each block of 12 samples.
7. Finally, the quantised samples, scalefactors, and control information are **multiplexed** together for transmission or storage.

The **decoder** unpacks this information, scales and interpolates the quantised samples as instructed via the control information, and passes the 32 bands through a synthesis filter to generate PCM audio samples. The decoder does not require a psychoacoustic model, so decoder complexity is reduced compared to the encoder. This is useful for broadcast applications, where a single (expensive) encoder must transmit to thousands of (inexpensive) decoders.

The decoder is specified exactly by the MPEG standard, but the encoder can use any coding strategy that yields a valid bitstream. For example, the psychoacoustic model may be arbitrarily complex (or non-existent if encoding speed is the only concern). In theory, this allows future developments in psychoacoustic knowledge to be incorporated into the encoder, without breaking compatibility with existing decoders. In practice, the fixed choice of filterbank parameters limits the fine-tuning that may be carried out.

2.2.5.2 MPEG-1 layer II

The layer II codec operates in a similar manner to layer I, but achieves higher audio quality at a given bitrate via the following modifications.

1. The 512-point FFT is replaced by a 1024-point FFT. This increases the frequency resolution of the masking calculation, at the expense of increasing the encoder delay.
2. The similarity between adjacent scalefactors in adjacent blocks is exploited, thus reducing the amount of control information that must be transmitted.
3. More accurate (smaller stepped) quantisers are made available.

MPEG-1 layer II coding is used by Digital Audio Broadcasting within the UK and much of the world (apart from America). It achieves near CD-quality at around 256 kbps stereo.

2.2.5.3 MPEG-1 layer III

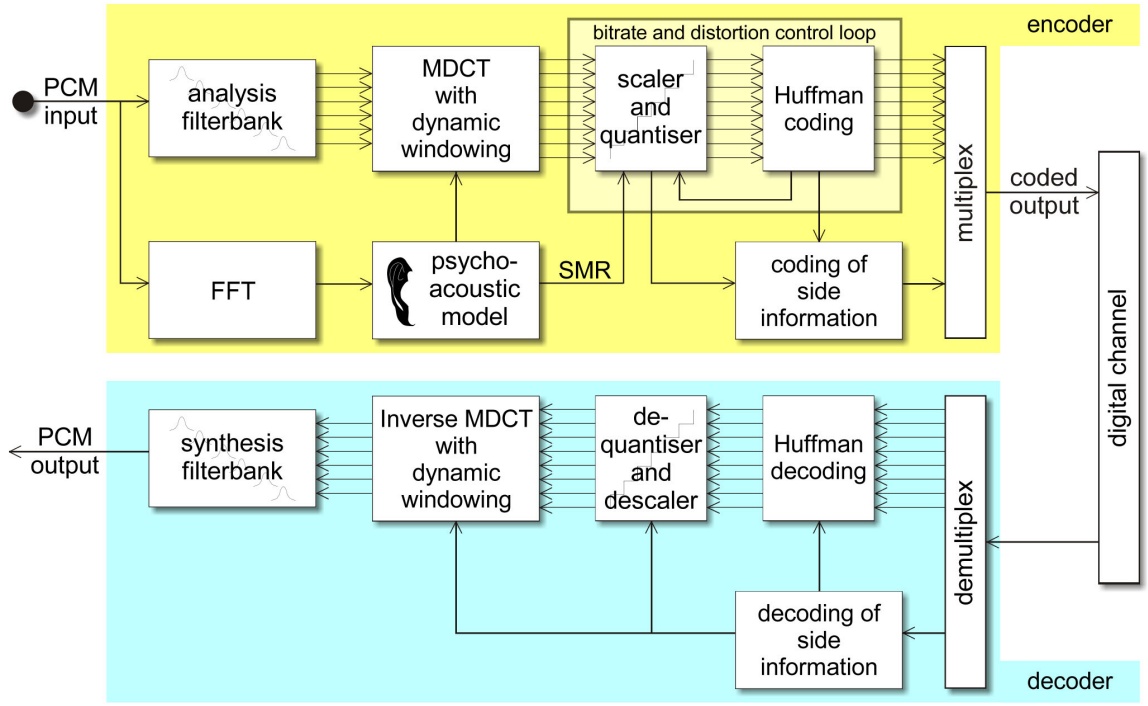


Figure 2.4: Structure of MPEG-1 layer III audio encoder and decoder

The layer III codec is significantly more complex than the lower layers. It uses both subband and transform coding, and is the only layer with mandatory support for variable bitrate coding. The layer III encoder is shown in Figure 2.4.

Each of the 32 frequency bands is sub-divided by a 6-point or 18-point Modified Discrete Cosine Transform. This gives a possible frequency resolution of up to 42 Hz, compared to 750 Hz for layers I and II. The layer III codec switches between the two possible MDCT lengths (often referred to as short and long blocks) depending on the input signal. This strategy is useful because, after quantisation of the coefficients, the temporal structure of the audio information within the MDCT block is often distorted. Hence, short blocks are used for encoding transient information to minimise audible temporal smearing, while long blocks are used for near steady-state signals to give increased spectral accuracy.

Three other significant improvements are included in the layer III encoder. A non-uniform quantiser is used to increase the effective dynamic range (in a similar manner to A-law or μ-law encoding, but operating upon a single frequency band). The quantised samples are

losslessly packed using Huffman coding. Finally, a bit reservoir is included in the layer III specification. This allows the encoder to increase the bitrate during brief “hard to encode” sections, so long as it can reduce the bitrate during a nearby “easy to encode” section. The overall bitrate is held constant, so the scheme is still referred to as “constant bitrate”. In this manner, the reservoir provides some of the advantages of variable bitrate coding, whilst maintaining compatibility with fixed bitrate transmission channels.

The layer III decoder is more complex than that required for layers I or II. However, the popularity of MPEG-1 and -2 layer III has led to low-cost single chip layer III decoders becoming available. Layer III is said to offer near CD quality at 128 kbps.

Many of the intricacies of the MPEG-1 layers are not covered here. Example encoders and decoders are described in the appropriate standards documents ([ISO/IEC 11172-3, 1993] and [ISO/IEC 13818-3, 1998]). One important feature is relevant to the present work, and is discussed in the next section.

2.2.5.4 Joint stereo coding

The redundancy sometimes found within two channel (stereo) signals allows for a significant bitrate reduction without a corresponding reduction in audio quality. MPEG-1 defines four modes:

1. Mono
2. Stereo
3. Dual (two separate channels)
4. Joint Stereo

In the first three modes, one or two separate channels are coded individually. In the fourth mode, the information in the two stereo channels is combined in one of two possible ways to reduce the bitrate.

Intensity stereo coding takes advantage of the human ear’s insensitivity to interaural phase differences at higher frequencies.

For each frequency band, the data from the two stereo channels is combined, and the resulting single channel of audio data is coded. Two coefficients are also stored to define the level at which this single channel should appear in each of the stereo channels upon decoding. This

procedure is only appropriate at higher frequencies, but it can offer a 20% bitrate saving compared to normal stereo. Unfortunately, the use of intensity stereo can be audible. Though the ear cannot detect the interaural phase of high frequency tones, the ear can detect interaural time delays in the envelope of high frequency signals. These time delays are destroyed by intensity stereo coding, and the stereo image appears to partially collapse. However, this effect is less objectionable than highly audible coding noise, so intensity stereo is useful at low bitrates, where it effectively frees some bits to reduce the coding noise.

Matrix stereo coding exploits the similarity between two stereo channels. Rather than coding the Left and Right Channels, the Sum (or "Middle") and Difference (or "Side") signals are coded instead, thus:

$$M = \frac{L + R}{\sqrt{2}} \tag{2-1}$$

$$S = \frac{L - R}{\sqrt{2}} \tag{2-2}$$

$$L = \frac{M + S}{\sqrt{2}} \tag{2-3}$$

$$R = \frac{M - S}{\sqrt{2}} \tag{2-4}$$

The transformation from L/R to M/S is entirely lossless and reversible via equations (2-3) and (2-4), though quantisation of the M/S signals will prevent perfect reconstruction in practice. For a signal with very little difference between the two stereo channels (i.e. an "almost" mono signal) the energy within the S channel is minimal, and the bitrate required for this channel is comparatively low. Thus, for a mono or 100% out of phase signal, the bitrate reduction is nearly 50%. For most audio signals, some bitrate reduction may be achieved by the use of joint stereo. It offers no benefit where the two stereo channels are completely uncorrelated. In some circumstances, it may cause problems.

For example, consider a stereo signal consisting of audio on the left channel only, with an *almost* silent right channel. The right channel may contain a hiss, or a quiet echo. The M and S

channels will be *almost* identical. However, the difference between the two channels is enough to ensure that the coding noise introduced into each channel is not identical. This coding noise is masked in both channels of the M/S representation. When the left and right channels are restored in the decoder, the right channel consists of the difference between the M and S signals. Hence, the right channel will contain very little signal information, but lots of coding noise. This occurs because the signal that masked the coding noise in the M/S representation is spatially separated from the coding noise in the decoded L/R output.

MPEG-1 layer III can use a combination of stereo techniques, in which the encoder switches dynamically between independent stereo, matrix stereo, and/or intensity stereo, depending on the incoming audio signal and the desired bitrate. This is yet another reason why layer III can achieve higher quality at a specified bitrate, or a lower bitrate at a given quality than layers I and II.

It is interesting to note the target bitrates of the three layers. The specifications suggest that layers I and II achieve CD quality at 256 kbps stereo; layer II at 192 kbps joint stereo, and layer III at 112-128 kbps joint stereo. Experience suggests that these recommendations are less than exact. Some audio signals are audibly degraded by some or all of the layers at *any* bitrate. Further, the suggested bitrate for layer III is especially optimistic; nearly twice this bitrate is often required to ensure CD quality over a wide range of material. The majority of layer III encoders deliver a bandwidth of 15-16 kHz at 128 kbps, which is by definition not CD quality. Whilst many audio extracts do sound acceptable at 128 kbps, a significant minority do not.

It is necessary to objectively measure the sound quality of audio codecs in order to verify manufacturers claims, to monitor broadcast sound quality, and to improve encoder performance. In the next section, some common audio quality measurements are described, and their application to psychoacoustic audio codecs is discussed.

2.3 Audio quality measurements

If a human listener auditions an audio device, and expresses an opinion that the device sounds “good” or “bad”, then this opinion represents a subjective judgement. Subjective assessment of perceived sound quality is very important, since an audio device that sounds subjectively “bad” is undesirable. However, subjective judgements are notoriously unreliable. The placebo effect often causes human listeners to perceive “audible” differences, even where there are none.

Two different listeners may not share the same opinion. In addition, subjective judgements carried out by the same listener on different days, or even in different moods, may contradict each other. Careful listening requires controlled conditions, and expert listeners, both of which are expensive to obtain. In summary, though the *subjective* audio quality of a device is of utmost importance, it is exceedingly difficult to quantify. For this reason, *objective* audio quality assessment is often preferred.

The audio industry has developed a variety of measurements over its hundred-year history. These measurements are objective and repeatable. They also give an *indication* of the perceived or subjective sound quality of the device under test. However, the relationship between the measured value and the perceived sound quality can be obscure, indirect, or even hidden due to differing measurement methods. Nevertheless, objective measurements such as the frequency response, signal to noise ratio, and total harmonic distortion, represent widely understood methods of quantifying the performance of an audio device.

Three audio quality measurements will be considered, and their application to the assessment of psychoacoustic based codecs will be discussed.

2.3.1 Frequency response

The frequency response of a device is defined as the gain or attenuation of the device as a function of frequency. Some measurement methods also produce the phase response as a function of frequency, though this is less often quoted. A graph of ideal frequency response is a straight horizontal line, indicating that all frequencies are passed equally by the device. Often the frequency response is quoted as a range of frequencies. This indicates that the response does not deviate from the mean by more than the specified amount (typically ±0.5 dB or ±3 dB) over this range. This is a useful and compact method of representing the frequency response for many audio components (e.g. amplifiers) which often have a flat response over the audible band, but attenuate very low and very high frequencies.

Several methods of measuring the frequency response exist, which rely on various signals being passed through the device. Possible signals include a swept frequency sinusoid, an impulse, or a maximum length sequence. The swept sinusoid will give the amplitude response directly as a function of the input frequency. The latter two methods require a Fourier trans-

form to be carried out upon output of the device in order to yield the amplitude and phase responses.

When measuring the frequency response of a conventional audio device, all methods yield similar results. One possible exception is the measurement of a loudspeaker's response within a real listening room, where standing waves can cause problems at low frequencies with tonal test signals. However, in general, the frequency response measurement acts as intended.

Ideally, a psychoacoustic audio codec should have a flat frequency response, though a low pass filter may be included at some high frequency. [Brandenburg, 1999] states that this is a positive design feature, since encoding high frequency inaudible signals wastes bits which could be used on lower frequency components. In addition, if the bitrate is constrained, reducing the bandwidth is preferable to adding large amounts of audible coding noise.

The frequency response measurement should provide this information about the audio codec. A tone sweep will reveal any fixed low pass filter, but may not reveal any dynamic low pass filtering that may come into play if the encoder "runs out of bits". The maximum length sequence stimulus consists of white noise, which is difficult to compress efficiently. Hence, this method of frequency response measurement may cause the encoder to activate any dynamic low pass filter, and this will be reflected in the frequency response measured by this method. Alternatively, a true random white noise signal may be fed into the encoder, and the spectrum may be calculated from the output of the decoder.

To evaluate the frequency response of an audio codec, both methods of frequency response measurement should be used. If the frequency response is flat up to a cut-off frequency, then the low pass frequency determined via each measurement is the only data that is required. If the frequency response is more complex, then a plot of amplitude against frequency obtained via each measurement may be appropriate.

To characterise an audio codec fully, further measurements are required. The most important aspect of the codec is the coding noise, which could be viewed as a type of non-linear signal dependent distortion. Three measurements which are appropriate for noise or distortion are now examined in turn.

2.3.2 Signal to Noise Ratio

If the RMS voltage of the maximum signal is V_S, and the RMS voltage of the background noise is V_N, then the signal to noise ratio, in dB, is given by:

$$SNR = 20\log_{10}\left(\frac{V_S}{V_N}\right) \tag{2-5}$$

The RMS noise voltage is measured in the absence of an input signal. A digital audio codec may easily have an infinite SNR, since a silent (digital zero) input signal will cause a silent decoded output signal. For this reason, the SNR measurement of a psychoacoustic digital audio codec is almost worthless. Where the codec does add constant noise, the SNR measurement will reflect this. However, almost all wide-band audio codecs can reproduce silent signals perfectly.

2.3.3 Total Harmonic Distortion (plus noise)

Where the device adds signal dependent distortion, this can be quantified by a THD+N measurement, as shown in Figure 2.5. A signal (usually a 1 kHz tone) is passed through the device. A notch filter centred on the signal frequency removes the test signal from the output of the device. The residue consists of the harmonic distortion plus noise. The THD may be specified in dB relative to the test signal, or as a percentage. If a pure THD measurement is required, the noise is measured in isolation, and subtracted from the THD+N value.

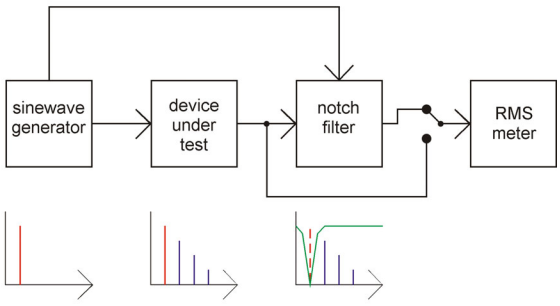


Figure 2.5: THD measurement technique

Most audio codecs do not exhibit significant harmonic distortion, though they do add much enharmonic distortion in the form of coding noise. Whereas the harmonic distortion components are found at integer multiples of 1 kHz above the test tone, the distortion added by the codec is centred on the 1 kHz tone. By definition, the THD includes only harmonic components. However, real world THD+N measurements of an audio codec will measure the coding noise, but the measured value will depend upon the characteristics of the notch filter. Hence, THD+N is not an accurate measurement of the coding noise.

Intermodulation distortion is another phenomenon that is often measured. However, like THD, it is less relevant to audio codecs because the signal components and the resulting distortion are at opposite ends of the audible spectrum, whereas the coding noise resides around the signal frequency.

2.3.4 Input Output difference analysis

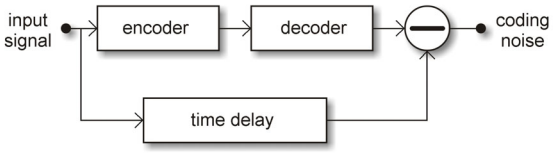


Figure 2.6: Input Output difference analysis

In a digital system, providing any delay due to the device is known and corrected for, the input signal can be subtracted exactly from the output signal, as shown in Figure 2.6. The residue consists of any noise and distortion added by the device. This technique may be used to determine the noise that is added by an audio codec in the presence of an input signal. If a test signal is applied, standard noise measuring techniques (e.g. [ITU-R BS.468-4, 1986] weighting followed by RMS averaging) may be used to calculate a single noise measurement. Alternatively, a Signal to Noise like Ratio may be computed, where the noise level is measured in the presence of the signal, rather than with the signal absent. This noise measurement may be used in equation (2-1), in place of V_N. The measurement is objective and repeatable.

Unfortunately, this measurement is almost useless for audio quality assessment. It is useless because the measured value does not correlate with the perceived sound quality of the audio codec. In fact, the noise measurement gives no indication of the *perceived* noise level.

The problem is that the noise measurement is quantifying inaudible noise. An audio codec is *designed* to add noise. The intention is to add noise within spectral and temporal regions of the signal where it cannot be perceived by a human listener. Subtracting the input signal from the output of the codec will expose this noise, and the noise measurement will quantify it. If the inaudible noise could somehow be removed from the measurement, then the resulting quantity would match human perception more accurately, since it would reflect what is audible. This task is complex, and many other approaches have been suggested which avoid this task. Some of these approaches, and the reasons why they are inappropriate, are discussed below.

A measurement of coding noise will include both audible and inaudible noise. Many analyses assume that all codecs will add equal amounts of inaudible noise. If this is true, then the codec that adds the most noise will sound worst, since it must add the most audible noise. However, a good codec may add a lot of noise, but all the noise may be masked. This codec will cause no audible degradation of the signal. Conversely, a poor codec may add only a little noise, but if the noise is above the masking threshold, then the codec will sound poor to a human listener. Hence, this approach is flawed, because the basic assumption is incorrect.

Many codec analyses published on the World Wide Web include plots of the long-term spectrum of the signal and coding noise. This approach assumes that where the coding noise lies below the signal spectrum, it will be inaudible, and where the noise is above the signal spectrum, it will be audible. Unfortunately, these assumptions are false. Noise above the signal spectrum may be masked, because masking extends upwards in the frequency domain. Noise below the signal spectrum may be audible, because the spectrum must be calculated over a finite time (ranges from 1024 samples to three minutes have been encountered). Hence, the signal that apparently masks the codec noise may not occur at the same time as the noise itself. This is especially true for sharp attacks, where many encoders generate audible pre-echo before the attack. This pre-echo is below the spectral level of the attack, so appears “masked” using this mistaken analysis method.

The problem with all these techniques is that they side-step the basic problem: it is necessary to determine which noise components are audible, and which are inaudible, before the audible effect of the codec upon the signal may be quantified.

In essence, the historical measurements that are discussed above are useful where an audio device is designed to change the signal as little as possible. However, audio codecs are designed to alter the signal significantly, but in a manner that is inaudible to a human listener. For this reason, a human listener must be the ultimate judge of the quality of an audio codec.

Subjective human opinion is notoriously unreliable. If it is to act as the ultimate judge, and provide a reliable quantitative indication of perceived audio quality, then some rigorous procedure must be employed. Such a procedure is described in the next section.

2.4 Subjective assessment

An international standard exists for "the subjective assessment of small impairments in audio systems" [ITU-R BS.1116-1, 1997]. The audible noise added by a psychoacoustic codec usually falls within the definition of a "small impairment". This includes all codecs that aim to be "transparent", where the difference between the original and coded audio signals may or may not be audible. If the perceived degradation due to the codec is large enough to be obvious to untrained listeners, then another assessment standard is appropriate. The MUSHRA standard, (proposed by the EBU [EBU, 2000], and now undergoing ratification by the ITU as [ITU-R draft BS.6/106, 2001]), addresses the assessment of medium quality audio codecs, where the audible impairment is obvious.

There are two reasons for the existence of two separate standards. Firstly, BS.1116 is very time consuming, mainly because it is necessary to prove that an audible impairment actually exists, before any quantification of the impairment can be considered valid. Within the MUSHRA testing procedure, it is assumed that the impairment is audible, which reduces the testing time considerably. Secondly, at some bitrates, it is impossible to achieve near CD quality, and codecs operating at these bitrates cannot be expected to sound transparent. Nevertheless, where the bitrate is severely limited it is useful to know which audio codec offers the best audio quality. Testing low bitrate codecs using BS.1116 will yield a set of results that are clustered within the worst quarter of the impairment scale, and the differences between codecs will be inaccurately represented. MUSHRA allows listeners to compare codecs directly, thus giving a more accurate prediction of relative quality of each codec.

There is a region of overlap between the two testing methodologies. However, all medium-to-high quality audio codecs are suitable for assessment via BS.1116, and this testing procedure is examined in detail.

In order to make all tests as "equal" as possible, BS.1116 contains strict guidelines on the following issues: choice of listeners, testing method, impairment scale, statistical analysis of results, listener training, choice of test material, specifications of loudspeakers and headphones, characteristics of listening environment including size, loudspeaker positioning, background noise, reverberation time etc, and listening level.

The choice of listeners is very important. All the listeners involved in the test should have experience in assessing audio quality, and must have demonstrated the ability to discern the types of audio impairment that will be auditioned in the test. All listeners must be trained extensively before the test is carried out. The results from a listener may be rejected after the test, if their results demonstrate that they were unable to reliably detect impairments. Approximately 20 listeners are desired, though fewer numbers are often used in practise.

The choice of test signal is critical. It is known that the noise added by audio codecs is significantly more audible within certain audio signals. It is suggested that appropriate material should be selected via an initial listening test. Only the most challenging audio extracts are suitable for use within a listening test. Extracts should be no longer than 1 minute. A duration between 15 and 30 seconds is ideal.

The test method is "double-blind triple-stimulus with hidden reference", as follows. A listener is presented with three audio extracts, identified as A, B, and C. The playback of these extracts is under the listener's control. They may listen to each extract as many times as they wish, and switch between extracts at will. Extract A is the original version, and extracts B and C are the original and coded versions, presented in a random order. The listener must identify whether B or C is the coded signal, and grade that extract on a scale from 1-4.9, shown in Figure 2.7. The listener *must* select either B or C as the coded version. This is to prevent conservative listeners from grading all extracts at 5.0, and is found to improve the accuracy of the test.

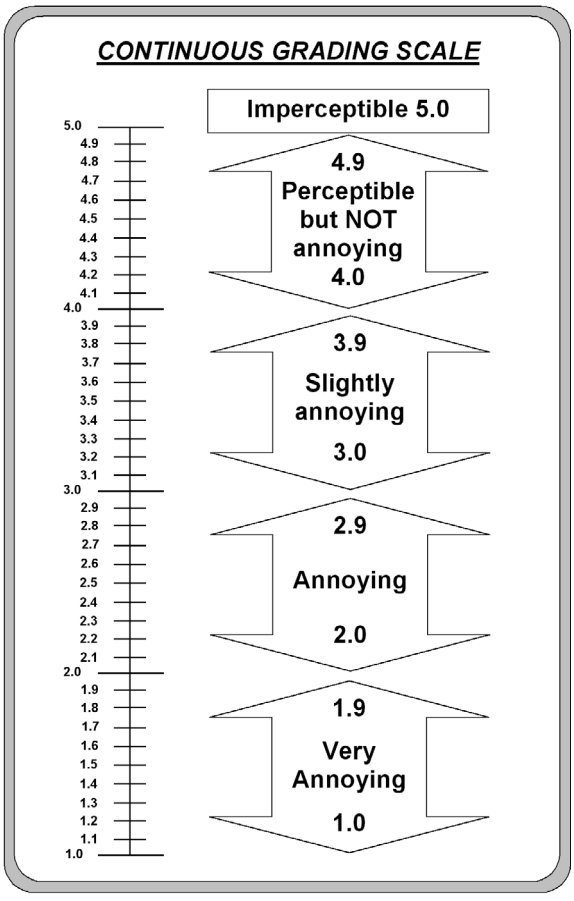


Figure 2.7: Scale used within BS.1116 (from [ITU-R BS.1284, 1997])

Where B or C is audibly different from A, but the coded version is preferred, this difference should still be assessed, and graded between 4.9 and 4.0 (perceptible but not annoying).

If the listener believes that B is the coded version, then by inference, C must be the original. Hence, C is implicitly given a score of 5.0. Some tests insist that the listener should grade both B and C explicitly, but since the listener is aware that one of B or C is the original signal, this is not strictly necessary. The raw score is not used as an indication of codec quality, but is transformed into a diffgrade, thus:

$$diffgrade = score_{coded} - score_{original} \tag{2-6}$$

where $score_{coded}$ is the score that the listener gave to the coded extract – **not** the score that the listener gave to the extract that they *believed* was the coded one. Thus, where the listener is mistaken in their choice, a positive diffgrade is generated, and where the listener correctly identifies the coded audio, a negative diffgrade is generated. The mean of all the diffgrades assigned to a given extract by all the listeners is a good indication of the perceived sound quality[5]. The transference of the scale in Figure 2.7 to the diffgrade scale (by subtracting 5 from each of the numbers) adds some descriptive information to the numerical diffgrade.

An excellent example of a BS.1116 test is reported in [Meares *et al*, 1998], and the first large-scale MUSHRA test is reported in [Stoll and Kozamernik, 2000].

2.4.1 Beyond subjective assessment

The BS1116 procedure is very time consuming, due to listener selection and training, and extract selection. The results are accurate, and surprisingly consistent, though some "anchor" extracts are sometimes required to ensure that the score-space is used correctly and consistently. However, the time and money involved prevents true BS.1116 compliant tests from being used in all situations where codec audio quality must be assessed.

In particular, it is very difficult to assess the audio quality of a codec "in service" within a broadcast chain using a subjective test. It is also painfully slow and expensive to use full scale BS.1116 subjective tests during the development cycle of an audio codec. For these reasons, it

[5] BS1116 strongly advises statistical processing to pre-filter rogue results, and ANOVA to determine the relationship between listener, codec, and audio extract. However, after the results have been filtered, the mean of the diffgrades across listeners does indicate the perceived quality of that codec/extract combination.

is desirable to develop an objective measurement that can mimic the performance of a human listener within a BS.1116 subjective test.

The objective measurement equipment must "listen" in a comparable manner to a human listener, comparing original and coded audio signals, and quantifying the *audible* difference between them. An ideal objective measurement would match the diffgrade, though more detailed information about the character of the audible differences would be useful in codec development.

An objective measurement tool which achieves these goals has been developed [Thiede *et al*, 2000]. At the heart of this tool is a psychoacoustic model that simulates human hearing. This tool, and many other psychoacoustic models, are discussed in Chapter 4.

Psychoacoustic models are based upon human hearing. Hence, before examining existing psychoacoustic models, the human auditory system will be discussed in detail in Chapter 3.

2.5 Conclusion

Psychoacoustic codecs reduce the amount of data required to represent an audio signal, by degrading masked regions of the spectrum. Conventional objective measurements of audio quality cannot predict the perceived quality of these codecs. Human listeners must be the final arbiters of quality judgements, and rigorous subjective test procedures have been prescribed which allow consistent qualitative assessments of audio quality to be carried out by expert listeners. However, it has been shown that this procedure is time consuming and expensive, and an objective alternative is required for many applications.

3

The Human Auditory System

3.1 Overview

In order to develop a model capable of simulating human hearing, it is desirable to understand the mechanisms present within the human auditory system (HAS). In this chapter the path of an audio signal is followed from free space, through the HAS, to its representation by neural impulses within the brain. At each stage along this path, it will be noted whether present knowledge is quantitative, qualitative, or theoretical. The sources of quantitative data and processing theories will be presented. Particular attention is given to the processes that give rise to the limits of human hearing, such as masking effects.

3.2 A walk through the human auditory system

Figure 3.1 (split across two pages) shows the main components of the human auditory system. The upper illustrations represent the physiology – the actual physical components that are present and identifiable within the human anatomy. The lower graphs indicate the functionality of each section. All frequency domain plots show amplitude in dB against log frequency. All time domain plots are linear on both scales. The illustration of the organ of corti at the top of Figure 3.1 (c) is taken from [Yates, 1995]; Figure 3.1 (d) is after [Patterson, WEB-1].

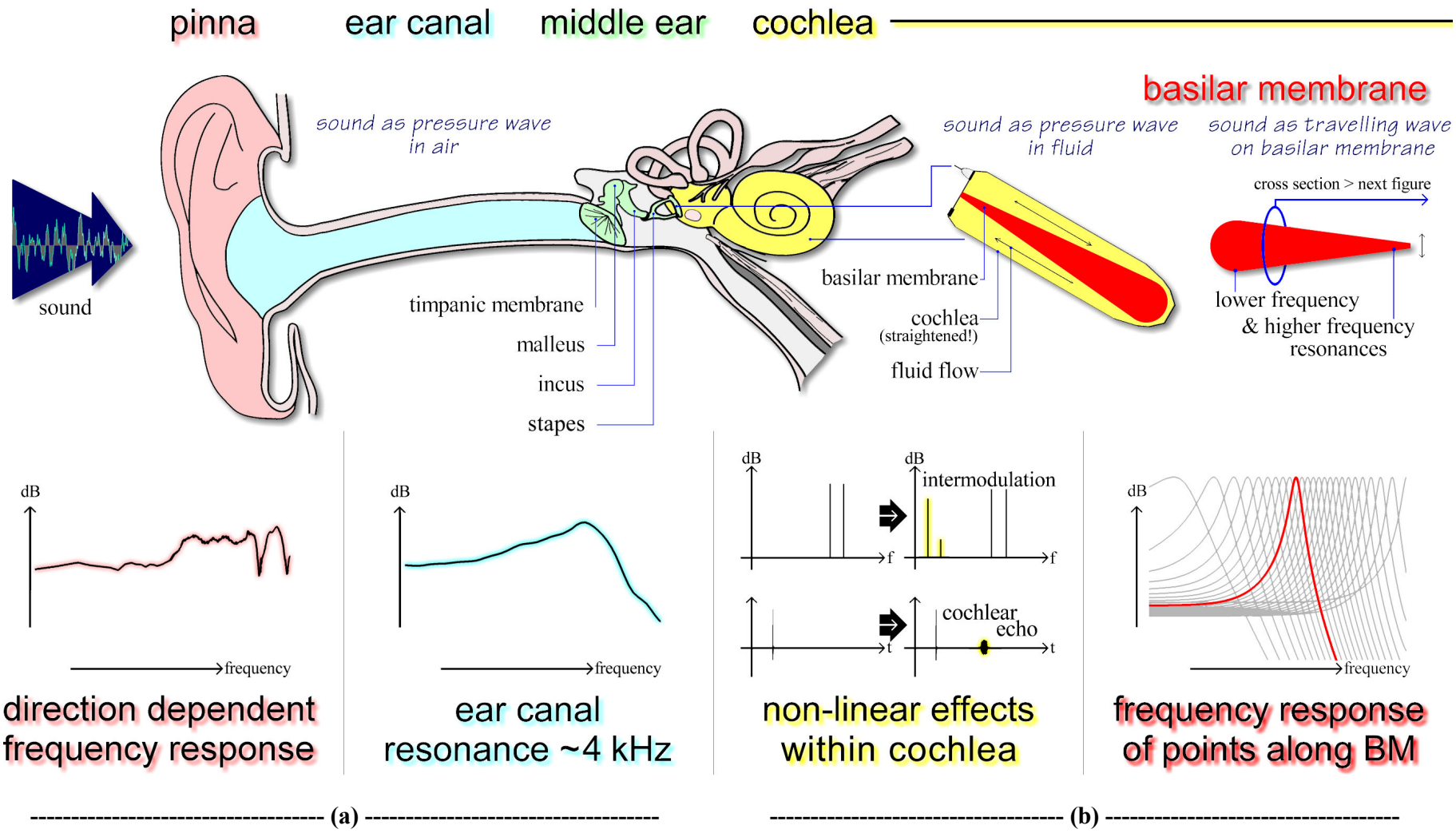


Figure 3.1: Signal path through the Human Auditory System

Upper Half: physiology Lower Half: function

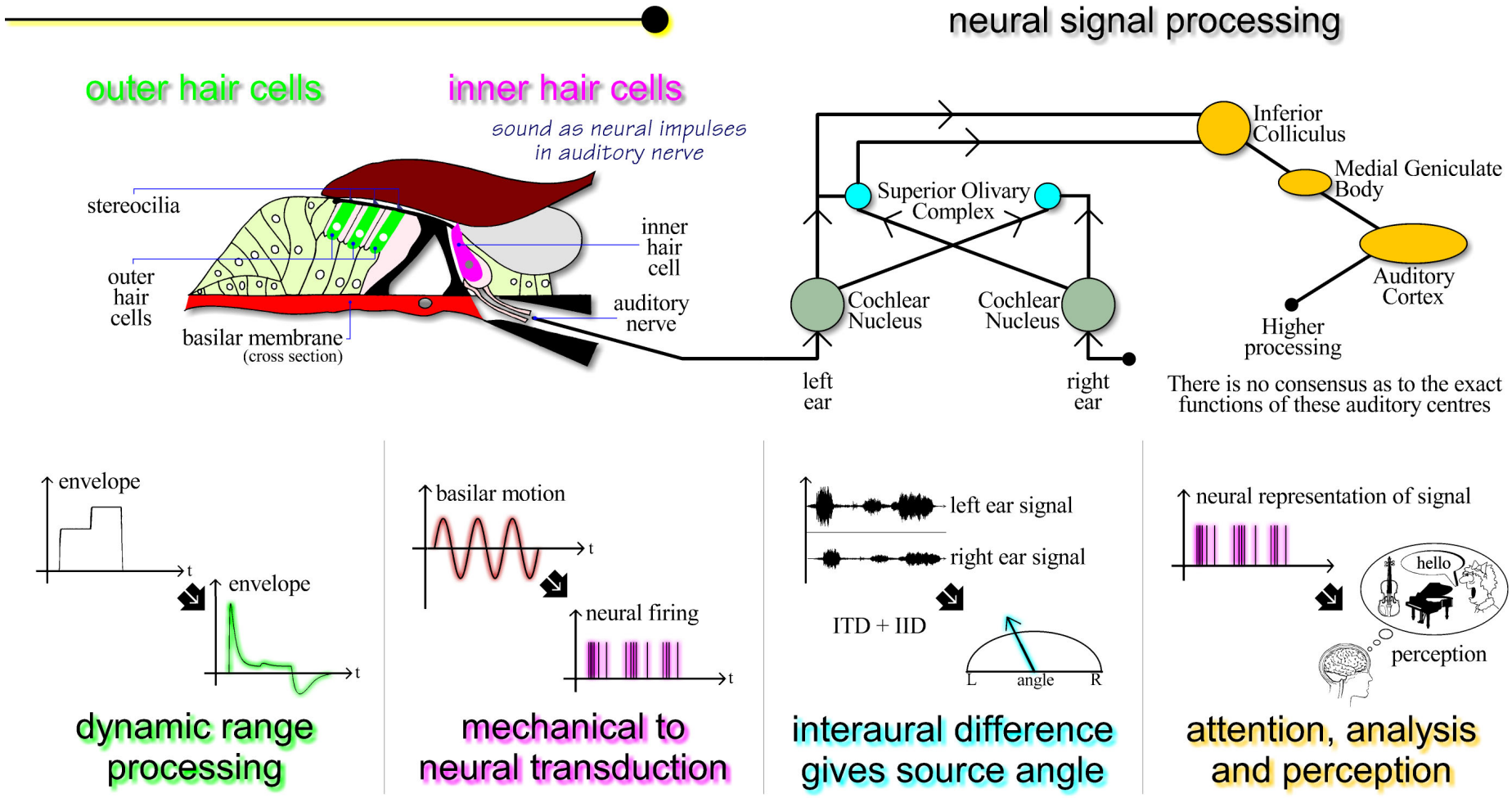


Figure 3.1: Signal path through the Human Auditory System

Upper Half: physiology Lower Half: function

Referring to Figure 3.1, the function of each section is as follows.

- The **pinna** is the flap of skin and cartilage found on each side of the human head. Some researches use the term **concha**. Together with the **ear canal**, it forms the **outer ear**. The function of the pinna is to direct sound into the ear canal. In doing so, the shape of the pinna causes the incoming signal to be filtered in a direction dependent manner.
- The **ear canal** forms a tube, linking the **pinna** to the **timpanic membrane**. The dimensions of the ear canal are such that it forms a resonance at around 2-5kHz, causing frequencies outside this range to be severely attenuated.
- The **timpanic membrane** (also known as the eardrum) lies at the base of the ear canal. The small ossicles bones known as the **malleus**, **incus**, and **stapes** act as an impedance-matching device, transmitting air-born sound pressure waves into fluid-born sound pressure waves within 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.

These processes will be considered in detail, before moving on to a discussion of the subsequent neural processing.

3.2.1 The Pinna

Until the late 19th century, the pinna was thought to be a simple funnel that directed sound into the ear. It is now known that the pinna is responsible for the ability to locate sounds in 3-D space. This ability is worthy of study in the present context, since all commercial systems which aim to create a virtual auditory environment rely on this property of human hearing to some extent. These include Dolby virtual surround, Lake theatre-phones, and any binaural [Robinson and Greenwood, 1998] or transaural [Foo *et al*, 1998] processing.

Humans can locate sounds in three dimensions. This is evident from the fact that sounds coming from outside our field of vision can be readily located. There are six mechanisms, or cues, by which humans locate sounds [Robinson and Greenwood, 1998], [Grantham, 1995]. These mechanisms will be examined by considering the sound that would reach a listener's ears from a source in front of them, slightly to their right.

1. If the source is nearer the right ear than the left, the sound will reach the closer ear first, yielding a time delay between the signals received at each ear. This is called the inter-aural time (or phase) difference, or ITD. Figure 3.2 [Green, 1976] shows a graph of inter-aural time difference against source angle in the horizontal plane. Note that the graph is symmetrical, so for each angular position in front of the listener, there is a corresponding position behind the listener that yields the same ITD.

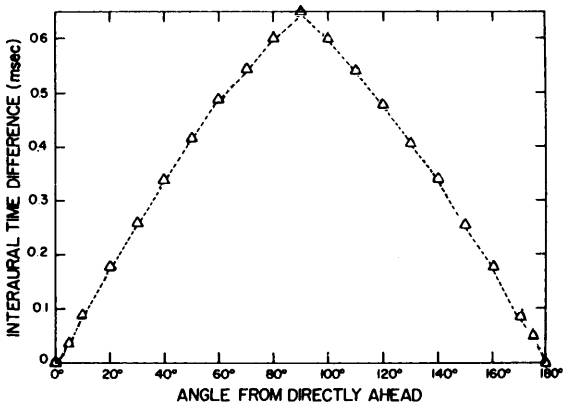


Figure 3.2: Variation of Interaural time difference with source angle

At around 1.5 kHz the wavelength of a pure tone is equivalent to the distance between the two ears. Hence, above half this frequency, the phase relationship between the signals arriving at each ear will be ambiguous, and it will be impossible for the auditory system to calculate a unique ITD. It is sometimes erroneously stated that the ITD provides no useful

localisation information for signals above 700Hz. However, this restriction only applies to pure tones. For complex signals containing only frequencies above 700Hz, the signal envelope may vary at a rate well below 700Hz, and hence the inter-aural time delay may be deduced with ease [Boerger, 1965].

2. The level of sound reduces as the square of the distance from the source ($1/r^2$ law). If the source is slightly nearer the right ear, the sound reaching that ear will be louder than that reaching the left. However, given the distance between the two ears, this is a small effect. More significantly, at higher frequencies, the head acts as a barrier to sound waves, so signals originating from one side of the listener will be attenuated at the opposite ear. This gives rise to the interaural level (or intensity) difference, or **ILD**.

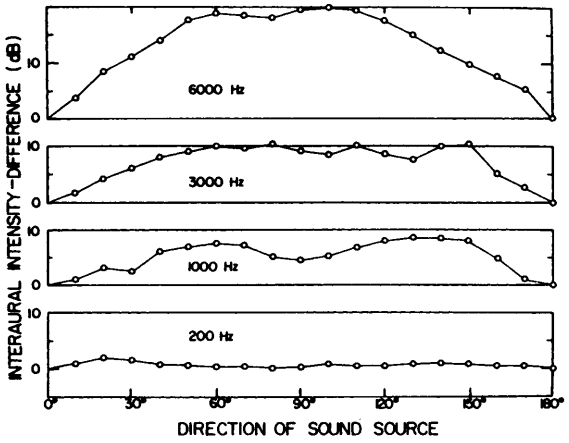


Figure 3.3: Variation of Interaural level difference with source angle

Figure 3.3 [Green, 1976] shows how the interaural level difference varies with source angle. Note that the effect is more pronounced for higher frequencies, where the head shadow effect is greatest.

3. The action of the pinna is third in order of importance. This will be discussed it in detail after the other cues.

4. If a listener is able to move their head, **all** the cues change dynamically, depending on the exact location of the source. This gives humans a more accurate bearing on the source location, and itself acts as a fourth localisation cue. It is often used subconsciously to resolve ambiguity in source location when all other cues are inconclusive.

The fifth and sixth cues are significantly weaker than the others.

5. In a reverberant environment (e.g. any normal room), the loudness of the direct sound from the source compared to the level of reverberation will give an indication of the source distance. Early reflections may also strengthen the ability of the HAS to perceive the distance of the source.

6. Finally, if the source is a great distance away, the large volume of air through which the sound must pass will attenuate high frequencies, yielding another distance cue.

If the listener keeps their head stationary, thus excluding cue four, **the remaining cues give no way of knowing whether the source is in front or behind**, or above or below the listener. Psychoacousticians talk of the "cone of confusion" which is the cone-shaped plane of possible locations inferred by the ITD cue alone [Mills, 1972]. Without further information, the auditory system has no way of accurately locating the sound source in real space. However, it is evident from everyday life, and has been proven experimentally, that humans are able to locate sounds without head movement, all-be-it with reduced accuracy.

This is possible because of the pinna. The pinna filters the incoming sound wave in a directionally dependent manner. Figure 3.4 shows the frequency response of the sound reaching the ear canal (via the pinna), from sources at various angles around a listener. The measurements in Figure 3.4 were made using a dummy head (a similar set of measurements are available from [Gardner and Martin, 1994]), but real human data are similar. These responses are called Head Related Transfer Functions (HRTFs) (see [Moller *et al*, 1995] for an extensive study). However, humans do not perceive this pinna filtering as changing the spectrum of the sound – the auditory system decodes the spectrum shape into spatial information.

For example, consider a sound source to the right of a listener, at a bearing of 45^{o}. There is a dip around 10 kHz in the HRTF for this angle, as shown in Figure 3.4. The auditory system detects this spectral signature, and decodes it into an estimate of source location. This process happens unconsciously, so the listener perceives a source at 45^{o}, and is unaware of the spectral coloration added by the pinna.

The surprising result is that humans do not need two ears to localise sounds. Blocking one ear proves that even without the ITD and ILD information, the spectral cues due to the pinna are sufficient to allow a human listener to localise sounds in 3-D space [Battaeu, 1967].

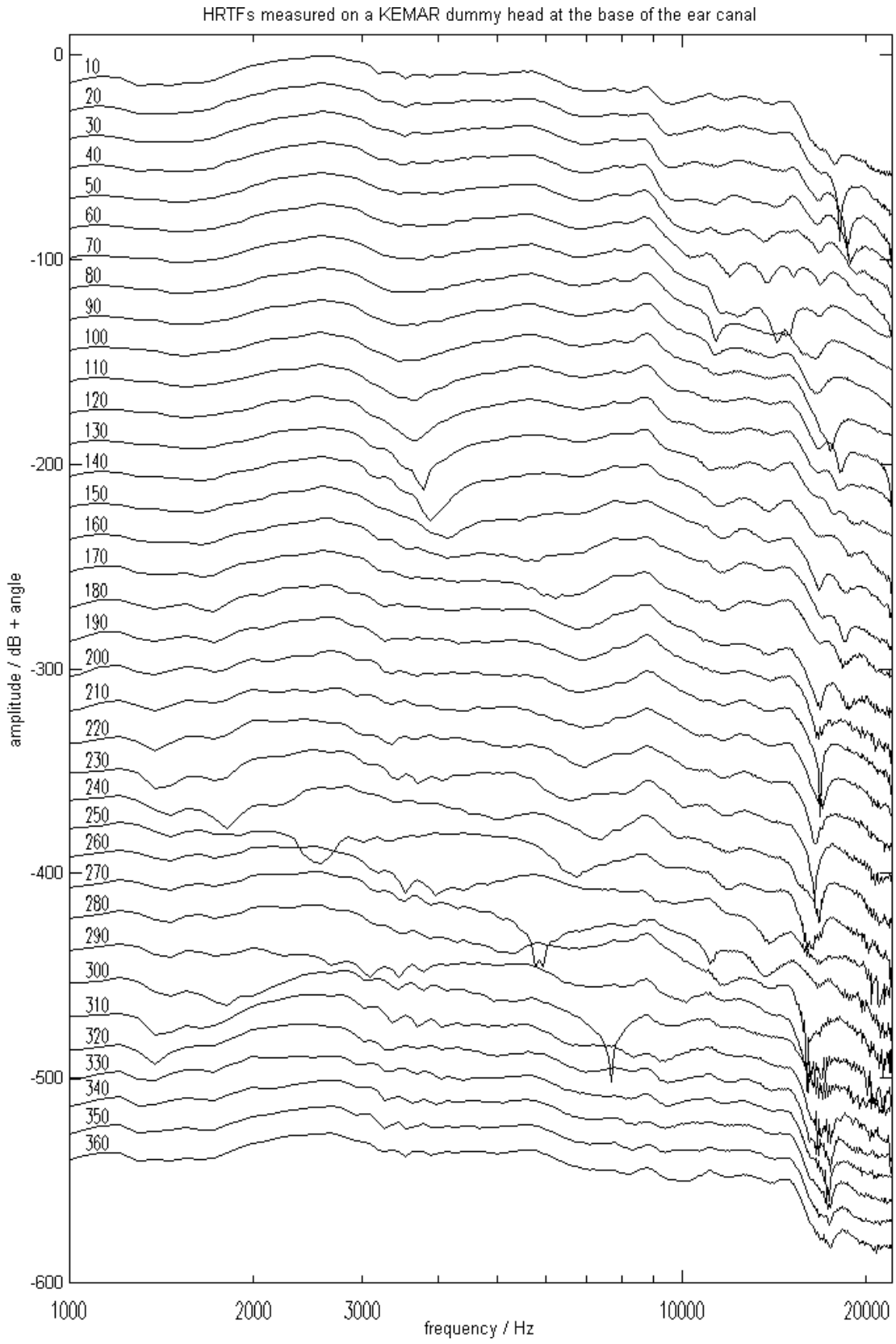
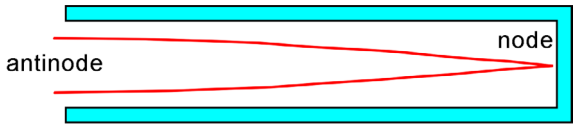


Figure 3.4: Head Related Transfer Functions of a dummy head

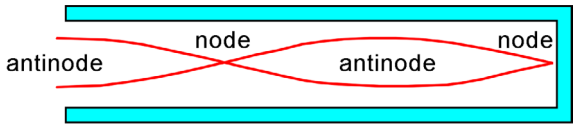
An interesting feature of human auditory localisation is that visual cues can be **stronger** than audio cues. Where the two contradict, the visual cue usually takes precedence. Though experiments in darkened rooms have proven that listeners can accurately locate sound sources to within ± 1°, a voice will often appear to originate from the lips of a person displayed on a T.V. or cinema screen, even though the sound originates from a loudspeaker several centimetres or even metres away. Multi-modal phenomena such as this are beyond the scope of the current research.

3.2.2 The ear canal

The resonant cavity between the outer and middle ear is called the ear canal (see Figure 3.1). It resonates in a similar manner to an organ pipe, since it is open at one end, and closed at the other, as shown in Figure 3.5 (after [Young, 2000]). The ear canal has a diameter of less than 1 cm, and is approximately 2.3 cm in length. Using this information, the resonant frequencies corresponding to the fundamental and 1st overtone can be calculated as follows.



(a) Fundamental



(b) 1st Overtone

Figure 3.5: Resonance of a closed pipe

The antinodes and nodes are displacement maxima and minima respectively.

An approximation to the speed of sound in air, v, is given by

$$v = 331 + 0.6T \tag{3-1}$$

where T is the temperature in degrees centigrade. The core temperature of the human body is typically 37 degrees centigrade, and the air within the ear canal is approximately 30 degrees centigrade. Hence, from equation (3-1) the speed of sound within the ear canal is approximately 350 ms^{-1}. If the length of the ear canal is given by l, then it can be seen from Figure 3.5 that the wavelengths of the fundamental and first overtone, λ_f and λ_o, are related to l thus:

$$l = \frac{\lambda_f}{4} \tag{3-2}$$

$$l = \frac{3\lambda_o}{4} \tag{3-3}$$

However, for any waveform

$$v = f\lambda \tag{3-4}$$

So the frequencies of the fundamental and 1st overtone are given by

$$f_f = \frac{v}{4l} \tag{3-5}$$

$$f_o = \frac{3v}{4l} \tag{3-6}$$

From this calculation, the frequency of the fundamental resonance is approximately 3.8 kHz, and the frequency of the first overtone is approximately 11.4 kHz. The measured frequency response of the ear canal matches this calculation, exhibiting peaks around 4 kHz and 12 kHz. See [Mehrgardt and Mellert, 1977] and [Moller *et al*, 1995a] for ear canal response measurements.

In summary, signal components at or near these resonant frequencies are boosted within the ear canal. This response contributes to the ears increased sensitivity to speech (mid-frequency) sounds. Figure 3.6 (from [Plack and Carlyon, 1995]) shows the level of sound at each frequency which humans *perceive* as being the same loudness. Examination of these

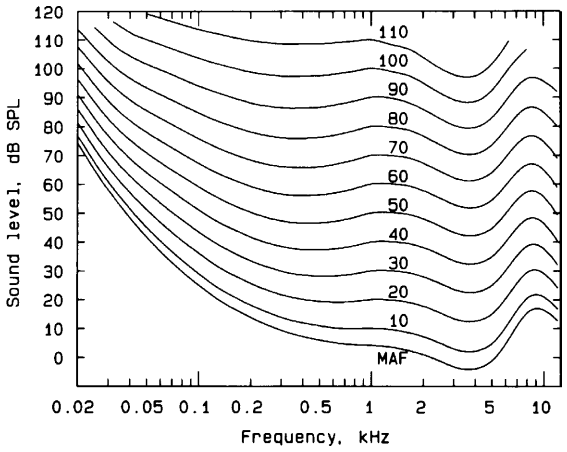


Figure 3.6: Equal loudness contours and Minimum audible field

The contours join sounds of equal *perceived* loudness. The reference for each contour is the perceived loudness of a 1 kHz tone at the SPL indicated on the graph. At all other frequencies, the contour indicates the SPL required for a sound of that frequency to match the perceived loudness of the reference tone.

curves shows the importance of the fundamental ear canal resonance, clearly visible as an increased sensitivity to sounds in the 2-6 kHz frequency region. The lower half of the resonance due to the 1st overtone is also visible at the extreme right of the graph.

3.2.3 Middle Ear

The **timpanic membrane** (eardrum), **malleus** (hammer), **incus** (anvil) and **stapes** (stirrup) transmit the sound pressure wave from the ear canal into the cochlea. These three ossicles bones act as a leverage system, converting the weak, high-amplitude oscillatory movement of the eardrum into a stronger, lower-amplitude force at the cochlea. This is transmitted into the cochlea by the pistonic action of the stapes against the oval window.

The frequency response and active non-linear properties of the middle ear are still open to debate. Historically, some researchers have assumed that the inner ear is equally sensitive to all frequencies. If this is true, then the changes in sensitivity with frequency within the HAS shown in Figure 3.6 are entirely due to the frequency response of the outer and middle ear. Other researchers (e.g. Zwicker and Fastl, 1990]) have suggested that internal noise at lower frequencies is entirely responsible for the ears decreased sensitivity to low frequency sounds. Recently, direct measurements of the middle ear transfer function in human cadavers [Puria *et al*, 1997] suggest that the middle ear transfer function gives rise to most of the variation shown in Figure 3.6. However, the outer ear, and (to a lesser extent) internal noise, both play a role. Thus, the averaged inverse of the Equal Loudness curves provides a good estimate of the outer and middle ear transfer function, especially at higher levels, where the effects of internal noise are less significant.

The recent measurements carried out upon dead subjects give no insight into active mechanisms which may or may not be present within living ears. Many of the non-linearities which have previously been attributed to the middle ear are now believed to be generated by processes *within* the cochlea. However, the leverage system within the middle ear is believed to stiffen and/or retract from the oval window in the presence of exceptionally loud sounds, so offering some protection to the delicate mechanisms found within the cochlea.

3.2.4 The Cochlea[1]

The fluid-filled cochlea is a coil within the ear, partially protected by bone. The sea water-like fluid vibrates with the incoming sound wave. The cochlea is semi-partitioned along its length by a thin flap called the basilar membrane.

3.2.4.1 The basilar membrane

The basilar membrane (BM) vibrates with the incoming sound, and acts as a spectrum analyser, spatially decomposing the signal into frequency components.

The movement of the BM is in the form of a travelling wave, as demonstrated by [von Békésy, 1960]. The wavefront travels from the oval window, down the BM towards the apex. The amplitude envelope of this travelling wave varies along the BM, as shown in Figure 3.7. The location of the maximum amplitude of oscillation marks a point of resonance, and the location of this resonance is dependent on the frequency of the incoming signal, as follows.

The stiffness of the BM decreases along its length. The resonant frequency of a point on the BM varies with stiffness, thus sounds of different frequencies cause different parts of the BM to vibrate.

At the base (near the stapes and oval window) the BM is narrow and stiff. The BM becomes broader and more flexible further into the cochlea (i.e. further along the BM). The differ-

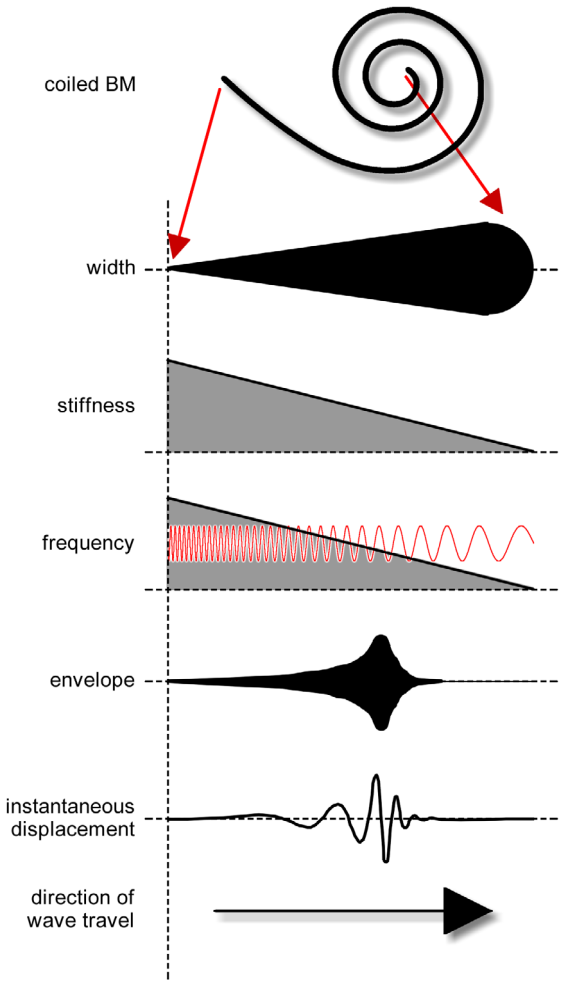


Figure 3.7: Properties of the basilar membrane

[1] A recent review of the structure and function of the cochlea can be found in [Yates, 1995].

ence from base to apex is a factor of about 3-4 times broader and 100 times more flexible. Higher frequencies cause oscillations near the base, where the BM is at its stiffest, and visa versa. This yields two interesting consequences.

3.2.4.1.1 Non-uniform frequency selectivity

The spacing of frequency resonances along the BM is not linear with frequency. For instance, if we take the point on the BM that resonates at 2 kHz, and measure the distance to the point that resonates at 4 kHz, we may find that it is 1 cm. If we now move a further 1 cm along the BM, we might expect to reach the 6 kHz resonant point. In fact we find a resonance of around 8 kHz. The resonant frequencies of various points along the BM are shown in Figure 3.8, from [Carterette, 1978]. These are known from probe experiments on mammalian cochlea, though some interpolation is necessary. The scale that relates the resonant frequency to position on the BM is called the Bark scale, or critical band scale, as shown in Figure 3.9. A formula for converting linear frequency to Bark scale is given in [Traunmüller, 1990]. Above 500 Hz, it is a reasonable approximation to a log scale [Zwicker and Terhardt, 1980].

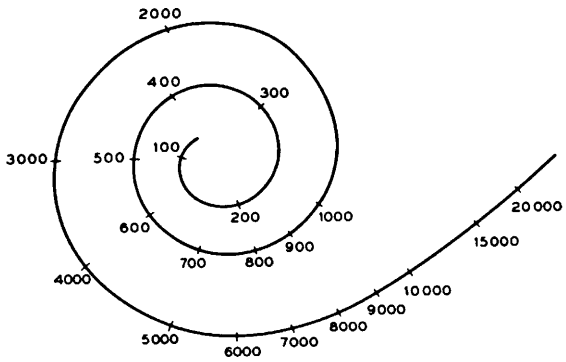


Figure 3.8: Position of resonant frequencies along the coiled length of the basilar membrane

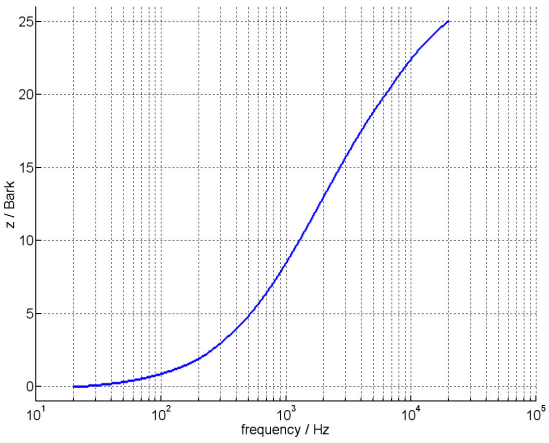


Figure 3.9: Relationship between frequency and critical band (bark) scale

At higher frequencies, this matches human pitch perception, which is logarithmic. Musical instruments are also tuned logarithmically. The reader is invited to listen to the demonstration of linearly spaced and logarithmically spaced tones included on the accompanying CD-ROM. Linear frequency steps sound very large at low frequencies, and very small at higher frequencies. Logarithmic frequency steps sound approximately equal across the audible frequency range.

Whilst there is a good correlation between frequency spacing on the BM and perception of pitch, this does not prove that pitch perception is entirely due to resonance location on the BM. This simplified view is correct for higher frequencies, but for lower frequencies, the firing rates of the hair cells upon the BM are more important. This will be discussed in Section 3.2.4.2.

3.2.4.1.2 Spectral masking

In the previous section it was shown that the mechanical structure of the BM causes it to act as a spectrum analyser, spatially decomposing the signal into frequency components. The auditory system can listen to the oscillation at any single point on the BM to the exclusion of all others. This gives the human auditory system the ability to filter out or ignore sounds away from the frequency of interest. The frequency resolution of such an analyser is limited by the width of the filter characteristic at each detection point. If the filters were infinitely narrow, then the presence of one spectral component would never impair the ability to detect another[2]. In reality, this is not the case, and the filters have a finite width.

The width of the filters varies with frequency. The relationship between resonant frequency and filter width is similar to the relationship between resonant frequency and filter spacing. The widest filters (giving the best time resolution, but poorest frequency selectivity) occur at high frequencies, whilst narrower filters (giving better frequency resolution) occur at lower frequencies.

[2] An infinitely narrow filter would also ring indefinitely, and introduce infinite delay.

Figure 3.10 shows the response at a particular point on the BM [Irino and Patterson, 1997]. The point on the BM graphed in black resonates at around 1 kHz, and it responds somewhat to all frequencies below the resonance, but rejects higher frequencies. Each point on the BM has a similar response, shifted up or down in frequency (shown in grey). If the BM is stimulated with a pure tone, the response is the excitation pattern show in Figure 3.11. The excitation pattern is the reversal of the frequency response. Mathematically, the frequency response (Figure 3.10) is convolved with a delta function (the pure tone), hence the reversal.

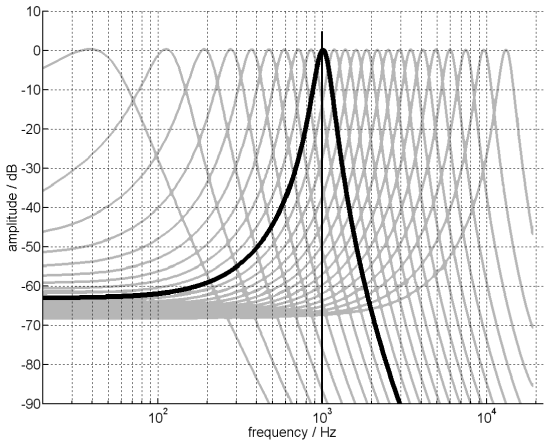


Figure 3.10: Frequency response of points on basilar membrane

The **black** curve shows, for a range of input frequencies (x axis), the amplitude of vibration of *one particular point* on the BM (y axis). The vertical black line indicates a 1 kHz pure tone. This excitation due to this tone is shown in the next figure.

To show this empirically, examine the responses of the resonances either side of the pure tone (i.e. the grey curves in Figure 3.10, either side of the vertical black line). The height at which they intersect the vertical black line corresponds to the amplitude of oscillation at that point on the BM. Note that all points *above* the tone are excited by it somewhat, but points significantly *below* it are not, which corresponds to the shape of the excitation curve. (This intuitive way of explaining the convolution was taken from [Hollier, 1996].)

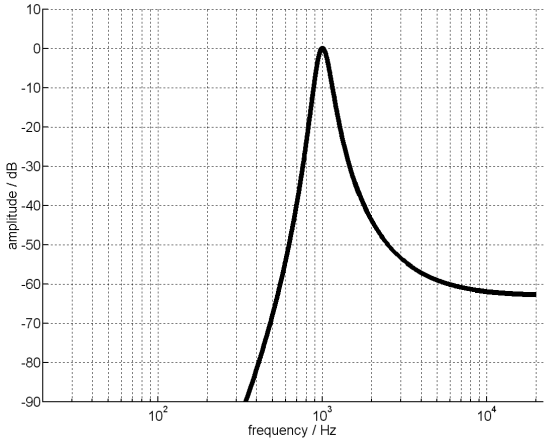


Figure 3.11: Excitation of basilar membrane due to a 1 kHz pure tone

This graph shows, for *each point* on the BM (x axis), the amplitude of vibration of each point (y axis) due to a pure tone.

Thus, a single pure tone excites a wide region of the BM. This is unsurprising; the BM is a continuous membrane, so it would be impossible for one point on it to move, whilst an adjacent point remained stationary, as that would cause it to rip or tear. Thus, even a pure tone (which is spectrally very narrow) must excite a finite region. This region extends upwards in

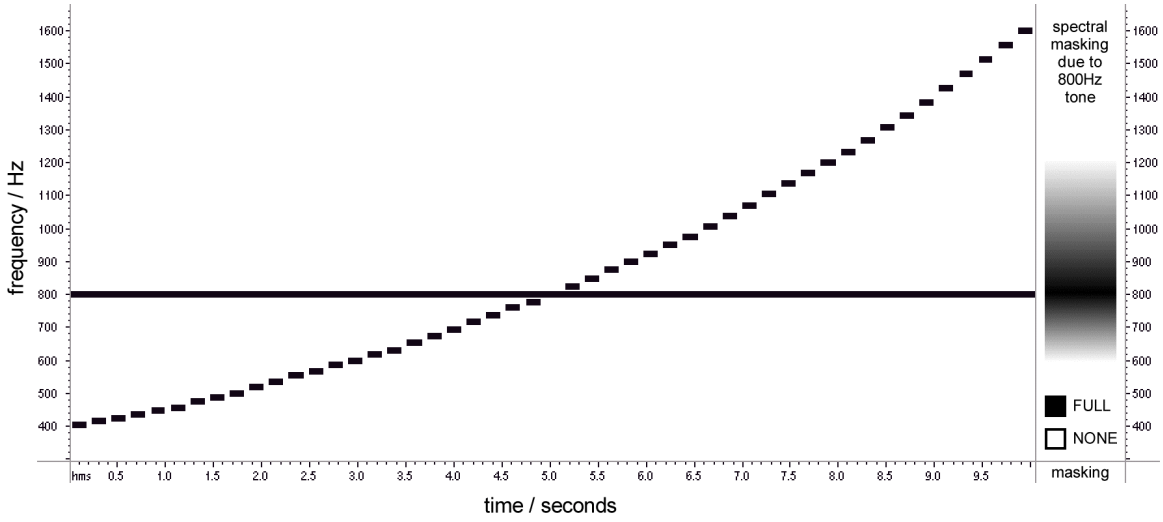


Figure 3.12: Left: Spectrogram of a demonstration of spectral masking
Right: Indication of masking due to 800 Hz tone

frequency because the sound wave enters the cochlea at the high frequency end of the BM, and travels down to the appropriate resonant place. Beyond this resonant place, the oscillation is rapidly damped.

The oscillating region covers the resonant peaks of several frequencies either side of the pure tone. Consider what would happen if a second pure tone is applied. This tone is of lower amplitude and slightly higher frequency than the first tone, and the region that it *should* excite on the BM is already excited by the original tone. If this new tone were played in isolation, the BM would vibrate, and it the tone be heard by the listener. However, the BM is already vibrating, and the presence of the new tone may not cause an increase in vibration. This means that the new tone will be *inaudible*. This effect is called spectral (or frequency) masking. No matter how hard or carefully the listener strains to hear the second tone, the transducing apparatus in the human ear does not have the capability to pass information about the second tone on to the brain. It can never be heard in the presence of the first (louder) tone [Moore *et al*, 1998].

This tone masking tone phenomenon can be demonstrated by generating a series of tones ascending in pitch, and then adding a louder tone, at an intermediate pitch, which persists through the entire sequence. A spectrogram of such a sequence is shown in Figure 3.12, and this demonstration is included on the accompanying CD-ROM. As the pitch of the stepped tones passes through the louder tone, the stepped tones become inaudible, demonstrating the

spectral masking due to the louder tone. Another feature is that the masking extends further above the tone than below it, due to the upwardly extended excitation pattern on the BM, as shown in Figure 3.11 – this is usually referred to as the upwards spread of masking.

In addition to the masking of one tone by another, there are three other basic masking phenomena: tone masking noise, noise masking tone, and noise masking noise. In each of these, the first signal is the *masker*, and the second is the *maskee* or *target*. The *masked threshold* is defined as the level at which the target signal is *just* audible in the presence of the masker. The *amount of masking* is defined as the difference between the masked threshold and the threshold in silence. The amount of masking depends on the nature of both the masker and the target, since the two can interact in complex ways within the ear. However, in all cases it is the action of the basilar membrane which is responsible for the masking.

Tone masking tone is the most complex combination, since the two tones can interact, generating beats and combination tones which undo the masking to a certain extent. A thorough study of this phenomena, and much useful psychoacoustic data, can be found in [Moore and Alcántara, 1998] and [Alcántara *et al*, 2000].

Tone masking noise: noise stimuli can be either narrow- or wideband. For a pure tone to mask noise, the noise must fall below the excitation due to the tone, and hence must be narrowband. The threshold is a function of the level, bandwidth, and centre frequency of the noise. Again, much useful data can be found in [Moore and Alcántara, 1998].

Noise masking tone: historically, experiments involving narrow band noise masking a pure tone have been used to discover the shape of the auditory filter, and much of the available data is in this form. The noise is centred on the frequency of the tone, and the bandwidth of the noise is increased. The tone becomes harder to detect, up until the point where the noise bandwidth matches that of the auditory filter. This is referred to as the critical bandwidth. Increasing the bandwidth of the noise beyond this causes no further reduction in the threshold of the tone. This is because the excess noise falls outside the filter characteristic of the point on the BM which detects the tone. This experiment was first performed by Fletcher in 1940, and has been repeated many times since (see [Bernstein and Raab, 1990] for a recent version).

Wideband noise masking a pure tone is the simplest of these synthetic stimuli for which to define the amount of masking, which is a function of target frequency, and (to a lesser extent) the level of the masker. A large amount of useful data is found in [Reed and Bilger, 1972].

Beyond these simple cases, almost all audio signals cause spectral masking. Spectral masking is a very important concept in the design of audio codecs. There are masked spectral regions around the loudest spectral components of every audio signal. These regions may be ignored, or filled with quantisation noise, without audibly altering the signal. For this reason, any model aiming to assess the quality of coded audio must take account of spectral masking. However, calculating the masked threshold for an arbitrary signal is a challenging task due to the non-linear nature of the ear.

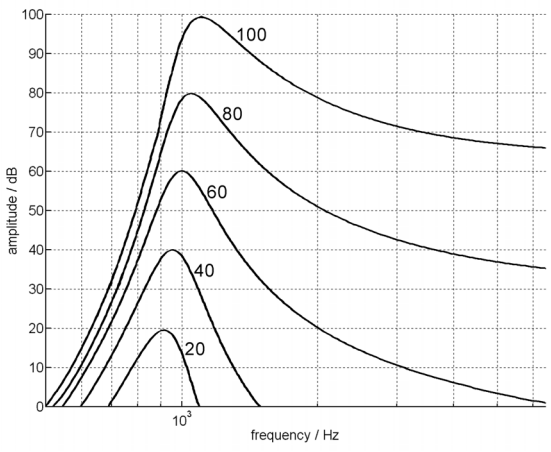


Figure 3.13: Variation of excitation on BM with stimulus amplitude

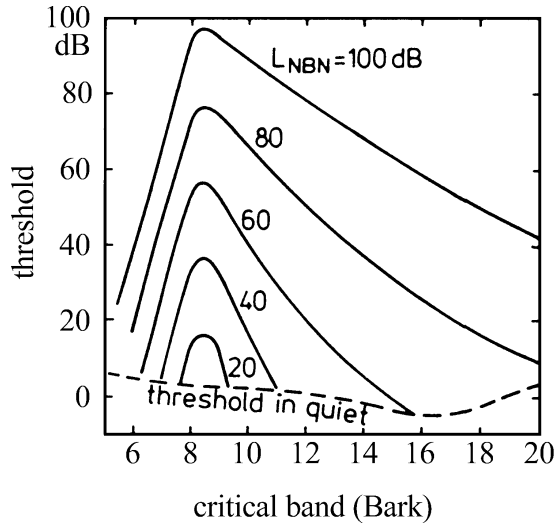


Figure 3.14: Variation of masked threshold with stimulus amplitude

3.2.4.1.3 Non linearities

The response of the basilar membrane to an incoming audio signal is amplitude dependent. The excitation pattern shown in Figure 3.11 is for a 90dB SPL stimulus. Figure 3.13 shows how this excitation varies with stimulus amplitude, while Figure 3.14 [Zwicker and Zwicker, 1991] shows how this effects the measurable masked threshold. Though the two graphs are drawn to different scales, the correlation between BM excitation and masked threshold (which has been heavily inferred throughout this discussion) is evident by comparing the two graphs.

Figure 3.14 shows that the shape of the masking curve is a function of stimulus amplitude, with a loud stimulus giving greater upwards spread of masking. Figure 3.13 confirms that this effect can be traced to the excitation upon the BM, which varies in a similar manner. Also, close examination of Figure 3.13 reveals that the position of the resonant peak changes slightly with stimulus amplitude. This latter phenomena does not have a direct effect upon masked thresholds, but it does affect pitch perception. For higher frequencies, for which the dominant pitch cue is the place of maximum vibration on the BM (see Section 3.2.4.1.1), the perceived pitch of a given frequency increases with stimulus amplitude, as first shown in [Stevens, 1935].

The non-linearity of the BM response has an important effect. The masking due to various simple stimuli was discussed in the previous section. Any arbitrary signal can be approximated as a series of tonal and noise-like components. The masking due to this arbitrary signal may be predicted from the linear supposition of the masking due to each component. However, excitations on the BM, and hence the masking due to them, do not add in a linear manner. This is very significant: many auditory models are introduced by noting that masking does not add in a linear manner, but this fact is subsequently ignored for the sake of computational efficiency.

The cause of the non-linear response of the BM is discussed in Section 3.2.4.3.

3.2.4.2 The Inner Hair cells

We now return to our journey through the human auditory system. The basilar membrane is moving in response to the incoming sound wave. This movement is detected by thousands of tiny hair cells, running along the length of the BM. There are many different types of cells on the BM, but the two most important are the inner and outer hair cells, labelled in Figure 3.1(c).

There are approximately 4000 inner hair cells along the length of the BM. They transduce the movement of the BM into neural impulses, which are carried onto the brain via the auditory nerve. At medium and high amplitudes, the inner hair cells only fire when the BM moves upwards, so the signal is effectively half wave rectified. This does not occur at lower amplitudes, so the process is sometimes referred to as “soft rectification”. Each cell needs a certain time to recover between firings, and the firing of any individual cell is pseudo-random, modulated by the movement of the BM. However, in combination, signals from groups of cells can give an accurate indication as to the motion of the BM.

At lower frequencies the firing of individual inner hair cells may phase lock to the incoming signal, thus the phase of the signal is preserved and transmitted along the auditory nerve. However, above approximately 1.5 kHz, the hair cells cannot lock onto individual cycles, and only the amplitude envelope is transmitted. The hair cells may phase lock onto sub-multiples of higher frequency stimuli (i.e. every 2^{nd} or 3^{rd} cycle), but the resulting phase information is weaker.

In the lower phase-locked frequency region, the repetition rate of the inner hair cells gives the auditory system a strong cue as to the frequency of the stimulus. This accounts for the ability to accurately discriminate pitch at lower frequencies, which would not be possible if the ear relied solely on the resonant place upon the BM.

The firing of the inner hair cells also transmits information about the amplitude of BM motion. The greater the amplitude of motion, the greater the probability of a given hair cell firing. However, for a single cell, this process rapidly saturates. To overcome this, three sets of cells of differing sensitivity are arranged across the BM. This extends the dynamic range to around 10^3. The short-fall between this, and the measured dynamic range of 10^6 is discussed in Section 3.2.4.3, after discussing the other features of the inner hair cells.

3.2.4.2.1 Temporal masking

The response of a group of inner hair cells to short and long tone bursts is shown in Figure 3.15 (calculated from [Meddis, 1986], [Meddis, 1988], [Meddis *et al*, 1990]). Three important features are apparent. Firstly, the hair cells are most sensitive to the onset of sounds. Secondly, the hair cells take a finite time to recover their full sensitivity after the end of a sound. Thirdly, the magnitude of this recovery depends on the duration of the sound. This gives rise to a process known as temporal

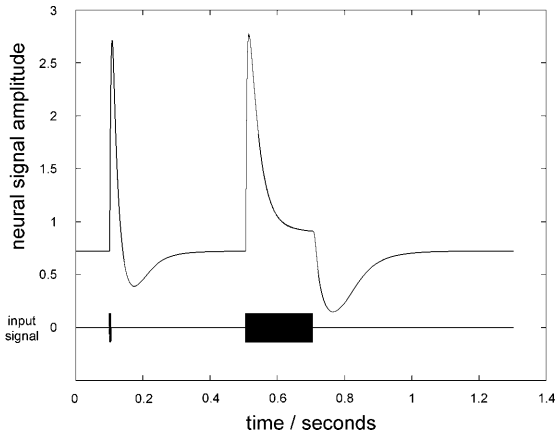


Figure 3.15: Overall response of hair cells to two 1 kHz tone bursts (5 ms then 200 ms)

masking[3]. If a sound (of a similar frequency) occurs during the period of recovery, the hair cells may be unable to register it, hence it may be inaudible.

This has been demonstrated experimentally [Zwicker, 1984]. A burst of white noise is followed by a short tone. The graph in Figure 3.16 shows the masked threshold of the tone at various times after the burst of noise. Note that the time scale on the graph is logarithmic. The two plots are for 200ms and 5ms bursts of noise. Tones below the level indicated are inaudible.

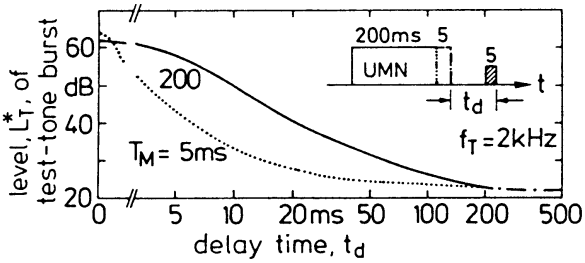


Figure 3.16: Temporal masking

Threshold due to a 200ms (solid line) and 5ms (dashed line) burst of white noise. Threshold shown for 5ms burst of 2kHz tone. [Zwicker, 1984]

This phenomenon is called post-masking, since the masking effect occurs after the masker. Pre-masking is also a measurable phenomenon, whereby a sound preceding a masker may be masked. Pre-masking is a small effect; untrained subjects sometimes exhibit pre-masking thresholds comparable to those found for post-masking. However, trained subjects typically exhibit pre-masking durations which can be accounted for by the ringing of the auditory filters (i.e. the motion of the BM, discussed in Section 3.2.4.1), hence the hair cells are thought to play no part in pre-masking. Despite this, the exact cause of pre-masking is still under debate.

Temporal masking is an important concept in the design of audio codecs. The masked regions following loud sounds are rarely used for the concealment of quantisation noise. However, the filterbanks found in audio codecs spread the audio energy in the time domain, and if this spreading exceeds the limits of temporal masking, it will cause audible damage to the signal. For this reason, any model aiming to assess the quality of coded audio must take account of temporal masking.

[3] At the time of experimentation, latest psychoacoustic knowledge suggested that temporal masking was due to the adpatation present within the auditory system, substantially the action of the inner and outer hair cells. Recent work [Oxenham, 2001] suggests that this *may* not be the case, and that a temporal integration process within the brain may be responsible. Without wishing to spoil the plot of the thesis, it is interesting to note that a temporal integration stage was necessary in the final model in place of higher processing.

3.2.4.2.2 Absolute Threshold

The inner hair cells fire at random, even in the absence of any incoming sound. In silence, the blood flowing around the regions of the inner ear becomes audible. These two factors combine to set an absolute minimum threshold of hearing. The shape of the threshold of audibility (Figure 3.17) is set by the resonance of the ear canal, the fall off in mechanical sensitivity of the BM at the frequency extremes, and the death of hair cells as humans age. The absolute minimum value, due to random hair cell activity, is 10^6 times less than that of the loudest sound humans can hear.

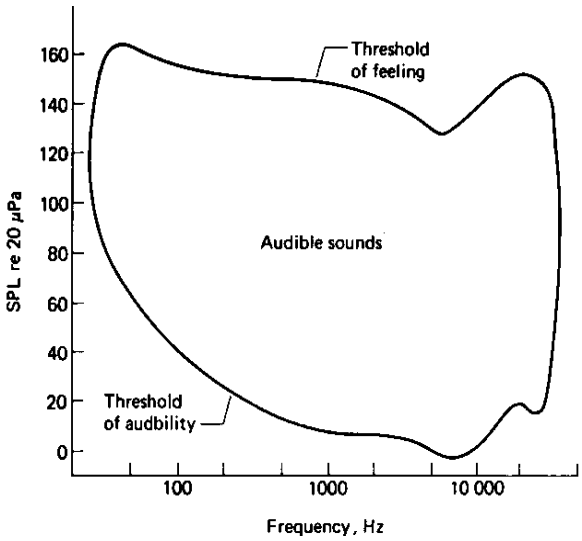


Figure 3.17: Minimum audible field for a range of frequencies.

3.2.4.3 The outer hair cells

There are approximately 12000 outer hair cells 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, and extends the dynamic range of the human auditory system.

Together, the inner and outer hair cells interact in an active feedback system, referred to as the cochlea amplifier, which increases the movement of the BM for quiet sounds, and suppresses it for loud ones. This mechanism is widely hypothesised, but unproven[4]. It is difficult to measure

[4] The stereocilia are directly above the outer hair cells, and are triggered by shearing forces between the outer hair cells and the tectorial membrane. One hypothesis [Yates, 1995] suggests that information from the stereocilia (rather than the inner hair cells) determines the length of the outer hair cells. Both hypotheses may be partially true, but either hypothesis yields a similar result; the outer hair cells change their length (and hence the resonant properties of the BM) in response to the motion of the BM. A third hypothesis suggests that the stereocilia, rather than the outer hair cells, may act to adjust the resonant properties of the BM (see [Yates,

directly as it only operates in live, undamaged cochlea. It is known that the inner hair cells are connected to the auditory nerve mainly by afferent fibres, which transmit neural signals to the brain. Conversely, the outer hair cells are predominantly connected to the auditory nerve by efferent nerve fibres, which receive neural signals from the brain. It is also known that the outer hair cells change their length by up to 10% due to feedback from the brain. Finally, the sharp tuning and amplitude dependence of resonances on the BM, and the large dynamic range of the human auditory system cannot be explained without some external feedback loop which returns energy into the system. Hence, though it is as yet unproven, the existence of the cochlea amplifier will be accepted as a probable mechanism of auditory perception throughout this project.

The cochlea amplifier acts as a sophisticated automatic gain control mechanism within the human ear, and causes human listeners to have a logarithmic perception of loudness. This should come as no surprise to audio engineers, who usually measure levels and responses in decibels (dB) which is a logarithmic scale ($20*\log_{10}$ amplitude). It is the AGC function of human hearing that causes humans to hear this way. It also extends the dynamic range of human hearing (the ratio of smallest to largest detectable pressure difference) from the 10^3 that would be achieved with the inner hair cells alone, to 10^6.

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.

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 humans *cannot* hear is attributable to this transduction process.

1995] for a review). As this hypothesis is the least likely of the three, it is not considered further within this research.

3.2.5 Neural signal processing

The function of each individual stage of the subsequent neural processing is less well understood (see Figure 3.1(d)). Some features can be predicted by comparison between the auditory nerve signal, and the final perception of sounds common to all listeners. Other processes can be inferred from the neural connections present between auditory processing centres. The action of individual cells within each processing centre can be recorded.

3.2.5.1 Cochlea Nucleus

The Cochlea Nucleus contains many cells, and has neural connections to the auditory nerve, superior olivary complex and the inferior colliculus. The cochlea nucleus is thought to sharpen and enhance features of the (now highly compressed) signal from the auditory nerve, and distribute it to the superior olivary complex and the inferior colliculus.

3.2.5.2 The Superior Olivary Complex

The Superior Olivary complex is believed to be responsible for the lateralisation of sound sources. As with all the processing discussed thus far, two SOCs are present within the auditory system, one corresponding to each ear. However, nerve fibres from *both* ears feed into each SOC.

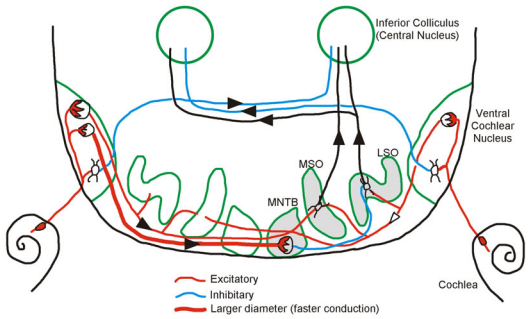


Figure 3.18: Neural connections of the binaural pathway

Within the SOC there is a region called the **Medial Superior Olive**, where nerve fibres from similar frequency regions in each ear meet. It is believed that the MSO calculates interaural time differences via a series of delays and co-incidence detectors. This concept was first proposed in [Jefress, 1948], and measurements carried out by [Patterson, WEB-3] appear to confirm the presence of this processing.

It is sometimes stated that the MSO is most effective at lower frequencies (below about 1.5 kHz), where interaural time difference is the dominant lateralisation cue. However, this assumption arises from the results of lateralisation measurements carried out using pure tones. If a signal contains no spectral components below 1.5 kHz, it can still be lateralised using the ITD cue alone *if* the envelope of the signal is modulated at a rate below 1.5 kHz. This meas-

ured response suggests that nerve fibres corresponding to frequencies higher than 1.5 kHz must be processed within the MSO.

Another region within the SOC, called the **Lateral Superior Olive**, calculates the interaural intensity difference. This is the dominant lateralisation cue for pure tones above 1.5 kHz, and is a useful cue for all sounds containing high frequency components. Within the right-hand SOC, inputs from the right-hand ear act to excite the LSO, and inputs from the left-hand ear act to inhibit the LSO. The opposite is true within the left-hand SOC. Thus, the ILD is reflected by the relative activity within the two MSO regions.

The density of detectors is not equal for all values of ITD and ILD. Rather, both sets of detectors are most plentiful for the region around zero interaural difference, and the number of detectors falls as the interaural difference increases. The result is that the human auditory system is less sensitive to large interaural differences, and sounds at source locations which yield these differences are localised with decreased accuracy. The best localisation performance occurs for sources directly in front of the listener, which yield approximately zero interaural time and level differences.

Whilst the physiology is unknown, experimental data proves that the outputs of the SOC are post-processed in a special manner. The binaural system is known to be "sluggish" – that is, it requires a finite time in order to detect any change in the binaural aspect of an input stimulus. This is true in both binaural localisation and detection tasks. For example, the minimum audible angle separating two sequential sound sources is approximately 1° where the two sounds are temporally separated by several hundred milliseconds [Mills, 1958]. However, [Perrot and Pacheco, 1989] show that by reducing the time between the two sounds to 40 ms, the minimum detectable angle is increased to 3°. Similar effects have been measured with respect to the Binaural Masking Level difference. For example, modulating the interaural correlation of the masking noise at a frequency higher than 2 Hz almost destroys the BMLD [Grantham and Wightman, 1979].

This binaural sluggishness is not a defect of the auditory system. On the contrary, forming immediate localisation determinations based upon the instantaneous output of the SOC would give rise to error. Whilst textbook determination of the interaural time delay of white noise yields a single sharp peak, correlation of real world signals gives rise to many brief false peaks. These may be due to the periodic nature of the input signal, the happenstance coincidence of

two unrelated auditory events, or the interference from early reflections. (It is worth nothing in passing that the precedence effect is chiefly accounted for by pre-processing, especially the response of the inner hair cells [Theis, 1999], rather than the subsequent integration). Integration of the correlator output removes most of this anomalous information, whilst strengthening the lateralisation cue.

3.2.5.3 Subsequent neural processing

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.

The neural processing which cannot be related to known auditory performance will not be discussed here. For more information about the auditory neural pathway, and the structure of the processing centres, the reader is directed to [Ehret and Romand, 1997] and [Anon, web].

3.3 Conclusion

In this chapter, the path of a sound wave has been followed from outside a listeners head, past the pinna and ear canal, into the cochlea, onto the basilar membrane, through the inner hair cells, along the auditory nerve, and into the higher processing centres.

The function of each stage of the human auditory system has been qualitatively discussed. Where available, experimentally measured human auditory performance data has been referenced. The frequency selectivity of the basilar membrane has been identified as the primary source of spectral masking. The adaptation of the hair cells has been suggested as the source of temporal masking. A neural contra-lateral coincidence detector has been suggested as the primary localisation cue.

3.4 Acknowledgements

This chapter draws on the following reviews of the human auditory system: [Yates, 1995], [Patterson, WEB], [Schubert, 1978], [Neely, WEB], [Allen and Neely, 1992], [Rosen and Howell, 1991], [Dallos, 1978], [Fantini, 1996].

4

Auditory Perceptual Models

4.1 Overview

An auditory model simulates human hearing. Such a model can be used to calculate the perceived difference between two audio signals. This chapter commences with a discussion of the various approaches to auditory modelling. As an example, the classic Johnston model is examined in detail, and the strengths and weaknesses of this model are assessed. Several models that are more recent are compared, and important areas for development in the present work are identified and discussed.

4.2 Auditory modelling

In order to explain approaches to auditory modelling, it is necessary to start with a generalisation, the limits of which will soon become apparent. The generalisation is this; there are two distinct approaches to modelling human hearing which have been pursued by researchers in different areas.

Psychoacousticians use the physiology of the human auditory system as a reference, and simulate the processing found therein. The resulting auditory models are often complex and computationally burdensome. The accurate simulation of auditory processing gives rise to measurable effects, such as spectral masking. For this reason, models of this type will be referred to as models of *cause*. The success of such models is judged by their ability to mimic human performance in psychoacoustic tests.

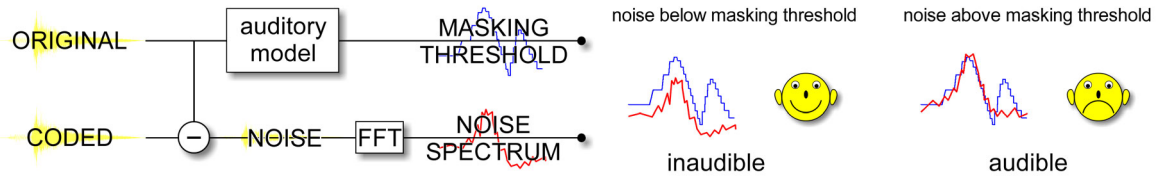
Audio engineers use the measured performance of human hearing as a reference, and directly model the known limitations. The resulting models are comparatively simplistic, and the effects of the human auditory system (e.g. spectral masking) are explicitly calculated within the model. For this reason, models of this type will be referred to as models of *effect*. These models are often computationally efficient. Their success is judged by their ability to determine the perceived quality of an audio signal.

In reality, most models lie somewhere between these two exaggerated extremes. Psychoacousticians often produce "signal processing" versions of complex models that yield similar results via simplified methods. For example, rather than modelling the response of each individual hair cell, the firing probability for large groups of cells may be computed. In this way, the "cause based" model becomes less like the system it aims to simulate, but the overall *effect* of the model is barely changed.

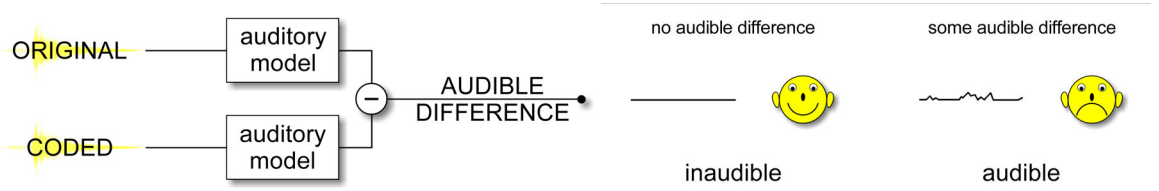
In contrast, audio engineers often refine models that simulate the "effects" of human hearing such that the model comes to match the processes present in the auditory system more closely. For example, all auditory models must transform data from the time domain into the frequency domain. An FFT may be used to transform blocks of data, but such an approach can often hide audible problems in the signal that the model aims to detect. One common solution is to use a filterbank, which gives a time to frequency transformation, without the block processing problems inherent with an FFT. The use of a filterbank mimics the processing which occurs within the auditory system, and moves the operation of the *effect* model much closer to that of a *cause* model.

Thus, when attempting to model human hearing, there are two different, but parallel approaches: model the cause or process that gives rise to some effect, or model the effect directly. In practice, there are many possibilities lying between these two extremes for each process (and associated effect) within the auditory system.

One useful approach that lies on a tangent somewhere between the two extremes is to map the signal into the perceptual domain. This approach is discussed in [Hollier, *et al*, 1993+1995]. As shown in Chapter 3, the human auditory system operates in a logarithmic domain in both frequency and amplitude perception. By transforming the signal into a domain that corresponds to the human auditory system's internal processing, a clearer idea of the human perception of the signal can be gained than by examining the linear time-domain signal. This is true even if



(a) Prediction: model of *effect* calculates masking threshold



(b) Comparison: model of *cause* calculates internal representation

Figure 4.1: Methods of determining perceived difference. (a): prediction (b): comparison

no attempt is made to model the masking due to the auditory system, or the processes that give rise to it.

The aim of the present research is audio quality assessment. The task is to calculate the perceived difference between two signals: the original signal, and a coded version of the same signal. The arithmetic difference between the two signals is coding noise, which may or may not be audible.

There are two ways to use an auditory model to calculate the perceived difference. The first method is to calculate the masking threshold due to the original signal. Any noise below the threshold is inaudible, whilst any noise above it is audible (Figure 4.1 a). This method is used by some models of *effect*.

The second method is to pass both signals through the auditory model in parallel (Figure 4.1 b). The auditory model transforms each signal into a representation of what is heard by a human listener, e.g. an internal representation. A comparison of the two internal representations yields a prediction of the audible difference between the two signals. The comparison may be made via a simple subtraction, or by complex analysis. If the result is zero, then the model predicts that there is no audible difference between the two signals. A numerical result other than zero indicates how large the perceived audible difference is likely to be.

Some models of *cause* use a detector (optimum or sub-optimum) to compare the original signal to the degraded signal after both have passed through the auditory model in parallel. Depending on the accuracy of the auditory model, and the nature of the difference between the two signals, the detector may need to be crippled in some manner to match human performance. The process is not usually deterministic, but stochastic. This matches human perception, and is often achieved by adding noise to the detection process itself.

The perceptual domain analysis gives an indication of the perceived difference between two signals, but since it does not include calculations of masking, it cannot determine whether any difference is audible or inaudible. Adding masking calculations to the perceptual domain analysis may turn it into a model of *effect*, allowing the masking threshold to be computed directly.

In conclusion, an auditory model may be based upon the processes found within the auditory system, or the effects to which these processes give rise. Most models use a combination of these two approaches. To predict the audible effect of a coding scheme, the coding noise may be compared with a prediction of the masked threshold due to the original signal. More commonly, both the original and coded signals are transformed into internal representations, and these are compared.

It is hoped that this overview has set the scene for the dissection of auditory models that follow. This will commence by examining a classic model of *effect*, the Johnston auditory model.

4.3 The Johnston model

In 1988, James Johnston published details of a model for calculating the “perceptual entropy” of audio signals [Johnston, 1988a]. This model calculates the masked threshold due to an audio signal in order to predict which components of the signal are inaudible. In this way, the model can be used to predict how much data is needed to transparently code the signal. The model is included in an audio coder [Johnston, 1988b], where the prediction of inaudible components is used to feed a bit allocation algorithm which reduces the data rate to 128 kbps for a mono signal.

The Johnston model is used in an audio coder, but the present research goal is a model to assess the quality of coded audio. Using the Johnston model for this purpose may be unwise –

any audible problems introduced by the model within the coder are unlikely to be detected by using the same model as an assessor. However, there are three reasons why the Johnston model is an excellent starting point for the current research. Firstly, examining the operation of a coding model will give valuable insight into areas where coded audio may be non-ideal, and hence the quality assessment model must be particularly vigilant. Secondly, the "basic" quality assessment model in the current ITU-PEAQ standard [Thiede *et al*, 2000] is similar to the Johnston model, indicating that for some applications, this model does an acceptable job of predicting perceived audio quality. Thirdly, most current auditory models are similar to the Johnston model, and some of the principles involved are universally accepted.

The Johnston model calculates the spectral masking due to an audio signal, but temporal masking is not addressed. Though the human auditory system is *continuous* in both time and frequency, the spectral masking estimate calculated by the Johnston model is *discrete* in time and frequency. The signal is split into short (64 ms) frames, and the spectral masking for each frame is computed as if the signal were steady state. The masking is computed for 25 frequency bins, spaced equally on the critical band scale. Thus one masked threshold is calculated for each frequency bin every 64 ms. This threshold is calculated by taking the FFT of a 64 ms frame, summing the energy in each frequency bin, spreading the energy to simulate spectral masking, adjusting for the nature of the signal, and normalising the result.

The following detailed walk through the Johnston auditory model is drawn from [Johnston, 1988a], [Johnston, 1988b], and [Rimell, 1996]. A MATLAB implementation of the model is included on the accompanying CD-ROM. This implementation is taken as a baseline for the current research; differences from the published sources are noted in *italics*. The plots show the progress of a synthetic signal (consisting of a 500Hz tone and a 5kHz tone) through the model.

4.3.1 Algorithm

4.3.1.1 Window and FFT

The audio signal is split into frames. In the original model, a frame length of 64ms is employed (2048 samples at a sampling frequency of 32kHz). In the present implementation, a default length of 23ms is used since this is a convenient and perceptually appropriate timespan, as discussed in [Paillard *et al*, 1992]. This is equivalent to a *window* length of 1024 samples at a sampling frequency (f_s) of 44.1kHz, though the MATLAB variable *window* is easily changed to accommodate any window length.

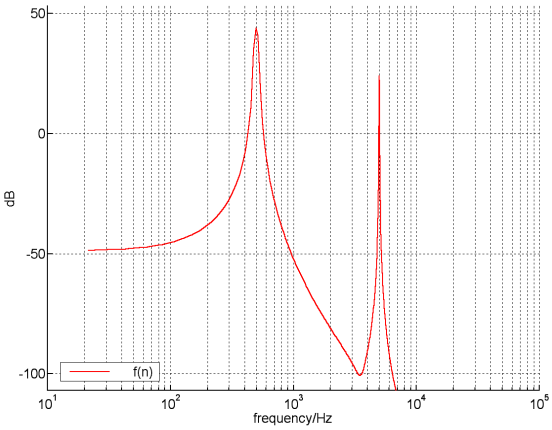


Figure 4.2: *f(n)* Signal Spectrum

The current frame is windowed with a Hanning (raised cosine) window and an FFT (Fast Fourier Transform) is performed. Each line (n) in the complex FFT refers to a spectral component of frequency f (in kHz), given by

$$f(n) = \frac{l.f_s}{1000 * window}, \tag{4-1}$$

which is valid for the first $(window/2)$ complex lines.

4.3.1.2 Critical Band Analysis

The real and imaginary components of the spectrum $\mathrm{Re}(n), \mathrm{Im}(n)$ from the FFT are converted to the power spectrum, $P(n)$, thus:

$$P(n) = \mathrm{Re}^2(n) + \mathrm{Im}^2(n) \tag{4-2}$$

This power spectrum is segregated in linear frequency, but the auditory system processes frequency on a near logarithmic scale, called the critical band scale, as discussed in the previous chapter.

The relationship between linear frequency, f in kHz, and the critical band, or Bark frequency, z_c in Bark, is given by

$$z_c = 1 + [13\arctan(0.76f) + 3.5\arctan\left(\frac{f}{7.5}\right)^2], \quad (4\text{-}3)$$

adapted from [Zwicker and Terhardt, 1980] to number critical bands from 1 to 25. If the lowest frequency component falling in critical band i, where $i = \text{int}(z_c)$, is given by bl_i, and the highest frequency component falling in critical band i is given by bh_i, then the summation of the energy in band i is given by:

$$B_i = \sum_{n=bl_i}^{bh_i} P(n) \quad (4\text{-}4)$$

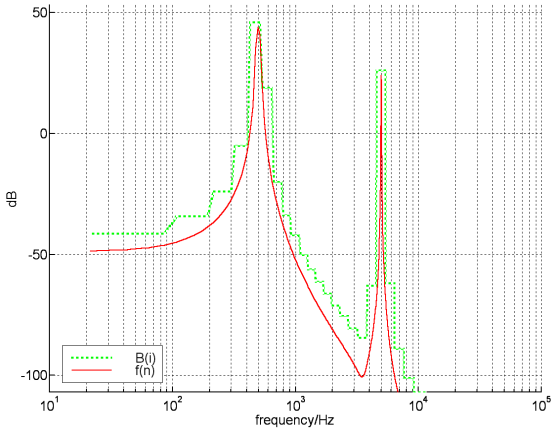


Figure 4.3: B_i Critical Band Energy

The energy in each critical band is summed in this manner. *The D.C. component of the spectrum is not included in this summation.*

4.3.1.3 Spreading function

The following spreading function (taken from [Schroder *et al*, 1979]) is used to estimate the effects of masking across critical bands.

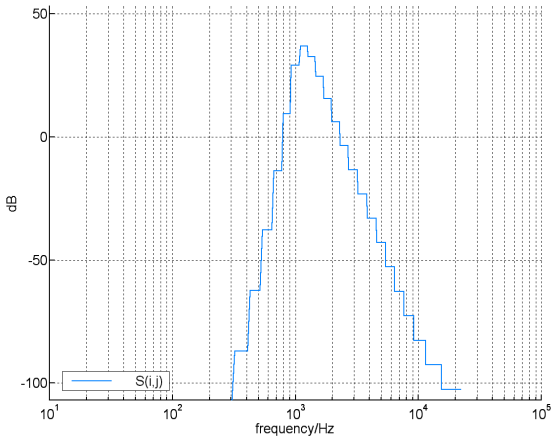


Figure 4.4: S_{ij} Spreading Function

$$S_{i,j}(\text{dB}) = 15.81 + 7.5(y + 0.474) - 17.5\sqrt{1 + (y + 0.474)^2}\ , \text{dB} \quad (4\text{-}5)$$

$$S_{i,j} = 10^{\left(\frac{S_{i,j}(\text{dB})}{10}\right)} \quad (4\text{-}6)$$

where:

y = $i - j$ *(not the modulus, as stated in some other papers)*

i = Bark frequency of masked signal

j = Bark frequency of masker signal

The spread critical band spectrum is calculated by convolving the spreading function with the critical band spectrum. This can be achieved by matrix multiplication, thus:

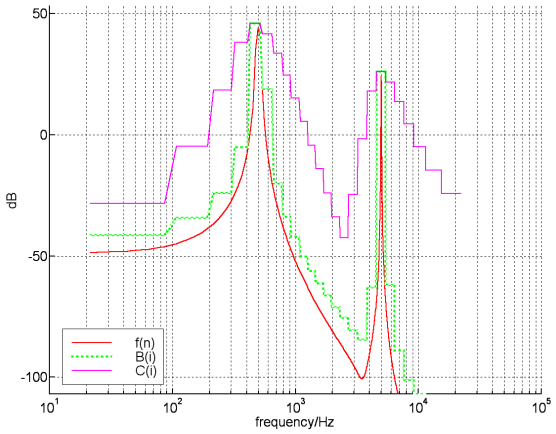


Figure 4.5: C_I Spread CB Spectrum

$$\begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ \dots \\ C_{25} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,25} \\ S_{1,2} & S_{2,2} & S_{2,3} & \dots & S_{2,25} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,25} \\ \dots & \dots & \dots & \dots & \dots \\ S_{25,1} & S_{25,2} & S_{25,3} & \dots & S_{25,25} \end{bmatrix} \times \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \dots \\ B_{25} \end{bmatrix} \tag{4-7}$$

4.3.1.4 Coefficient of tonality

The masking threshold for noise masked by a tone is taken to be $14.5 + i$ dB below C_i, but the masking threshold for a tone masked by noise is taken to be 5.5 dB below C_i [Johnston, 1988b]. Johnston uses the Spectral Flatness Measure (SFM), calculated from the geometric and arithmetic means of the power spectrum, to determine how tone-like or noise-like the signal is. The SFM is given by

$$SFM_{dB} = 10 \log_{10} \left(\frac{GM}{AM} \right) \tag{4-8}$$

However, this calculation cannot be performed correctly using 32-bit floating point arithmetic. *In the current implementation, the SFM is calculated using logarithms*, thus:

$$SFM\,(\text{dB}) = 10[\log_{10}(GM) - \log_{10}(AM)] \tag{4-9}$$

where

$$\log_{10}(GM) = \frac{1}{N}\sum_{n=1}^{N}\log_{10}\big(P(n)\big), \tag{4-10}$$

$$\log_{10}(AM) = \log_{10}\left[\frac{1}{N}\sum_{n=1}^{N}P(n)\right] \tag{4-11}$$

and $N = window/2$. An SFM of zero dB would indicate that the signal is entirely noise like, while an $SFM >= SFM_{dB\max}$, where $SFM_{dB\max} = -60$ dB, would indicate that the signal is entirely tone like. Most tone like signals, such as organ, sine waves, or flute have an SFM that is close to or over the limit. A coefficient of tonality is calculated as follows:

$$\alpha = \min\left(\frac{SFM_{\mathrm{dB}}}{SFM_{\mathrm{dB\,max}}},1\right) \tag{4-12}$$

This coefficient is used to geometrically weight the two thresholds, yielding O_i, the threshold offset, thus:

$$O_i = \alpha(14.5+i) + (1-\alpha)5.5 \tag{4-13}$$

Though this final equation is *linear*, equations (4-9) to (4-13) yield the required *geometric* weighting.

4.3.1.5 Spread threshold estimate

The offset O_i is subtracted from the spread critical band spectrum C_i to give the spread spectrum estimate T_i, thus:

$$T_i = 10^{\log_{10}(C_i)-(O_i/10)} \tag{4-14}$$

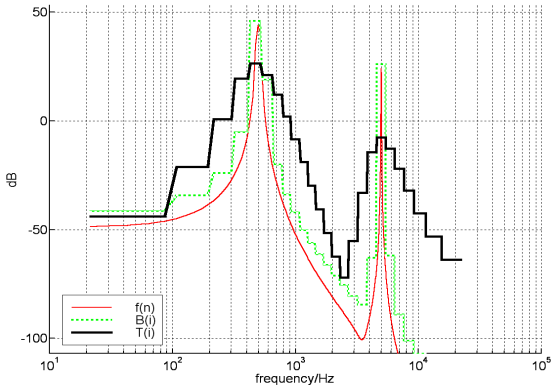


Figure 4.6: T_I Spread Threshold Estimate

4.3.1.6 Re-normalisation of the threshold estimate

This process is suggested in [Johnston, 1988b], though the details in [Johnston, 1988a] seem incomplete. The overall effect is comparatively small (2.04 dB by this calculation). The implementation here is as follows.

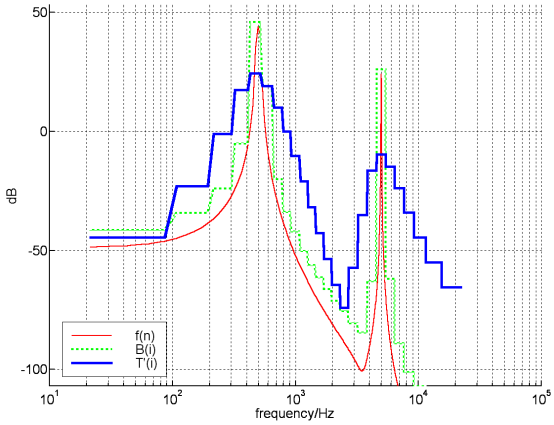


Figure 4.7: T_i' Normalised threshold estimate

The spreading function described in Section 4.3.1.3 increases the overall energy, where as the psychophysical process that we are attempting to model spreads the energy by dispersing it. For example, examine the behaviour with a hypothetical stimulus with unity energy in each critical band. The actual spreading function of the ear will result in no overall change to the level of energy in any critical band[1]. However, the spreading function presented here will cause the energy in each band to increase, due to the additive contributions of energy spread from adjacent critical bands.

The solution presented here is to normalise the threshold estimate at this stage. A hypothetical stimulus, with unity energy in each critical band, is used as the B_i in equation (4-7), to give the spread spectrum error, C_{Ei}, thus:

$$\begin{bmatrix} C_{E1} \\ C_{E2} \\ C_{E3} \\ \dots \\ C_{E25} \end{bmatrix} = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & \dots & S_{1,25} \\ S_{1,2} & S_{2,2} & S_{2,3} & \dots & S_{2,25} \\ S_{3,1} & S_{3,2} & S_{3,3} & \dots & S_{3,25} \\ \dots & \dots & \dots & \dots & \dots \\ S_{25,1} & S_{25,2} & S_{25,3} & \dots & S_{25,25} \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} \qquad (4\text{-}15)$$

[1] In reality the lowest and highest bands will loose energy by this process, but all other bands will loose and gain equal amounts of energy by dispersion, hence the total level of energy in each band will remain unchanged.

The normalised threshold estimate $T_i^{'}$ is calculated by converting C_{Ei} into dB, and subtracting it from the threshold estimate, thus:

$$T_i^{'} = T_i - 10\log_{10}(C_{Ei}) \tag{4-16}$$

4.3.1.7 Conversion to dB SPL

In the original model, Johnston set the absolute level such that a signal of 4kHz, with peak magnitude of ±1 least significant bit in a 16-bit integer, is at the absolute threshold of hearing. A more sophisticated approach is taken in the present implementation which takes into account the replay level of the audio signal.

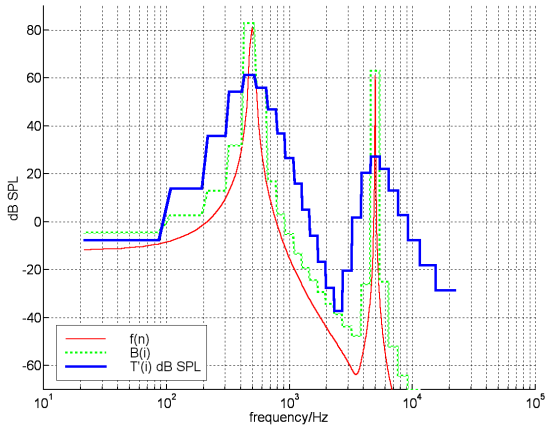


Figure 4.8: $T_i^{'}$ Normalised Threshold Estimate in dB SPL

To include the absolute threshold of hearing (the Minimum Audible Field, or MAF) in the masking threshold estimate, it is necessary to relate the digital audio signal to a real listening level. It is suggested in [Stuart, 1990] that a 105 dB sound pressure level (SPL) at 1 m from the speaker is a typical maximum level for listening in the home. The author pities Bob Stuart's neighbours! A quieter 90 dB SPL reaching the listener is chosen as the default value here.

The threshold estimate is corrected relative to this level as follows. A full scale 1 kHz sine wave undergoes the windowing, FFT, and conversion to power spectrum as described in Sections 4.3.1.1 and 4.3.1.2. This yields the calculated power of the sine wave, $P_{ref\,(\mathrm{dB})}$. The difference between this calculated power and the specified loudness level of 90dB is added to the normalised threshold estimate (in dB), thus:

$$T_{i(\mathrm{dB}spl)}^{'} = T_i^{'} + (90\mathrm{dB} - P_{ref\,(\mathrm{dB})}) \tag{4-17}$$

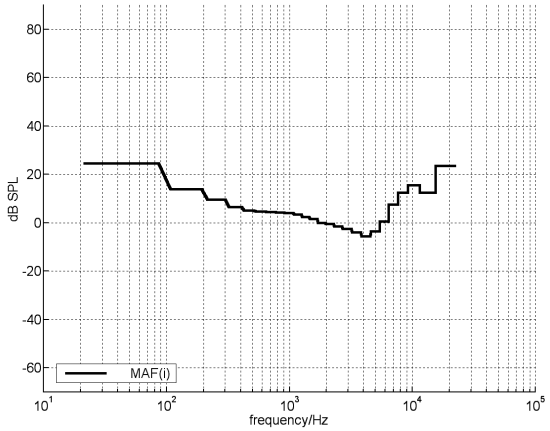


Figure 4.9: M_i Minimum Audible Field

in dB SPL

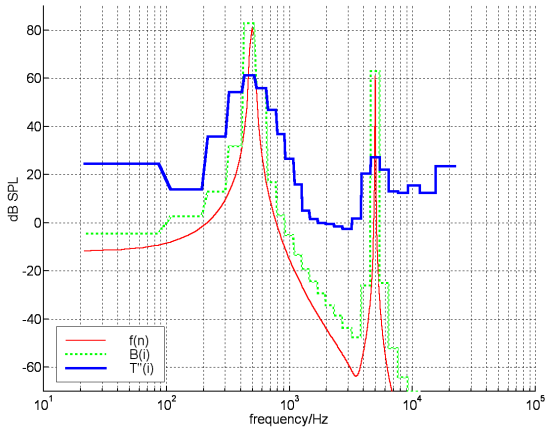


Figure 4.10: T_i'' Final Threshold Estimate

in dB SPL, including MAF

4.3.1.8 Inclusion of Minimum Audible Field threshold information

The minimum audible field information is taken from [Robinson and Dadson, 1956]. The conversion to dB SPL is carried out using the reference of the MAF threshold at 2kHz being equivalent to 0 dB SPL [ISO 389-7, 1996]. The minimum threshold in each critical band, M_i is taken to be the lowest value of the MAF curve falling within that band.

Johnston uses the median MAF value in each band. His approach may predict a sound as being masked when in reality it is audible, whereas the approach chosen here may predict that a sound is audible when in reality it is masked. This is preferable for detecting the noise added by an audio coder. However, these differences occur only at very low (and very high) frequencies, where the slope of the MAF curve gives rise to a large discrepancy between the median and the minimum value.

Hence, the final threshold estimate is given by

$$T_i'' = \max\left(T_{i(\mathrm{dB}spl)}', M_i\right). \tag{4-18}$$

4.3.1.9 Threshold interpretation

The Johnston model was designed to identify the least important spectral regions of an audio signal in order to control the bit allocation algorithm of an audio codec. The present task is to judge perceived audio quality. Both tasks require the identification of inaudible signal components, but whereas the former task requires prediction of such from a single signal, the latter requires a comparison between original and processed signals to determine the presence of an audible difference. In this section, the use of the Johnston model threshold estimate in these tasks is discussed.

There are at least three ways in which to interpret the threshold estimate, as follows.

1) All spectral components below the masked threshold are inaudible, and may be removed at will, without audibly changing the signal

2) Any spectral components which, on their own, would fall below the masked threshold, may be added to the signal without causing an audible change.

3) Two signals will sound identical if the spectral components lying above the masked threshold are identical.

Interpretations 1 and 2 are the ones for which this model is designed, and are used by audio coders, including [Johnston, 1988b]. In reality, the coder does not remove all spectral components below the threshold estimate, nor does it fill the entire spectral space below the threshold estimate with noise. However, the coder does treat all spectral components below the threshold estimate as being less important, as discussed in Chapter 2. In extreme circumstances, it may remove some of them entirely, but it is unlikely to strip all masked components neatly away – that is not its function. This fact is mentioned here to dispel this common misconception, which is a gross simplification. The usual mode of operation is to represent the spectral components below the threshold estimate with decreased accuracy by using coarser quantisation steps. This has the effect of adding noise within the spectral regions below the threshold estimate. However, the aim is not to add as much noise as possible, but rather to incur as much noise as is necessary to meet the bitrate requirements, whilst constraining the noise to be below the threshold estimate.

In an audio coder, interpretations 1 and 2 are not pushed to their limits. If the bitrate is sufficiently high, then coding noise will be significantly below the threshold estimate. If the bitrate is too low, then noise may be added above the threshold estimate. However, since the coder will breach the threshold estimate to meet bitrate requirements, the absolute value of the threshold estimate is less important than the shape (in the time and frequency domains) which should cause the coding noise to be minimally audible. Conversely, in the assessment of perceived audio quality, the absolute value of the threshold is vitally important, since the aim is to judge whether coding noise is audible or inaudible.

When assessing the perceived quality of coded audio, we assume access to the original signal, the coding noise, and the resulting coded signal[2], thus:

$$original + noise = coded \tag{4-19}$$

Interpretations 1 and 2 only require knowledge of the original signal and the noise, be it additive or subtractive. From this knowledge, the audible difference between the original and coded signals may be inferred.

The third interpretation is different. It says, "two signals will sound identical if the spectral components lying above the masked threshold are identical." This requires knowledge of the original signal and the altered signal, and implies a comparison of the two. Though this interpretation appears similar to the first two, it is a fundamentally different concept. Consider, for example, the case where the original signal and the coding noise are highly correlated. The noise may be below the masked threshold (hence judged inaudible by 2), but when summed with the original signal, some components may add constructively, thus yielding new or altered spectral components *above* the masked threshold (hence judged audible by 3). Interpretations 1 and 2 rely on rules to predict what will be audible. Interpretation 3 relies on a comparison between two signals to measure the difference.

In conclusion, the masked threshold estimate can be used to determine whether there is an audible difference between two signals *either* by examining the original signal and predicting

[2] Only the original and coded signals are usually available. For high quality coded audio, the synchronisation of these two is trivial, hence the noise may be calcualted via subtraction.

the allowable coding noise, *or* by examining the difference between the original signal and the coded signal. These two possibilities will be referred to as the predictive measure, and the comparative measure.

4.3.2 Verification: Psychoacoustic test.

To verify the performance of the model, a simple, repeatable test of audibility is required, in which the performance of the model can be compared to that of a real human listener. A suitable psychoacoustic test is the determination of masked threshold. In such a test, a human listener is required to detect one sound (the target) in the presence of another (the masker). The level at which the target is *just* audible is called the threshold of masking, or masked threshold. To verify the performance of the Johnston model, two simple tests are used: noise masking tone, and tone masking noise. These two properties are incorporated in the model, as described in Section 4.3.1.4, so it is reasonable to expect the model to match human performance in these tasks.

For the predictive measure, the masked threshold estimate calculated by the Johnston model is used directly to determine whether a target is audible; any target above the threshold estimate is judged to be audible. For the comparative measure, the masker and masker + target are compared, and any differences above the threshold estimate are judged to be audible.

4.3.2.1 Noise Masking Tone

This experiment was carried out using human listeners in [Reed and Bilger, 1973] and the results presented in that paper are used as a reference. The masking noise is at a spectrum level of 35 dB/Hz. The target, a 1 kHz tone, is just audible at a level of 45 dB. Figure 4.11 shows the threshold estimate calculated by the Johnston model, and also shows the 45 dB tone. The tone is nearly 10 dB below the threshold estimate. Using the predictive measure, the model incorrectly judges the tone to be inaudible.

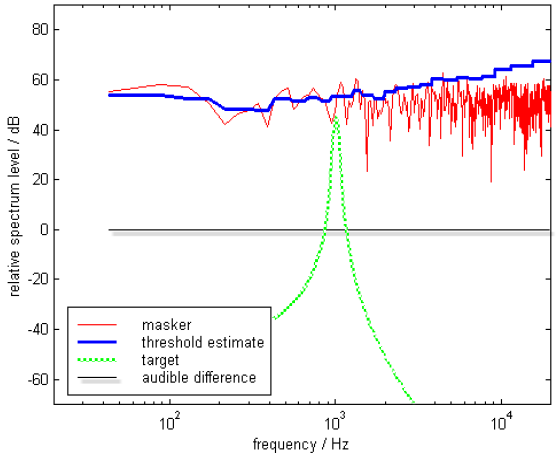


Figure 4.11: Noise Masking Tone

target below threshold estimate

model prediction = inaudible (incorrect)

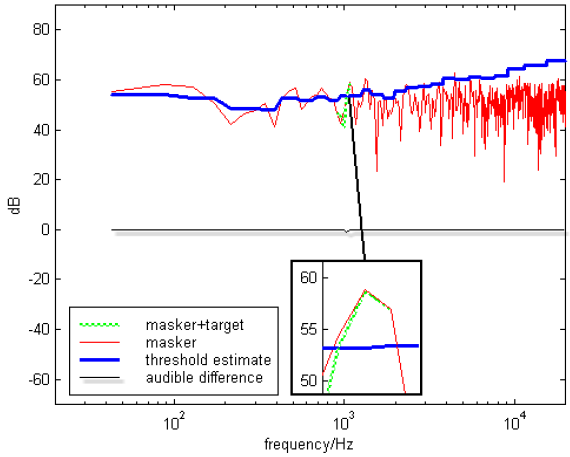


Figure 4.12: Noise Masking Tone

masker vs masker + target difference above threshold estimate

model prediction = audible (correct)

The results of the comparative measure are more promising. Comparing the spectrum of the masker with that of the masker + target (Figure 4.12) reveals that there is a small difference *above* the threshold estimate, indicating that this difference is just audible. The audible difference plot (in black) is calculated by subtracting the spectra of the two signals where they lie above the threshold estimate. The blip visible at the centre of the audible difference plot shows that there is a 1 dB difference between the masker and the target + masker – this is enough to be detected, and matches the performance of a human listener.

For this example of noise masking tone, the predictive measure did not give a direct prediction of masked threshold. However, the comparative measure correctly delineated the audible difference between the masker and masker + target signals.

The model is designed from the results of simple masking experiments, yet it apparently cannot reproduce the results of such experiments when the threshold estimate is interpreted as originally intended (as a prediction measure). Before rejecting the predictive measure, other possible sources of error will be examined.

The short-term windowing of a noise signal gives a false indication of its spectrum, a fact which may be interfering with the masking calculation. The true long term spectrum of the noise is flat from 20 Hz to 20 kHz, but the spectrum shown in picture Figure 4.11 is far from flat. Subsequent snapshots give differing spectra. However, at no time does the threshold

estimate fall enough to cause the 45 dB tone to be judged audible, so this problem is not causing the discrepancy in the prediction measure.

Another explanation for the failure of the model is the windowing and FFT process, because it disperses the energy of the signal in the spectral domain. The frequency domain representation of a pure sine wave is infinitely narrow, but the FFT of a windowed pure sine wave has a finite spectral width, determined by the shape and length of the window function. As the length of a sine wave decreases, it becomes less well defined, since fewer wavelengths fall within the window. Hence, a shorter window will cause greater spectral dispersion. Narrow- and wide-band noise signals are also dispersed, but with one important difference. The spectrum of white noise is distributed evenly, so dispersion does not alter its peak or average value. Unfortunately, the peak value of the sine wave spectrum is significantly reduced by the dispersion. The result is that white noise with a spectrum level of 35 dB/Hz appears at the same level after the windowing and FFT. A pure tone of 45 dB appears to have a much lower peak power when passed through the same process.

In the model, the conversion to dB SPL is carried out with reference to a sine wave. Consequently, the conversion corrects for the reduction in peak sine wave level, but in doing so, it over amplifies noise-like signals. This is clearly visible in Figure 4.11, where a 35 dB/Hz band of noise appears at 50 dB/Hz.

Increasing the window length may help to solve the problem. The original Johnston model uses a 64 ms window, whereas the present implementation uses a 23 ms (1024 sample) window. In the current test, increasing the window length to 4096 samples (93 ms) reduces the error from 8.2 dB to 6.1 dB. However, the variance from one snapshot to another is slightly greater than this value, so the improvement is not significant. This problem will be discussed further after the next test.

In conclusion, for this psychoacoustic experiment, the Johnston model does not predict human perception if the prediction measure is used, but does so if the comparison measure is employed.

4.3.2.2 Tone Masking Noise

This experiment was carried out using human listeners in [Moore and Alcántara, 1998] and the results from this paper are used as a reference. The masking tone is at a level of 85 dB SPL and a frequency of 1 kHz. The target noise band ranges from 960 Hz to 1040 Hz, and is just audible at a level of 45 dB/Hz. Figure 4.13 shows the Johnston model threshold estimate for the masking tone. Using the predictive measure, the model predicts that the target noise is audible, since the spectrum of the target (green) goes above the threshold estimate (blue), as shown in Figure 4.13. If the level of the noise is reduced by 1 dB, the model still predicts that it is audible. A reduction of 2 dB is required for the model to predict that the noise is inaudible. Hence, the model over estimates the masked threshold by 1.5 dB, but such a margin of error may be acceptable in a masking threshold determination. If the window length is increased to 4096 samples, the model under estimates the masked threshold by 0.5 dB. This is a trivial error, but again indicates that the choice of window length can be important.

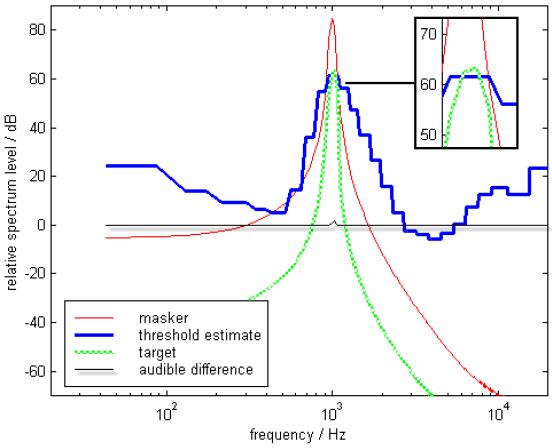


Figure 4.13: Tone Masking Noise

target above threshold estimate

model prediction = audible (correct)

The comparative measure works less well in this case. With a window length of 1024 samples, the model predicts a just audible difference, but in entirely the wrong frequency region (Figure 4.14). Switching to a window length of 4096 samples, the model predicts an audible difference of 0.5 dB with the noise 5 dB below threshold (i.e. inaudible to a human listener). At the true masked threshold, the model still predicts an audible difference of 0.5 dB (Figure 4.15). As the level of the noise is increased beyond this, for every dB above masked threshold, the predicted audible difference also increases by 1 dB.

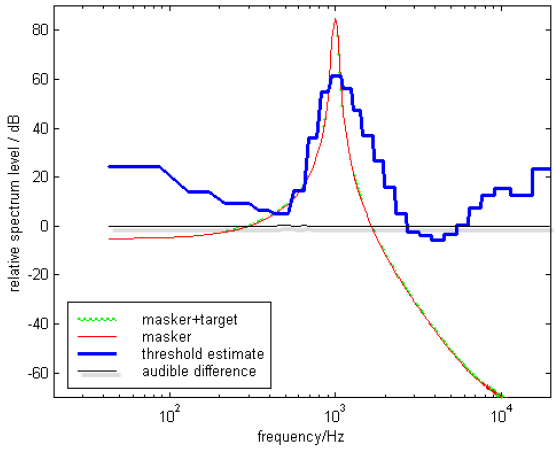


Figure 4.14: Tone Masking Noise

masker vs masker + target difference above threshold estimate, but at wrong frequency!

model prediction = audible (correct)

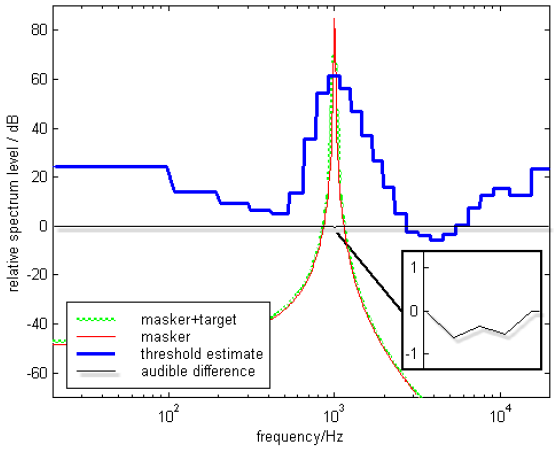


Figure 4.15: Tone Masking Noise

masker vs masker + target difference above threshold estimate

model prediction = audible (correct)

In this experiment, a predicted audible difference of 0.5 dB is inaudible to human subjects. Having an *in*audible "audible difference" may sound like nonsense, but it may indicate that a different mechanism is at work here. The threshold estimate is a calculation of masking. However, the comparative measure requires that, within audible spectral regions, the sound must be identical to a second sound in order to be perceived as identical. Studies have shown that this is not the case. If two tones of the same frequency are presented sequentially, it is sometimes impossible to detect any amplitude difference between the two tones of less than 1 dB. For this reason, it may be reasonable to expect "audible differences" (i.e. differences above the masked threshold) to be inaudible if they are less than 1 dB. In this experiment, adding a 1 kHz tone and a 1 kHz band of noise *may* simply increase the amplitude of the tone, making the task similar to the one outlined above.

For this example of tone masking noise, the threshold estimate gave a direct prediction of masked threshold, showing that the prediction measure gives a correct result for this experiment. Comparing the masker with the masker + target, the threshold estimate correctly delineated the audible difference between the two, if a difference threshold of 1 dB was introduced. This demonstrates that a modified version of the comparative measure gives a correct result for this experiment.

4.3.2.3 Discussion

Three significant issues are raised by these tests.

Firstly, the predictive measure doesn't correctly predict the threshold of the noise masking tone experiment. As discussed in Section 4.3.2.1, this may be due to the dispersion of energy in the spectral domain inherent in the windowing and FFT process. In this test, one signal is infinitely spectrally narrow, while the other occupies the maximum spectral bandwidth of the system. This difference in signal bandwidth maximises the discrepancy due to dispersion. Such a large discrepancy would not occur in the intended use of the model, where the codec noise is designed to occupy a similar spectral space to the original signal.

Secondly, the comparative measure requires a threshold value in order to predict the masked threshold correctly.

Thirdly, the Johnston model commences with a windowing and FFT operation. This test has shown that, even in a simultaneous steady state masking experiment, where temporal effects are relatively unimportant, the window length *is* important.

In conclusion, the comparative measure, with a threshold value around 1 dB, may be used with some confidence. The effect of window length is discussed further in Section 4.3.4.3.

4.3.3 Assessing Audio Quality using the Johnston model

In this section, the Johnston model is used to predict perceived audio quality. An existing subjective test employing a panel of human listeners is used as a reference. The test is repeated here with the human listeners replaced by the Johnston model. The quality assessments of the Johnston model are compared to those of the listening panel to evaluate how accurately the model predicts human perception.

The human perception of the quality of various audio codecs is assessed in [Soulodre *et al*, 1998]. The results are presented as diffgrades, which indicate the audible difference between the original and coded signals. The scale ranges from 0 to –4, where 0 indicates that the original signal and the coded signal sound identical, and –4 indicates that the differences between the original and coded signals are "very annoying".

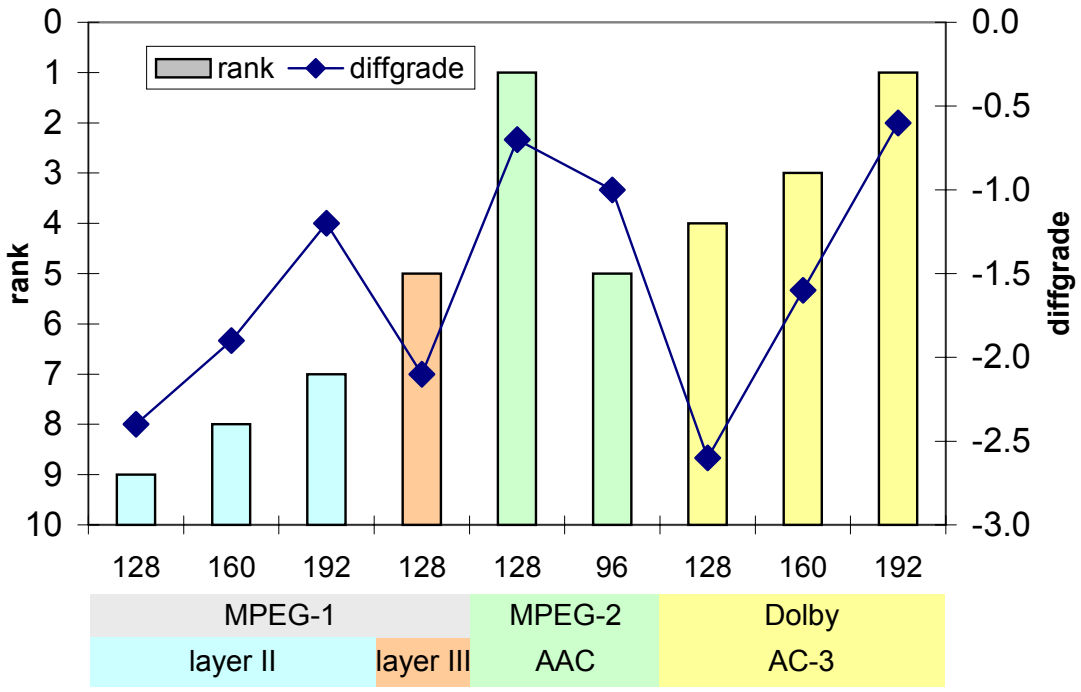


Figure 4.16: Human subjects assessment of audio quality

A variety of audio codecs are tested in the paper, using eight different audio samples. One of the most challenging audio samples is the Harpsichord (Track 40 from [EBU, 1988]). This audio sample is used in the present test.

Unfortunately, the reference versions of the audio codecs employed in the published test are not generally available. However, MPEG-1 layer III, MPEG-2 AAC, and AC-3 software codecs have been obtained from the same companies who provided the reference codecs used in the published test. An MPEG-1 layer II software codec has been obtained from the only company currently offering such a codec. Thus, the coded audio that will be judged by the Johnston model is similar, but not identical, to that auditioned by the human listeners in the original test.

The human perception of audio quality for each codec is shown in Figure 4.16. The diamond points show the diffgrades from [Soulodre *et al*, 1998], using the reference codecs.

The author auditioned the decoded audio from each codec, and ranked the results in the order of preference indicated by the shaded bars in Figure 4.16. This ranking does not correspond to the diffgrades. This indicates that either the author has a very different perception of sound

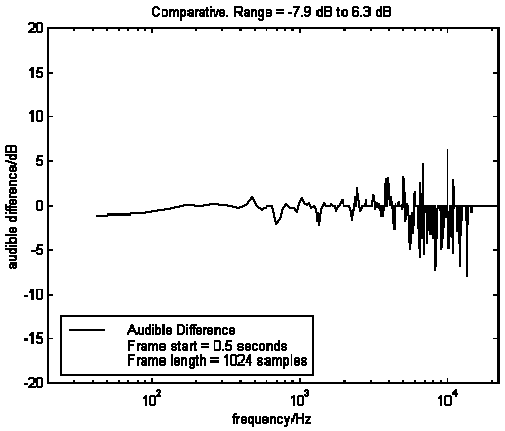


Figure 4.17
Instantaneous audible difference

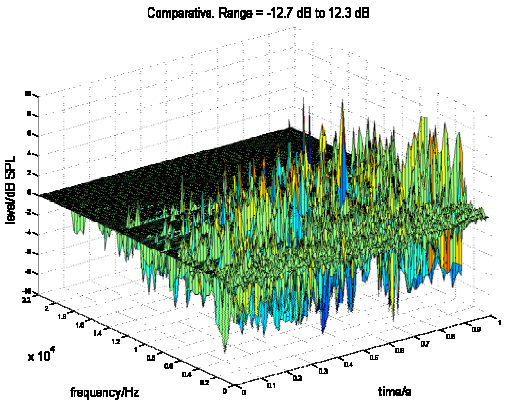


Figure 4.18
Audible difference surface

quality from the listening panel, or the codecs used here are very different from those used in the original test. Whilst no two people perceive sound in an identical manner, the author believes that the codecs are more likely to be at fault. This is borne out not only by over four years subjective assessment experience, but also by the fact that the most mis-ranked codec (MPEG-1 layer II) is the one which was *not* obtained from the same source as the reference version. The sound of the MPEG-1 layer II codec is an order of magnitude worse than that of the others on this test. Since the diffgrades do not correspond to the audio samples that will be presented to the model, the ranking must be used as a reference instead.

The Johnston model will be used with the predictive measure and the comparative measure, as outlined in Section 4.3.1.9. The audible difference calculated for a single FFT frame is shown in Figure 4.17. Calculating this over successive frames gives the audible difference surface, which varies over time and frequency, as shown in Figure 4.18. This rendering of the 3-D audible difference surface onto a 2-D medium is difficult to interpret. As an alternative, Figure 4.19 shows the same audible difference plotted as a colour image, where the colour at each point represents the audible difference for that time and frequency, as indicated on the colour scale to the right. The spectrum of the original signal can also be represented in this manner, as shown in Figure 4.20. This is similar to a conventional spectrogram, except that the lowest frequencies are shown at the top of the graph, and the frequency scale is linear, rather than logarithmic.

Figure 4.20 shows the first second of the original signal. This section of the signal consists of the first two notes of an arpeggio. The onset of the first note occurs at t=0 seconds, whilst the

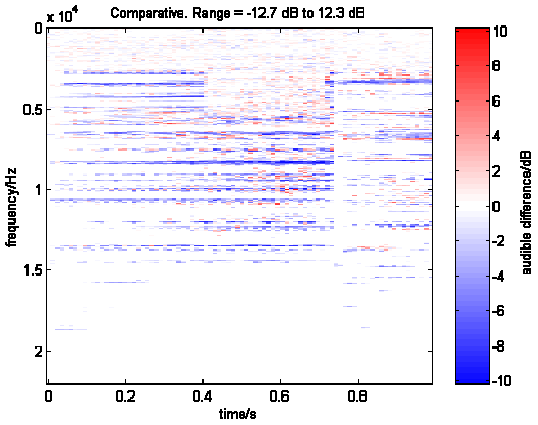


Figure 4.19

Audible difference colour image plot

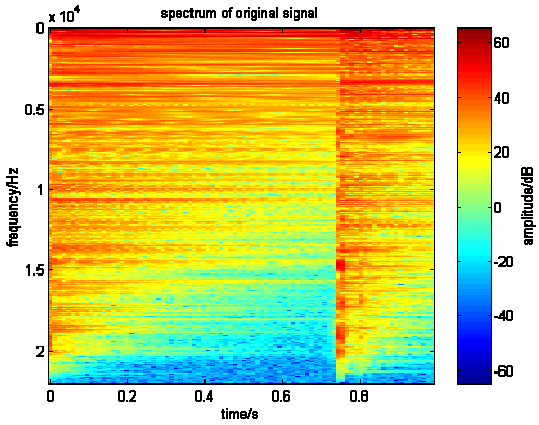


Figure 4.20

Original signal colour image plot

onset of the second note occurs at t=0.75 seconds. The reader should note that both these onsets are apparent in Figure 4.20, as is the decaying high frequency spectrum as each note dies away. The features of the audible difference colour image plot (Figure 4.19) can also be interpreted. The horizontal blue bands represent spectral components that have been removed by the codec, whilst any red areas represent noise that has been added by the codec. If this audible difference plot matches human perception, then the codec shown in Figure 4.19 is removing many audible spectral components. There is no predicted audible difference at high frequency (the bottom of the plot). The reason is that, whilst there are differences between the original and coded signals within this frequency region, the Johnston model predicts that these will all lie below the minimum audible threshold, and hence will be inaudible.

The following audible difference plots represent the Johnston model's assessment of the coded audible from the subjective test. The codecs are presented in the order in which a human listener ranked them, with the best first. If the Johnston model does match human perception, then some trend should become apparent. The left-hand plots are calculated by the predictive measure, the right-hand plots by the comparative measure.

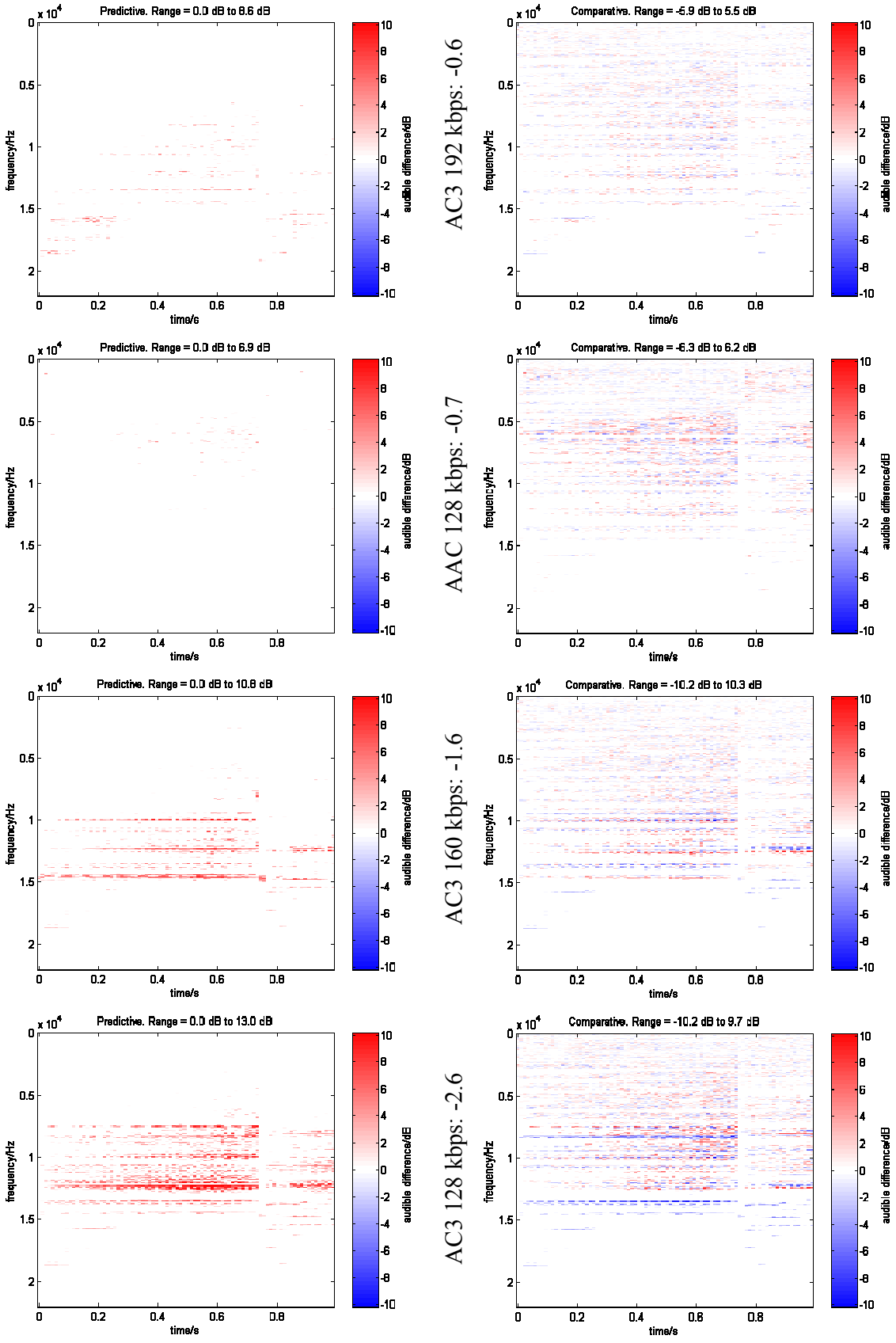


Figure 4.21 (a) Johnston Model audible difference predictions for codecs under test

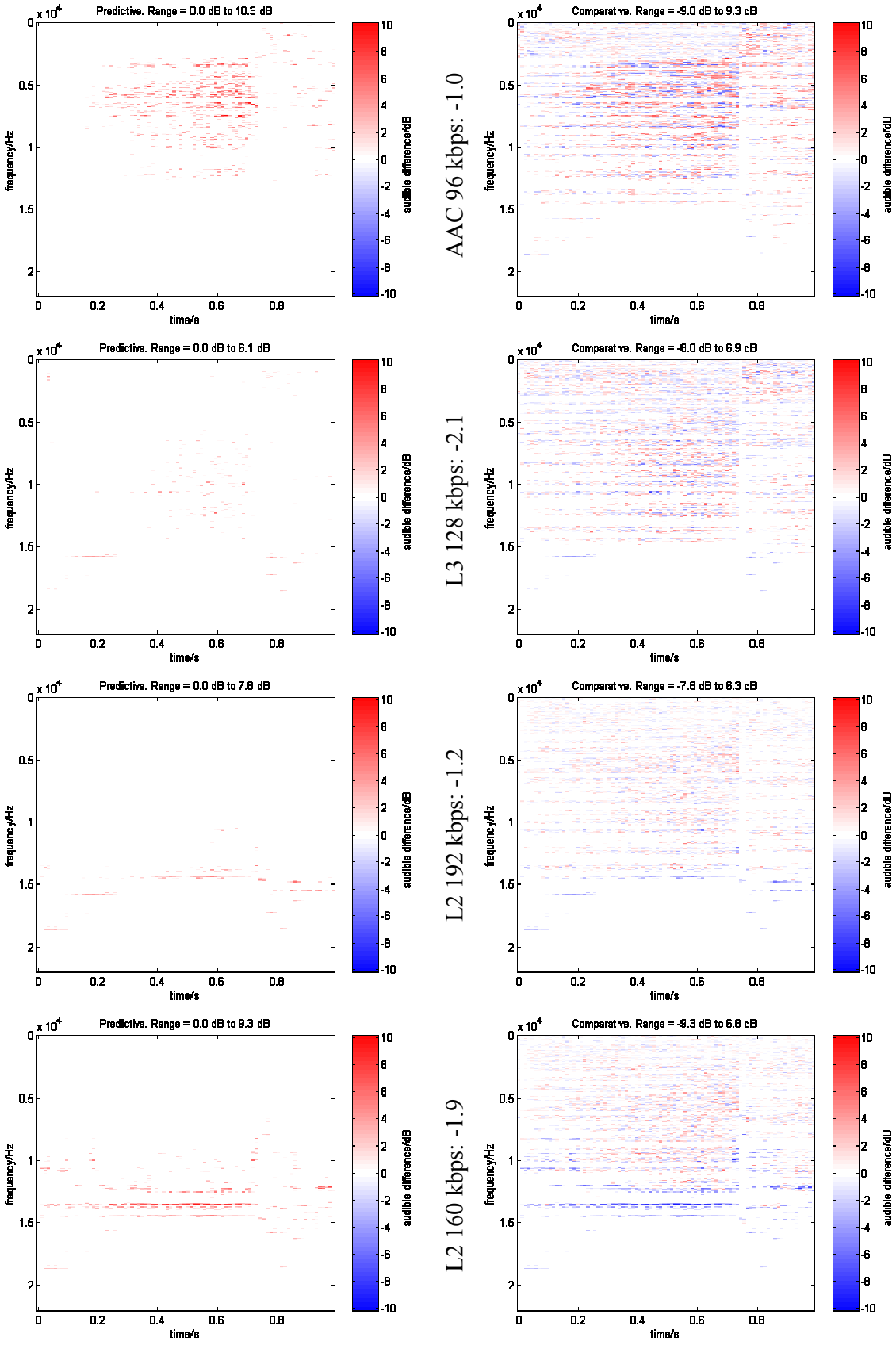
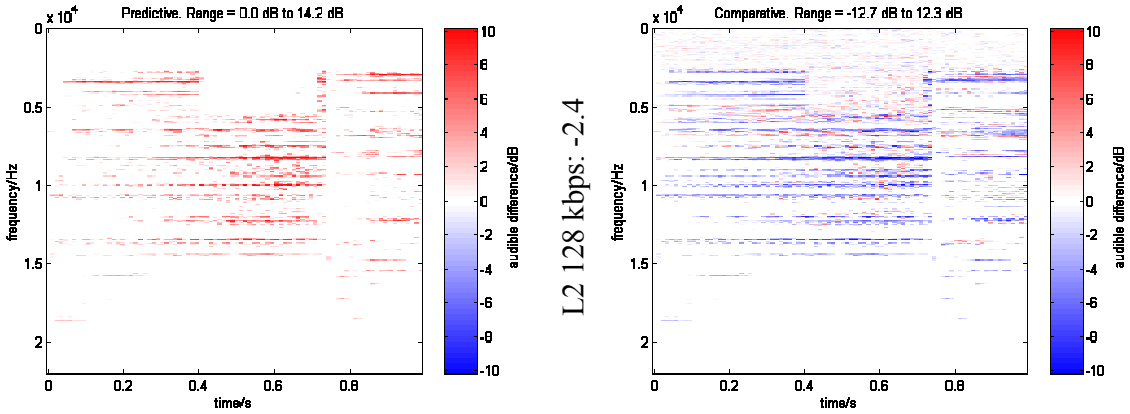


Figure 4.21 (b) Johnston Model audible difference predictions for codecs under test



predictive measure comparative measure

Figure 4.21 (c)

Johnston Model audible difference predictions for codecs under test

Meaningful analysis of the audible difference surface is a complex task. Some numerical solutions are discussed in [Hollier *et al*, 1995]. However, for our present purpose of judging the ability of the Johnston model to predict human perception, visual examination will suffice.

If the model correctly predicts human perception of audio quality, then some aspect of the audible difference will become more pronounced as the audio quality is reduced. Unfortunately, this is not universally true. This expectation does hold within each codec family; as the bitrate is reduced, the sound quality deteriorates, and the audible difference surface becomes more active. However, across codecs, no relationship between perceived sound quality and the activity of the audible difference surface is apparent.

A notable case where the Johnston model fails to predict human perception of audio quality is the AAC 96 kbps codec. Both the listening panel and the author ranked this codec as significantly better than any of the layer II codecs. However, the Johnston model predicts a large audible difference for the AAC codec, where as the audible difference surface for the 192 kbps layer II codec is quieter overall, with less obvious peaks and fewer spectral components.

This reveals a problem with the Johnston model in particular, and all auditory models in general. The auditory model built into the layer II codec is similar to the Johnston model itself. When the Johnston model assesses the quality of layer II coded audio, it uses the same criteria as the original codec. Hence, any short comings of the layer II codec are missed by the Johnston model. For this reason, the auditory model used within a perceptual measurement

system must always be more advanced and accurate than those used within the codecs it is designed to assess.

Interestingly, the AAC codec contains a more advanced auditory model than that of Johnston, and in this instance the Johnston model "hears" problems that are not apparent to human listeners. This demonstrates that, where the assessment model is inferior to a given codec, the model may mistake *in*audible noise as being audible, since the codec is aware of a type of masking which the model does not calculate.

In conclusion, the Johnston model fails to correctly assess perceived audio quality because it is less advanced than those found within the codecs it has assessed.

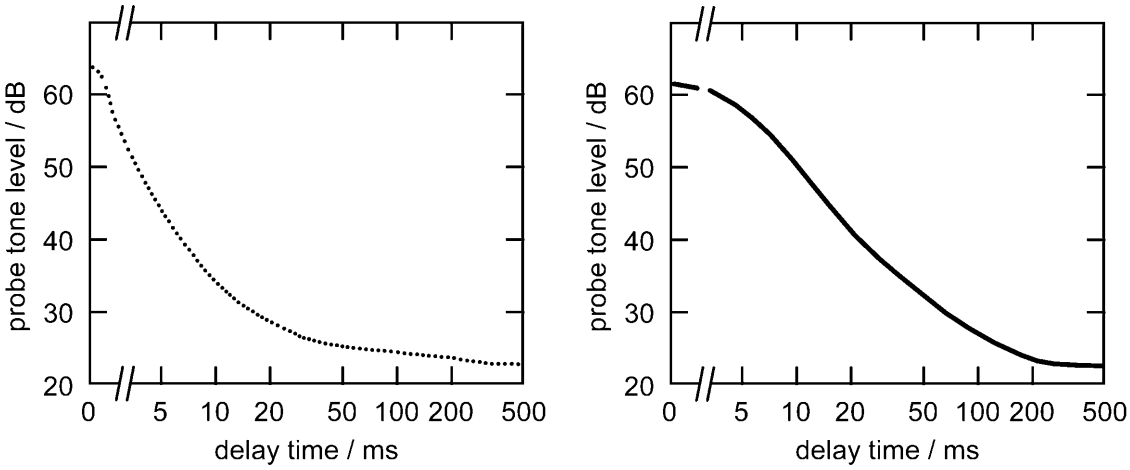
4.3.4 Criticism of the Johnston model

The following features are absent, or inadequately addressed in the Johnston model. The aim is to address all of these issues in the work that follows, in order to develop a model that is capable of assessing state-of-the-art audio codecs in accordance with human perception.

4.3.4.1 Spectral Masking determination

The Johnston model estimates the spectral masking in each critical band, yielding 25 individual masking estimates over the entire audio bandwidth. Thus, the masking estimate surface is constant over each critical band, but discontinuous at the critical band boundaries.

Whilst it is true that 25 critical bandwidth filters would cover the entire audible spectrum, there are *not* 25 such filters within the auditory system. Rather, there are many *overlapping* filters. The filter characteristic is defined by the response of the basilar membrane at any given point, as discussed in the previous chapter. The basilar membrane is continuous, though its movement is transduced by individual (discrete) hair cells, which act as detectors. The Johnston model uses one detector per critical bandwidth, whereas the human auditory system contains over one hundred. Though the human auditory system is a discrete system, the resolution is so fine that it is usually assumed to be continuous, since the spectral spread of movement along the BM, rather than the individual hair cells, is the limiting factor. For this reason, any accurate estimate of masking should be free from the discontinuities that are inevitable in the Johnston model.



(a) – masker = 5 ms noise burst (b) – masker = 200 ms noise burst

Figure 4.22: Temporal Masking – threshold of detecting 2 kHz tone after white noise

4.3.4.2 Temporal Masking determination

The Johnston model does not include a calculation of temporal masking. The window length sets the time resolution of the model, and this is chosen to be "perceptually appropriate". However, temporal masking is highly signal dependent, and is not accounted for by windowing alone.

The largest temporal masking effect is that of post-masking, where the target signal occurs after the masker. Figure 4.22 shows the threshold of detecting a 5 ms, 2 kHz tone after an 80 dB burst of white noise. Figure 4.22 (a) shows the threshold after a 5 ms burst of noise, whilst Figure 4.22 (b) shows the threshold after a 200 ms burst of noise (from [Zwicker, 1984]). The time scale is logarithmic, and the noise burst ceases at t = 0 ms. Comparison of the two figures reveals that the duration of the masker has a significant effect on the decay of masking.

The Johnston model may fail to predict any masking in this situation, a fact that can be demonstrated by example. The Johnston model calculates the masking estimate on a frame by frame basis. The default frame-length gives an inter-frame time of about 20 ms. If the noise burst ceases in the middle of one frame, and the target falls within the subsequent frame, the Johnston model does not attempt to calculate the masking of one sound by the other. Figure 4.22 (b) shows that, for a human listener, the target tone is masked below a level of 40 dB, 20 ms after the end of the noise burst. This masking is ignored by the Johnston model, and illustrates why temporal masking should be included in the present model.

4.3.4.3 Pre-echo detection

A masker may mask a signal that precedes it. This is known as pre-masking. The effect is much smaller than that of post-masking (5 ms compared to 200 ms). Whereas the Johnston model under estimates (or more correctly, doesn't estimate) post-masking, the effect of windowing is to over estimate pre-masking. This is because a 20 ms time window is used by default. Whilst any temporal smearing of the signal up to 20 ms will remain undetected by the Johnston model, such smearing is easily detected by a human listener, due to the 5 ms limit on pre masking.

Temporal smearing is a common problem with the MPEG-1 layer III audio codec, most readily demonstrated with the Castanets track from [EBU, 1988]. Typical listener impressions of this codec include "a softening of transients" and reports that "the sounds of castanets are spread out in time, sounding more like a whoosh than a sharp crack". In a subjective test ([Meares *et al*, 1998]), it is reported that the diffgrade for this item is –1.6. The audible artefacts include "temporal distortion, high frequency loss, high frequency distortion".

The results of the Johnston model's assessment of this item, via the predicted measure, are shown in Figure 4.23 and Figure 4.24.

Using the default 2048-sample window length, and the predictive measure, the model calculates that there is no audible difference between the original and coded audio signals. This is shown in Figure 4.23 where the coding noise (shown in green) falls below the threshold estimate (shown in blue) for every frame. Halving the window length, to the value suggested in [Paillard *et al*, 1992] fails to reveal any audible difference. Switching to a 512-sample window length, the model finally detects an audible difference between the original and coded signals. Figure 4.24 shows that a large audible difference is detected in the 3rd and 4th frames, where the coding noise (green) is above the threshold estimate (blue).

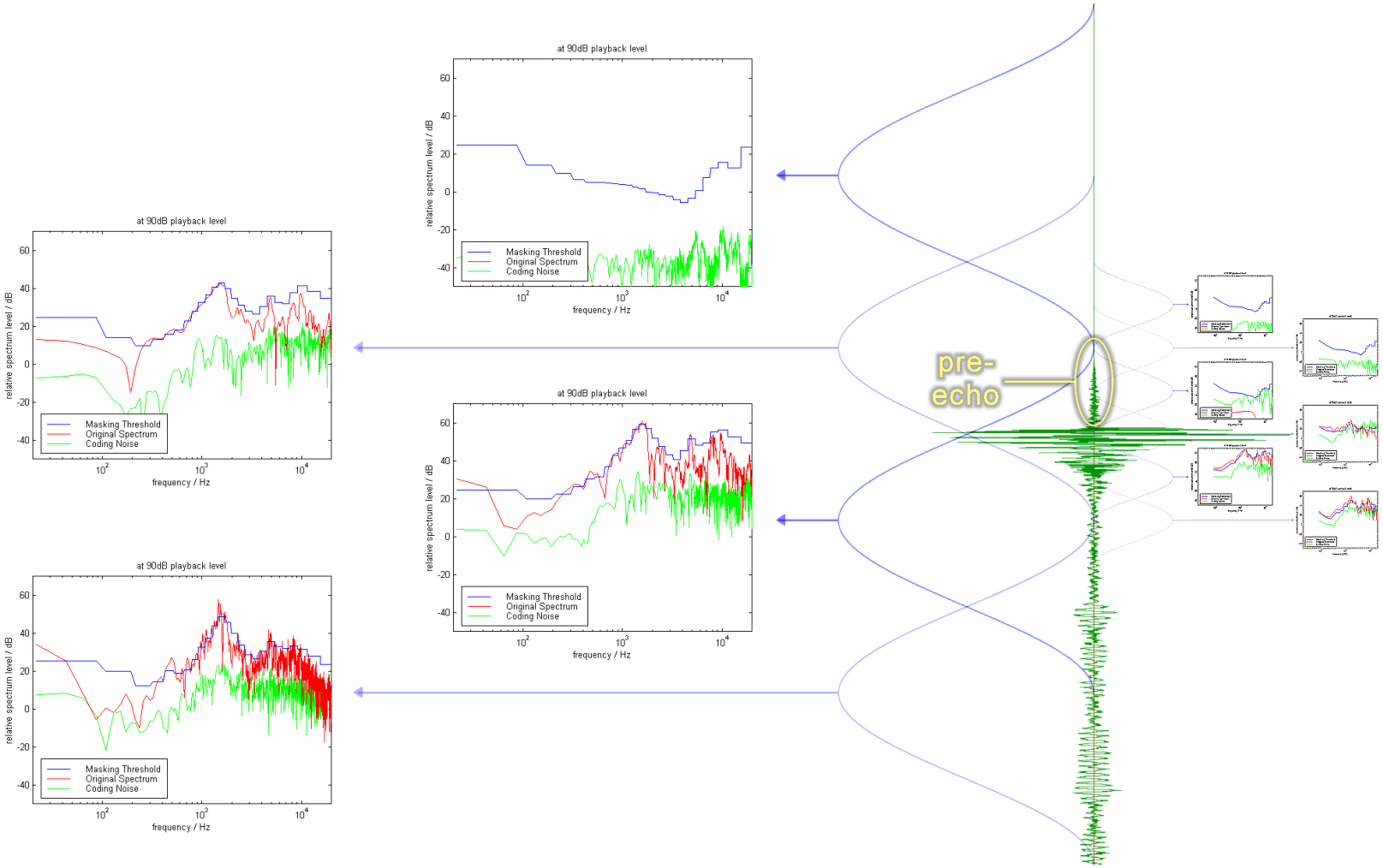


Figure 4.23: 2048-sample window length Johnston analysis of Castanets signal

Mid right: time domain waveform (coded audio), running vertically downwards, with Hanning window functions shown in blue.

Far right: 512-sample analysis (Figure 4.24) shown for comparison.

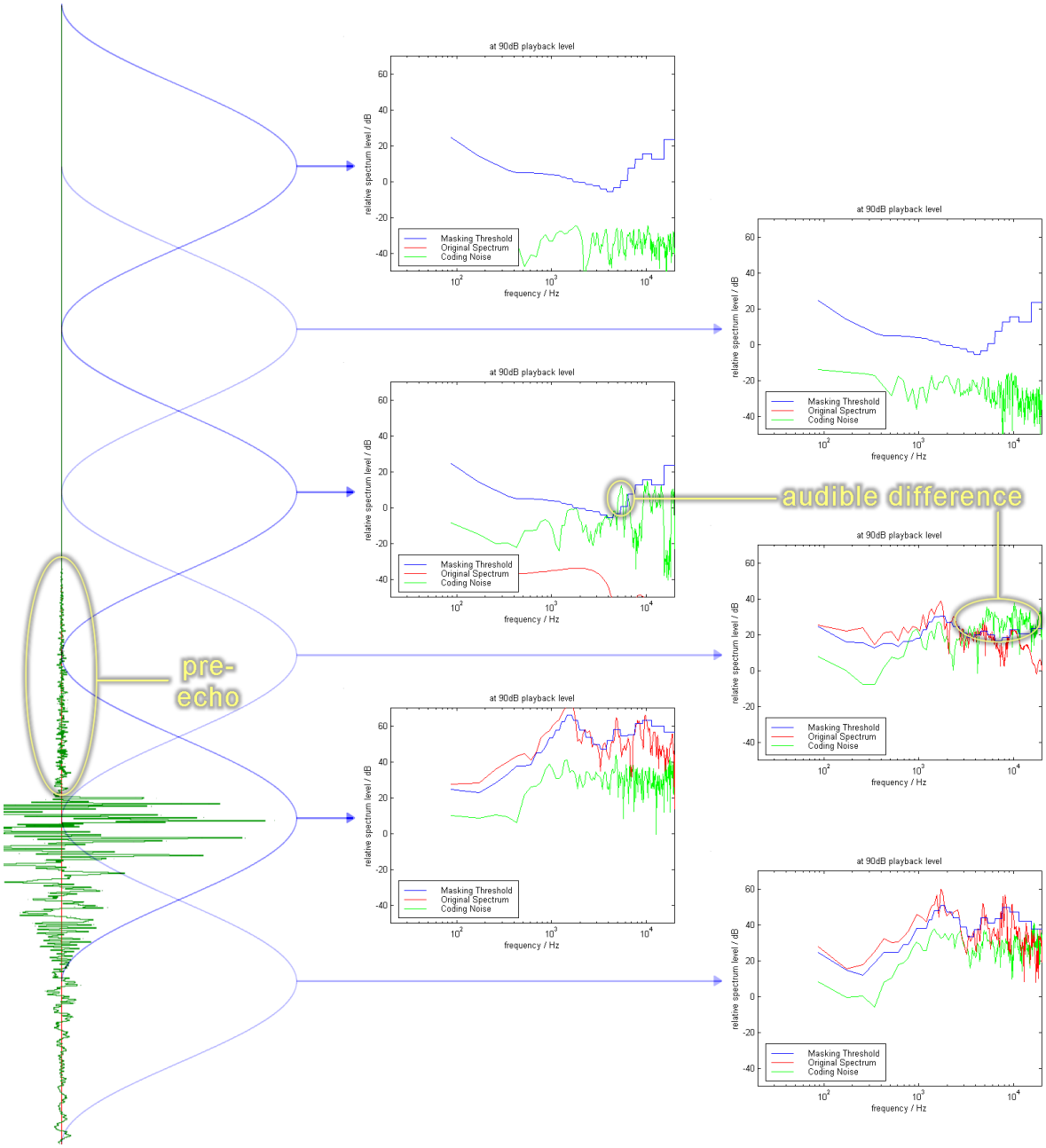


Figure 4.24: 512-sample window length Johnston analysis of Castanets signal

Left: Time domain waveform (coded audio), running vertically downwards, with Hanning window functions shown in blue.

Unfortunately, a window length of 512-samples gives a frequency resolution of 86 Hz, and yields unreliable information about spectral content below 86 Hz. This would be a serious restriction for any high-quality audio assessment tool, so the use of a 512-sample window in isolation can be ruled out.

One solution to the many windowing problems may be to perform the calculations of the Johnston model in parallel using several different window lengths. However, a less cumbersome solution may be to use a basic transformation that matches both the time *and* frequency resolution of the human auditory system. This approach will be considered in the present model.

4.3.4.4 Non-linearity

The human auditory system is non-linear. In engineering terms, this means that the internal representation of signal A plus the internal representation of signal B is not equal to the internal representation of signal A+B. Whilst the masking effects discussed so far may be termed non linear, there are two specific areas where measurable non-linear behaviour may be relevant to an audio quality assessment model.

Firstly, the amount of masking varies with the amplitude, frequency, and duration of the masker. None of these relationships are monotonic, but the variation with amplitude is the most non-linear. The shape of the spectral masking curve is dependent on the level of the masker, such that a louder masker produces *more* masking than a linear masking model would suggest.

Secondly, intermodulation distortion occurs within the cochlea. Two pure tones presented to a listener give rise to distortion components that can be heard by the subject, and detected with a probe microphone. The frequencies of the distortion components are easily calculated, since they are integer difference frequencies (e.g. $2[f_1-f_2]$). However, the amplitude of distortion components does not follow any simple rule. This is discussed extensively in Appendix A.

4.3.4.5 Binaural hearing

Human beings listen via two ears, and most current audio material is stereophonic (2-channel). If this material is replayed via headphones, then a simple channel to ear mapping occurs. Where the material is auditioned using two loudspeakers, a mix of sound from each speaker reaches each ear, the exact ear signals being dependent on the position of the listener and the geometry of the speakers and listening room.

The Johnston model is explicitly a monophonic, or single channel model. The near universal practice when using single channel models with stereophonic signals is to employ two identi-

cal, independent models in parallel, one for each channel. This approach does not match human auditory perception. The neural connections present in the auditory system indicate that information from both ears is compared centrally, as discussed in Section 3.2.5.2. Also, listening tests with human subjects prove that our hearing ability is better than the independent performance of two ears would suggest.

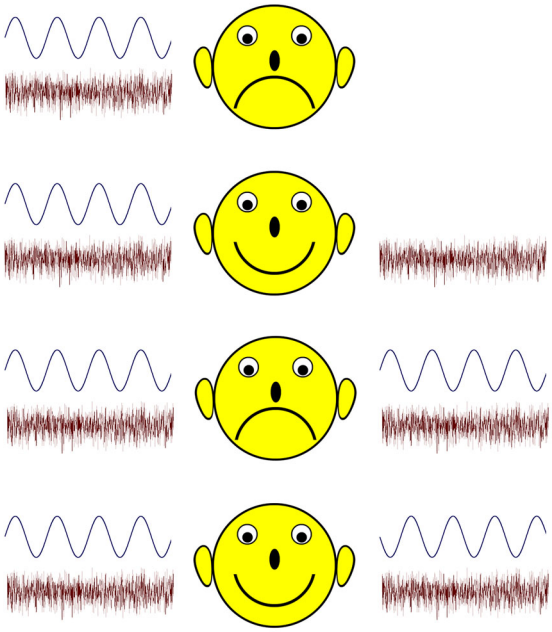


Figure 4.25: Binaural Masking Level Difference

For example, consider the experiment illustrated in Figure 4.25. A tonal signal and a noise masker are presented at one or both ears; the smiling face indicates the condition in which the listener can hear the tone. With both the tone and noise presented to the left ear only, the tone is below threshold, and is inaudible. If, in addition, the noise is presented to the right ear, the listener can now detect the tone in the left ear. At first, this seems counter-intuitive: the total noise power has increased and yet the tone becomes audible. The reason is that the human auditory system compares the signal at each ear, and whereas the noise is common to both ears, the tone is present at one ear only. Therefore, the tone is easier to detect.

This is verified by presenting both signals to both ears; now the listener fails to detect the tone, since there is no advantage gained in comparing the signal at both ears. Finally, the tonal signal to one ear is inverted, and the listener can hear the tone again. In this experiment, the threshold of detecting the tone when it is inverted at one ear is 15 dB lower than that of detecting the tone when it is identical at each ear [Moore, 1997].

Note that if the comparison between ears was perfect, then the threshold for detecting differences between the signals at each ear would be equivalent to the absolute threshold. This is not the case. However, the comparison does yield a significant change in masked thresholds that cannot be predicted by two monophonic models operating independently. This concept will be addressed in the present model.

4.4 Other models

The Johnston model is a single, relatively primitive masking model. Many other models exist, designed for psychoacoustic research, audio coding, and audio quality assessment. Some of the issues raised in the previous section are addressed in some of these models. The purpose of reviewing so many models is to determine which features can be usefully incorporated into the present model, and which features have not been incorporated into any model to date.

The table on the following three pages summarises the features of the most important auditory models.

General notes to Table 4.1

All of the models are purely monophonic, apart from PEAQ advanced [Thiede *et al*, 2000], which is designed to find the loudest error out of the two channels.

The models are compared *as published*. In reality, some of the computational models may contain details and refinements that are not fully described in the short space permitted in an international publication.

In many of the models, frequency dependent weighting and/or pre-filtering accounts for the *shape* of the absolute threshold, whereas internal noise and/or a fixed internal offset accounts for the *level* of the absolute threshold. None of the models account for impaired hearing, or a raised absolute threshold.

Reference	[Colomes *et al*, 1995]	[Dau *et al*, 1996a+b]	[Hollier, *et al*, 1993+1995]	[Johnston, 1988a+b]
Purpose	audio codec + objective perceptual meter	psychoacoustic tool	assessing perceived audio quality[3]	audio codec
outer + middle ear	pre-filtering: ear canal and LPF	-	-	-
frequency transform or **basilar membrane response**	**Fourier transform** *2048 samples* grouped into 600 bands equally spaced on Bark scale according to [Zwicker and Feldtkeller, 1967]	**filter bank** linear basilar mambrane filter bank model, after [Strube, 1985]	**filter bank** one-third-octave filter bank, energy averaged over 4ms	**Fourier transform** *2048 samples* grouped into 24 bands equally spaced on the Bark scale according to [Zwicker and Terhardt, 1980]
spectral masking	**convolution** with excitation curve from linear formula[4]	**arises from** filter bank response and internal noise	**computed directly** via model of frequency discrimination	**convolution** with triangular excitation curve + offset due to tonality of signal
amplitude transform or **hair cell response**	**conversion to dB**	rectification **5 feedback loops control gain**	**mapping via equal loudness curves**	**conversion to dB SPL** fixed replay level inclusion of MAF
temporal masking	-	**arises from** 5 feedback loops and internal noise	**computed directly** via model of temporal masking decay	-
detector	**comparative** psychometric function	**comparative** optimum detection unit	**comparative** advanced error surface analysis	**predictive** direct threshold calculation

[3] This model assumes an audible error (i.e. low quality coded audio) and attempts to judge the subjective impairment due to the error.

[4] In this paper, an excitation curve with a level dependent upper slope is tested, but this approach is rejected, as it decreases performance.

Reference	[Thiede *et al*, 2000] basic	[Thiede *et al*, 2000] advanced	[Paillard *et al*, 1992]
Purpose	"PEAQ" - measuring perceived audio quality	"PEAQ" - measuring perceived audio quality	"PERCEVAL" - Perceptual Evaluation of audio quality
outer + middle ear	frequency dependent weighting function	pre-filtering and frequency dependent weighting	multiplication of energy spectrum
frequency transform or **basilar membrane response**	**Fourier transform** *2048 samples* grouped into 0.25 Bark bands according to [Schroeder *et al*, 1979]	**filter bank** 40 filters, equally spaced on Bark scale according to [Schoeder *et al*, 1979]; upper slope level dependent after [Terhardt, 1979], with temporally smoothed dependency.	**Modulated lapped transform** *1024 samples @ 44.1 kHz* sliding average taken for each freq component over 3 frames. Transformed to 2500 point Bark scale.
spectral masking	**computed directly** non-linear supposition of local level + frequency dependent smearing of excitation pattern (plus fixed offset simulating internal noise)	**arises from** filter bank response, effective rectification, plus frequency dependent offset simulating internal noise.	**computed directly** Bark scale representation filtered *along* frequency bands to spread energy. Internal noise added.
amplitude transform or **hair cell response**	**conversion to dB** adjusted for replay level.	**conversion to dB** adjusted for replay level	**logarithm** of basilar spectrum
temporal masking	**temporal smearing** frequency dependent IIR filter smears excitation pattern Unfiltered value used if larger	**temporal smearing** 8ms FIR low pass and 4ms first order IIR low pass filters smear temporal signal	-
detector	**comparative**	**comparative** probability calculated from smoothing probability in each band over time and taking maximum	**comparative** statistical detection in each band and across all bands

Table 4.1: Comparison of auditory perceptual models

Reference	[Patterson *et al*, 1992a,b]	[Rimell and Hawksford, 1996] [Beerends and Stemerdink, 1992]	[Robert and Eriksson, 1999][5]	[Zwicker and Zwicker, 1984]
Purpose	"Auditory Image Model" – psychoacoustic tool	perceptual audio quality assessment	psychoacoustic tool	general perceptual audio transform[6]
outer + middle ear	middle ear filtering	Full room simulation and HRTF filtering	first order butterworth bandpass filter 1.4 kHz – 20 kHz	-
frequency transform or **basilar membrane response**	**filter bank** gammatone filterbank *or* transmission line filterbank, spectral sharpening and compression	**Fourier transform** *40ms, Hanning windowed* convert to power warp frequency scale to Bark scale	**filter bank** active gamma-tone filters filter shape varies with complex feedback from adjacent frequency channels	**frequency to Bark transform**
spectral masking	***may* arise from** frequency transform, which *may* cause spectral masking if some internal noise or uncertainty is added.	**convolution** with a unique (to this model) spreading function	**arises from** filter bank and random nature of inner hair cell firing	**computed directly** excitation level implies it
amplitude transform or **hair cell response**	**compression** and 2-D adaptive thresholding *or* **inner hair cell simulation**	**specific compressed loudness** calculated via an equation which simulates equal loudness curves	**inner hair cell simulation** includes half wave rectification, low pass filtering, and adaptation.	**specific loudness transform**
temporal masking	***may* arise from** strobed temporal integration *or* correlogram which *may* give rise to some masking	**temporal smearing** Frequency dependent temporal smearing between frames	**arises from** adaptation and random nature of inner hair cell firing	**temporal smearing** specific loudness smoothed over time
detector	-	**comparative** perceived noise = log (difference between internal representations of inputs, averaged over time)	-	-

[5] This model post-dates the monophonic section of the model developed in the present thesis [Robinson and Hawksford, 1999].

[6] This paper does not present a model *per se*, however, it contains principles for transforming audio into a perceptually relevant internal representation.

4.4.1 Comparison of models

Whilst the models vary in complexity and intended purpose, Table 4.1 shows that many have common features. Most pre-filter the audio signal; all use some kind of time to frequency transform, with a near logarithmic internal frequency scale; all convert the signal amplitude into some kind of logarithmic representation. These processes correspond to stages within the auditory system, discussed in the previous chapter.

Some of the models calculate spectral masking explicitly, and most simulate temporal masking with a smearing, spreading, or low pass function. However, some of the models simulate neither spectral nor temporal masking directly. Rather, they simulate the processes within the HAS which give rise to masking, and let the masking arise as a consequence (see Section 4.2).

The number of frequency bands or bins varies dramatically, from nine [Hollier, *et al*, 1993+1995] to 600 [Colomes *et al*, 1995]. The models employing hundreds of bands do not appear to exhibit better performance than those using tens of bands[7]. Both PEAQ versions [Thiede *et al*, 2000] predict human perception admirably. The filter bank version of PEAQ yields the more accurate prediction of human perception, though the filter bank version uses only 40 bands, while the inferior FFT version uses 100. However, the amount of information generated by 40 filters is much greater than that generated by a 1024 sample FFT summed into 100 frequency bins. In particular, the filter bank generates continuous time-domain signals, whereas the FFT generates discrete snapshots. Post-processing may recover further frequency information from the output of a filter bank, whereas FFT components, once summed into coarser bands, cannot be decoded to yield further time or frequency domain information. For this reason, if a filter bank model is chosen, fewer frequency bands will be employed than if an FFT model is developed.

All the models in Table 4.1 except Johnston use a comparative measure to calculate the audible difference between two signals. This approach will be used in the present model, since the predictive measure is unreliable, as demonstrated in Section 4.3.2.

[7] This conclusion carries the following caveat; all authors conclude their papers with positive reports of their model's performance, but few include a critical comparison with other models.

[Rimell and Hawksford, 1996] (from [Beerends and Stemerdink, 1992]) and [Thiede *et al*, 2000] (basic) represent the most advanced attempts to simulate the effects of the human auditory system, *without* simulating any components of the auditory system directly. This is one possible approach for the present work. However, [Thiede *et al*, 2000] (advanced) move away from this approach, and closer to the actual workings of the auditory system by replacing the FFT by a filter bank. This development reduces computational efficiency, but improves performance. There is extensive data in [Thiede *et al*, 2000] demonstrating that the advanced model matches human perception more closely than the basic model. For this reason, the present model will be designed to closely match the processes found within the human auditory system.

The use of a filter bank removes all the problems associated with windowing, which were amply demonstrated throughout Sections 4.3.2 and 4.3.3. However, the level-dependent non-linearity of spectral masking is easier to simulate in an FFT based model, where the spectral masking is calculated directly. To incorporate this non-linearity into a filter bank based model, the filter bank itself must be non linear, as in [Thiede *et al*, 2000].

Most of the models in Table 4.1 include a filter that smears the signal envelope in order to simulate temporal masking. Whilst this yields a first-order approximation, the aim is to predict temporal masking more accurately in the present model.

Finally, none of the models in Table 4.1 address spatial masking or binaural hearing. This is a significant area in which the present model aims to improve upon the performance of existing models.

4.5 Conclusion

In this chapter, the concept of auditory modelling has been introduced. A simple model has been tested, and shown to give satisfactory performance in simple psychoacoustic tests. The model was unable to reliably assess the audio quality of codecs containing a more advanced auditory model. Current state-of-the-art models have been compared, and suggestions have been made for areas where the present work may supersede the performance of existing models. These include spatial masking, temporal masking, and non-linearity.

5

Monophonic Auditory Model

5.1 Overview

In this chapter, an auditory perceptual model for assessing the perceived quality of coded audio signals is described. This model simulates the functionality of the physiology found within the human ear. The passage and transformation of sounds from the free field to their neural representation is simulated. The neural signal is processed by a novel detection network. Finally, the model is calibrated using known psychoacoustic data.

5.2 Introduction

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, as discussed in detail in the previous chapter. 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 of 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.

If the processing within the model is similar to that present within the human auditory system, then the model may accurately predict the perception of a human listener. For this reason, the present model will simulate 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, arise as a consequence of these processes.

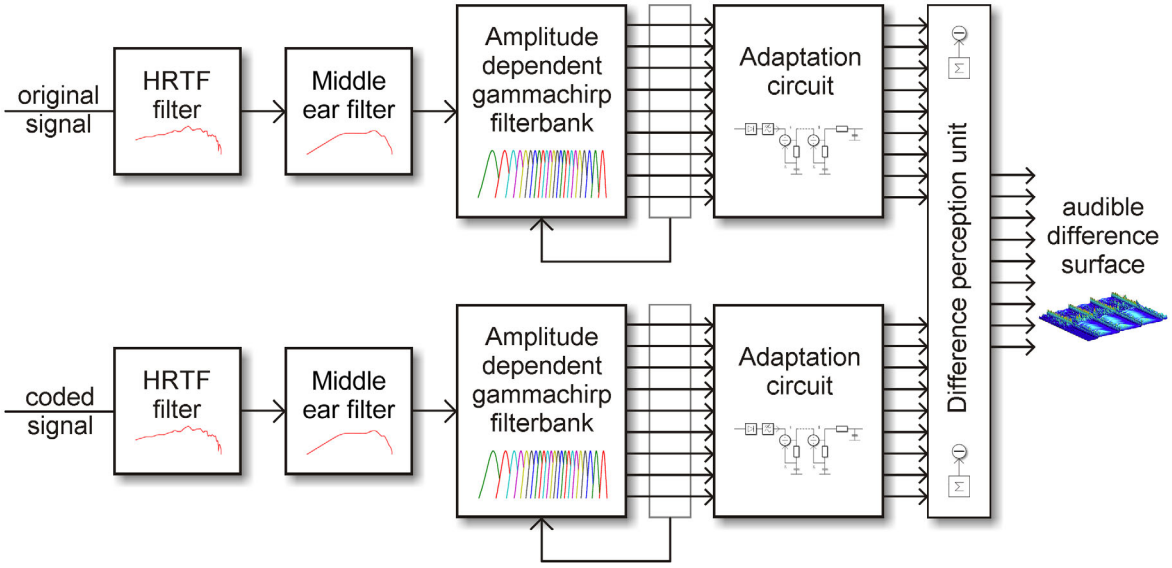


Figure 5.1: Overall structure of the monophonic model

5.3 Structure of the model

The model presented here is based upon the processing present within the human auditory system, as described in Chapter 3. The structure of the auditory model is shown in Figure 5.1. Each individual component is described in the following sections.

5.3.1 Signal Preparation and Implementation

The model is programmed using the MATLAB environment, and all code is included on the accompanying CD-ROM. All input signals are fed to the model via Microsoft .wav format files, and where appropriate, intermediate results are stored in 32-bit accurate .wav files. The MATLAB functions for reading and writing .wav files have been modified by the author to allow the use of 32-bit .wav files, and to enable truly random read/write access to the files.

Any D.C. offset is removed from the input signal before further processing. Where two signals are presented for comparison, the signals are time aligned to the nearest sample. This process is carried out either by manipulating the two waveforms within a sample-accurate audio editor [Syntrillium, 2000], or automatically by cross correlation, using the MATLAB routine *waveta*, which is included on the accompanying CD-ROM.

Source Material	*ms*	gain relative to SMPTE RP 200
Default setting	90 dB	-13 dB
Chart CD single	89 dB	-14 dB
Pop CD album	91 dB	-12 dB
Classical CD	99 dB	-4 dB
Audiophile CD	101 dB	-2 dB
SMPTE RP 200	103 dB	0 dB

Table 5.1: Suggested values of *ms* for different musical genres, and corresponding gain adjustments relative to SMPTE RP 200 reference

The relationship between the digital representation of the waveform and the real world sound pressure is controlled by the variable *ms* (maximum SPL). The ear's response to a signal presented at 40 dB is very different to the same signal presented at 90 dB, so this information is required by the model. *ms* is defined as the sound pressure created by a digital full scale sine wave at the centre of the listener's head, given in units of dB SPL. The default value is 90 dB. This is chosen because it is a typical listening level for popular music, and the theoretical noise-floor of CD quality audio is inaudible at this level.

The value of *ms* is effectively a measurement of the monitor gain of the audio system through which the signal is (hypothetically) being auditioned. In simple terms, it communicates the position of the listener's volume control to the model. Surprisingly, there is no accepted standard within the audio industry which relates digital signal amplitude to real world SPL. However, the film industry does have such a standard [SMPTE RP 200, 1999]. Table 5.1 lists approximate values of *ms*, and the appropriate gain adjustments relative to the SMPTE standard, for various styles of music. The different values are not due to musical style *per se*, but due to the manner in which different musical genres are typically mastered at different levels on CD.

It is very important to set *ms* correctly, especially when testing classical music. If the default value of 90 dB is employed, the lowest 10dB of dynamic range will fall below absolute threshold. All signals below absolute threshold are ignored by the model, so any "audible" problems in this range will be missed. However, a real human listener would increase the volume to

listen to the quieter classical CD, thus raising any problems within the lowest 10dB up into the audible range. Hence *ms* must be increased correspondingly, or the model will give misleading results.

In yet another scenario, *ms* may be lowered to simulate quiet (late night) listening, where lower quality audio may be acceptable. Thus, the specification and inclusion of *ms* within the model yields new applications and flexibility.

5.3.2 Pre-filtering

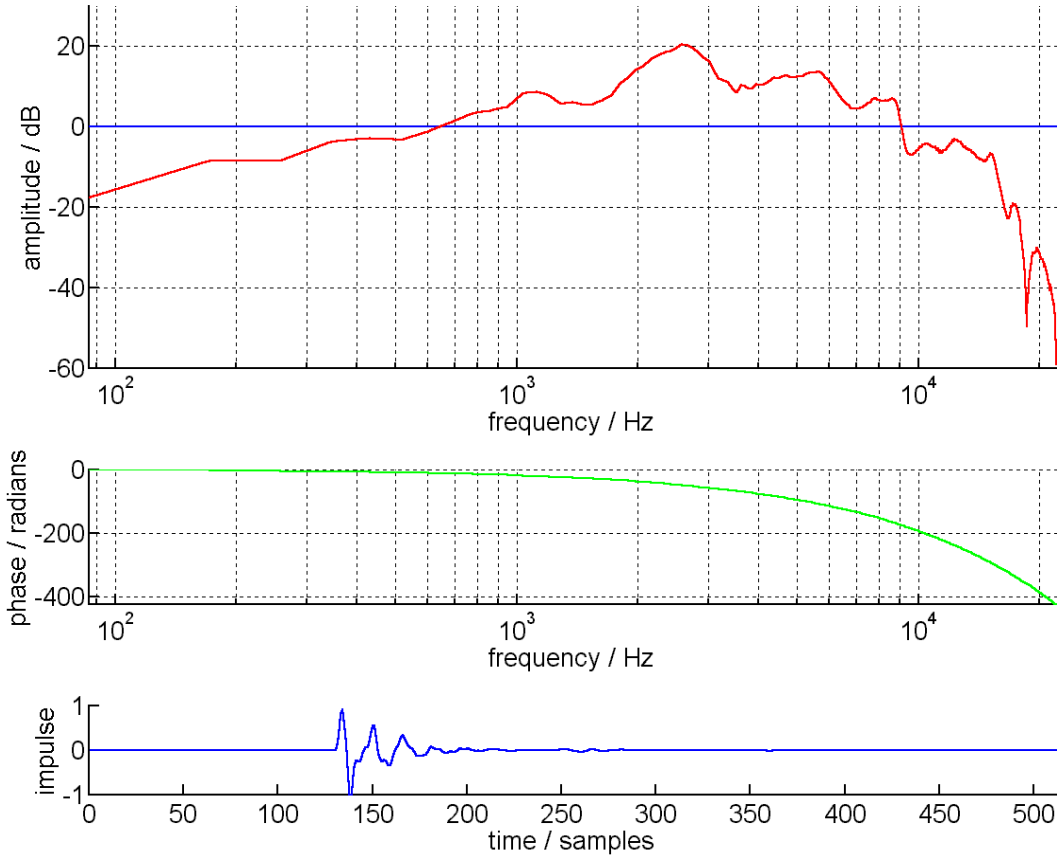


Figure 5.2: HRTF 30°

amplitude, phase, and impulse response; f_s = 44.1 kHz

The filtering of the pinna and ear canal is simulated by an FIR filter, derived from measurements carried out via a KEMAR dummy head. An arbitrary angle of incidence is chosen, in this case 30°, yielding the response shown in Figure 5.2. The KEMAR measurements are used because they are readily available for this research. Measurements from human subjects could

be used as a more accurate and realistic alternative (e.g. [Hammershøi *et al*, 1992] and [Møller *et al*, 1995b]).

5.3.3 Middle Ear filtering

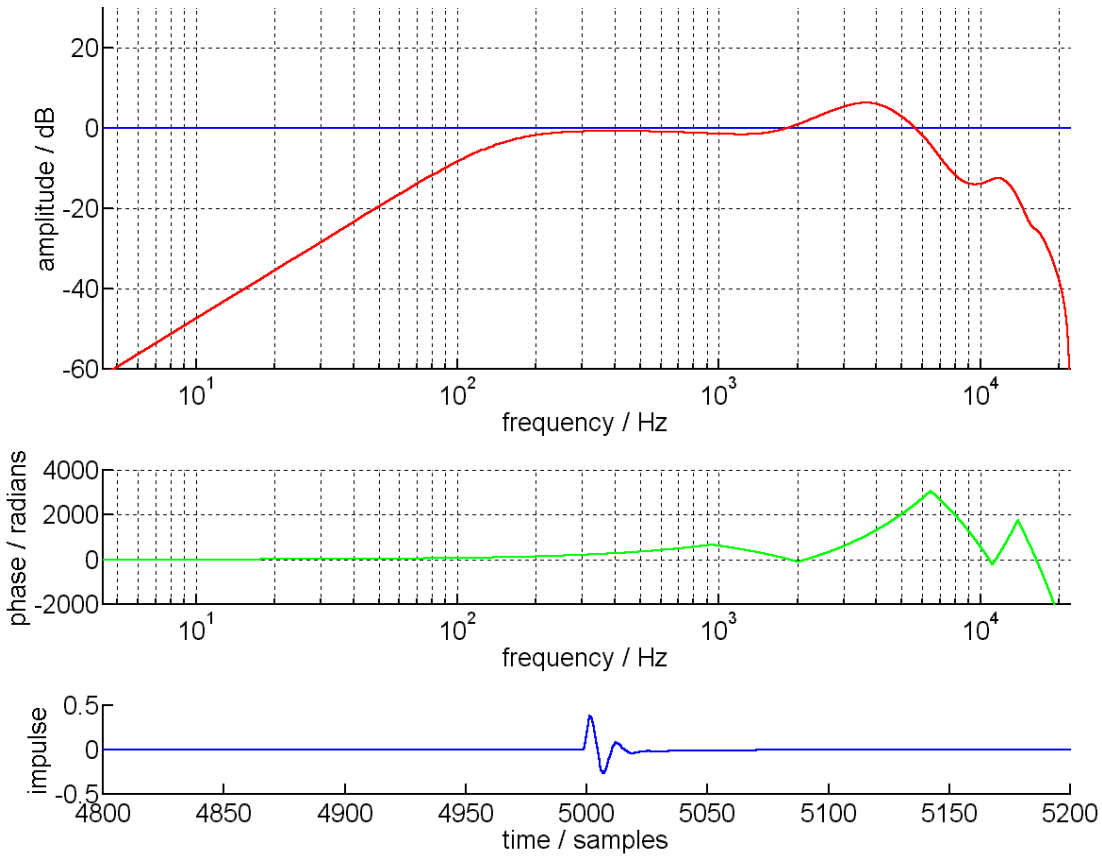


Figure 5.3: Equal Loudness filter to simulate outer- and middle-ear frequency response

A filter based upon inversion of the Equal Loudness curves [Robinson and Dadson, 1956] provides a good approximation to the response of the outer *and* middle ear (see [Moore *et al*, 1997] for further discussion). The equal loudness curves were discussed in Chapter 3, and a plot of these curves is shown in Figure 3.5.

An IIR filter was designed to match the target response using the MATLAB `yulewalk` function. The `yulewalk` function only considers the amplitude response when optimising the filter, hence the phase response is erratic. A stable and computationally efficient filter was produced by cascading a 10-point yulewalk-designed IIR filter with a 3-point 2nd order butterworth high-pass IIR filter. The latter was employed to produce the desired low frequency attenuation, which would have required an excessive number of taps if implemented within the

first filter. This dual-filter method has the advantage that, should the low frequency response be in error (due to internal noise, as discussed in Chapter 3), then it can easily be altered without re-designing the entire filter. The response of the combined filter is shown in Figure 5.3.

This filter simulates the response of the ear canal and middle ear. However, the HRTF filter also includes the ear canal response, and comparison of Figure 5.2 and Figure 5.3 confirms that both filters contain a common band-pass component. If the pinna response is not required (for example, when simulating headphone listening) then the HRTF filter should be dispensed with. If the pinna response is required, then a diffuse normalised HRTF may be used[1], since diffuse normalisation removes the ear canal response. (See [Robinson and Greenfield, 1996] for further details.) Both raw and diffuse normalised HRTFs are included on the accompanying CD-ROM.

5.3.4 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 [Irino and Patterson, 1997], and simulates the response of the BM at a given point.

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

$$g_c(t) = at^{n-1}\exp(-2\pi b\,\mathrm{ERB}(f_r)t)\cos(2\pi f_r t + c\ln(t) + \phi) \qquad (5\text{-}1)$$

[1] The MAF measurements were carried out using a single loudspeaker in front of the listener. For this reason, a frontal HRTF response should be embedded within these measurements. Whilst there are some traces of the 0° HRTF visible in the MAF curve, it is not an accurate representation of the response. Either the MAF data has been smoothed, or the listener subconsciously corrected for the HRTF. The latter is probable. In everyday life, the varying frequency response of sounds from different directions is hidden from us by the auditory system. The direction dependent frequency response (HRTF) is decoded, such that we perceive the correct direction, and a flat frequency response. For this reason, it is assumed that the MAF does not contain an HRTF response, although some trace of the HRTF may be present. Hence, a diffuse normalised HRTF is employed in the simulation of free-field listening. If a future MAF measurement clearly contains an accurate frontal HRTF response, then it may be use in isolation, or with a front-normalised 30° HRTF. The reader is referred to [Robinson and Greenfield, 1996] for an extensive discussion of HRTF normalisation.

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

$$\mathrm{ERB}(f_r) = 24.7 + 0.108 f_r \tag{5-2}$$

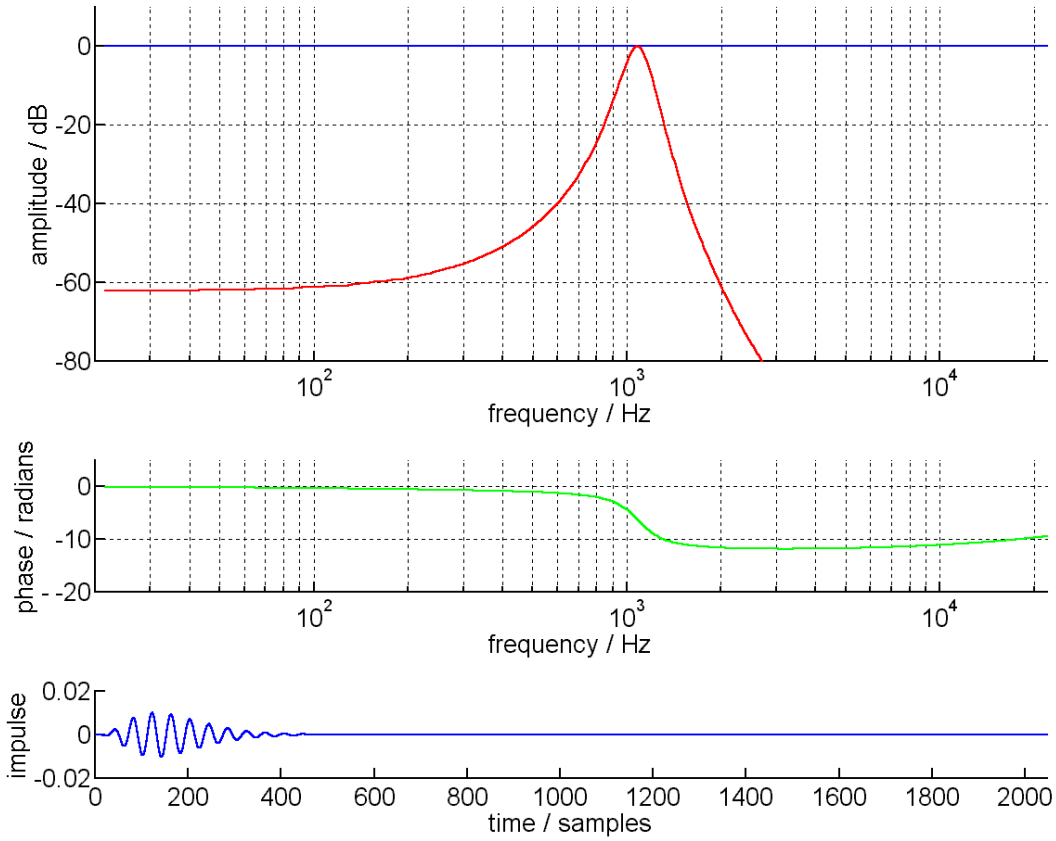


Figure 5.4: Gammachirp filter

amplitude, phase, and impulse response f_s = 44.1 kHz

In [Irino and Patterson, 1997], 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 (though an implementation issue of this approach will be discussed in Section 5.3.4.2). The term *c* ln(*t*) causes the amplitude dependency, where *c* is proportional to the amplitude of the signal in each band. This will be discussed in Section 5.3.4.3. A single gammachirp filter is shown in Figure 5.4.

5.3.4.1 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

$$z = [26.81 f /(1960 + f)] - 0.53 \tag{5-3}$$

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

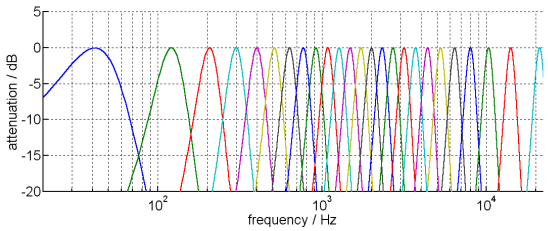
$$z' = z + 0.22(z - 20.1). \tag{5-4}$$

This definition of the Bark scale is taken from [Traunmüller, 1990]. Comparison with measured data [Zwicker, 1961] shows that this formula is accurate to within ±0.1 Bark. This is more accurate than any other available formula (c.f. the auditory models reviewed in the previous chapter). Formulae (5-3) and (5-4) have two further advantages. Firstly, they are computationally simple, and secondly, they are easily inverted to give a transformation from Bark to linear frequency, ensuring that a conversion from frequency > Bark > frequency will yield the original frequency.

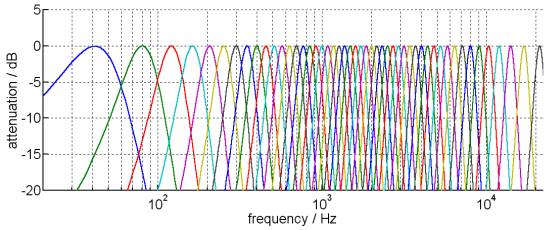
It is debatable as to what spacing (and hence number) of filters are needed to match the frequency resolution of the human ear. It has been suggested that humans can differentiate about 620 frequencies, equally spaced in the bark domain [Colomes *et al*, 1995] from [Zwicker and Feldtkeller, 1967]. However, 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 (or pitch, since this is what humans perceive), irrespective of the resolution of the BM. In addition, a pure tone of *any* audible frequency will excite a region of the BM wider than one Bark. Thus, if the spacing of the filters is equal to one Bark, it will be possible to estimate the frequency of a single tone by comparing the energy levels in adjacent filter-bands.

Hence, it is unnecessary to simulate 620 points on the BM via a bank of 620 gammachirp filters in order to match human perception. Instead, the number of filters is chosen such that all frequencies are transduced without significant inter-filter gaps.

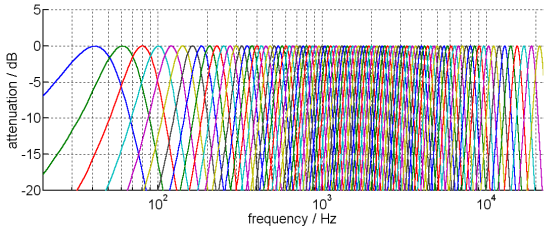
Choosing a spacing of ½ Bark causes all frequencies between 100 Hz – 16 kHz to be within 3 dB of the resonant peak of a filter (See Figure 5.5 (b)). Decreasing the filter spacing to ¼ bark causes *all* audible frequencies to be within 1 dB of the resonant peak of a filter (See Figure 5.5 (c)). At this spacing 96 filters are needed to cover the full audible range.



(a) – 1 Bark spacing (24 filters)



(b) – ½ Bark spacing (48 filters)

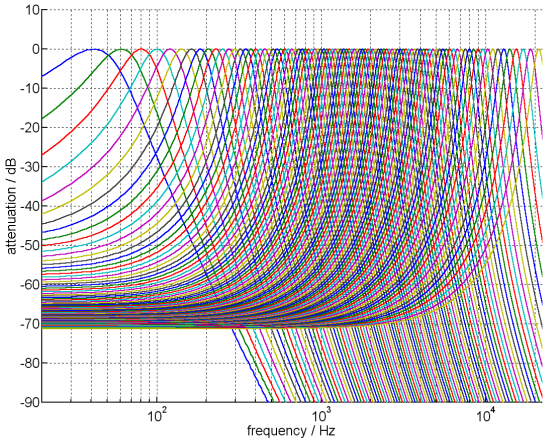


(c) – ¼ Bark spacing (96 filters)

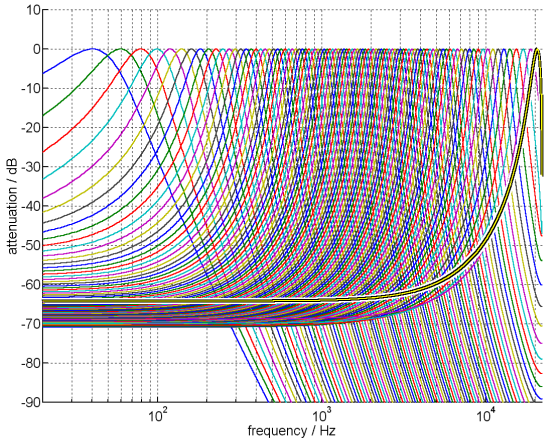
Figure 5.5: Spacing of gammachirp filters

5.3.4.2 Implementation issues

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.



(a) – ideal gammachirp filter bank



(b) – sampled filterbank, f_s = 44.1 kHz

Figure 5.6: Effect of Nyquist sampling on gammachirp filter bank

Two problems arise when a gammachirp filter bank is generated in a digitally sampled domain. Firstly, filters acting at or near the Nyquist frequency (half the sampling rate) can behave unexpectedly. Figure 5.6 compares the response of an ideal filter bank with that of a filter bank sampled at 44.1 kHz. The response of all the filters increases slightly near the Nyquist limit (right hand side of the graph), but this deviation is trivial. More significantly, the lower stop-band of the highest frequency filter is increased by 7 dB compared to the ideal filter. When auditioning a typical coded audio signal (which is down at the noise floor by 20 kHz), this will cause the level in this band (due to leakage from lower frequency components) to be twice as loud as the adjacent band. Examining the impulse response of the filters reveals that the highest frequency filter lies closest to $t = 0$, and has a relatively high amplitude first sample. Windowing the start of the filter does not help, because any reasonable length window function will obliterate the majority of the non-zero impulse response. Using a sample rate of 96 kHz solves the problem, as shown in Figure 5.6 (a). This is one possible solution to the problem, though it is more convenient to operate the model at the same sample-rate as the input signal (usually 44.1 kHz).

However, a filter centred over 20 kHz is irrelevant for most audio signals and listeners, so may be ignored unless specifically required (e.g. simulating response of young listeners[2] or a high sample-rate system). In addition, the hearing of older listeners can be crudely simulated by discarding several of the highest frequency filters, as appropriate.

[2] Many researchers have argued that there are no auditory filters centred above 15 kHz, and that the ability of some listeners to hear higher frequency sounds can be attributed to the output of the 15 kHz centred filter (e.g. [Moore *et al*, 1997] and [Zhou, 1995]). Whilst this may be true for the majority of young listeners, it is difficult to believe that the exceptional individuals who can hear up to 25 kHz are detecting such frequencies through a 15 kHz centred filter, since such a filter attenuates 25 kHz by 90 dB! A 20 kHz centred filter would attenuate 25 kHz by 40dB, so the presence of this filter in some gifted individuals is a more plausible explanation.

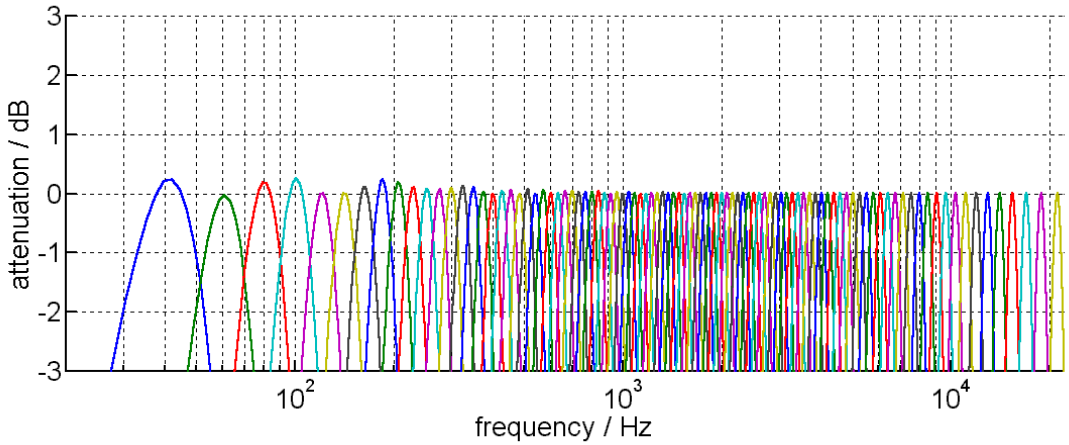


Figure 5.7: Ripple in filterbank due to finite frequency resolution of FFT

The second problem in the sampled implementation of the gammachirp filter bank arises from the normalisation of the filters to 0 dB at their centre frequency. This is carried out by FFTing each impulse response, and normalising to the maximum value. However, the FFT yields a discrete frequency domain representation. As with any discrete representation, the peak value of the FFT does not necessarily coincide with the true peak value of the filter response, which may lie between discrete points. The effect is most pronounced at low frequencies, where fewer frequency domain values define each filter. This results in errors in the normalised peak values, as shown in Figure 5.7. These errors are small, and could be ignored. However, they may be avoided by padding the impulse response with zeros before the FFT, which has the effect of oversampling the response in the frequency domain, and hence causes a value closer to the true peak value to be present.

5.3.4.3 Amplitude dependence

The frequency response of each point on the basilar membrane is amplitude dependent. In the model, this is simulated by adjusting the shape of each gammachirp filter continuously in response to the input stimulus.

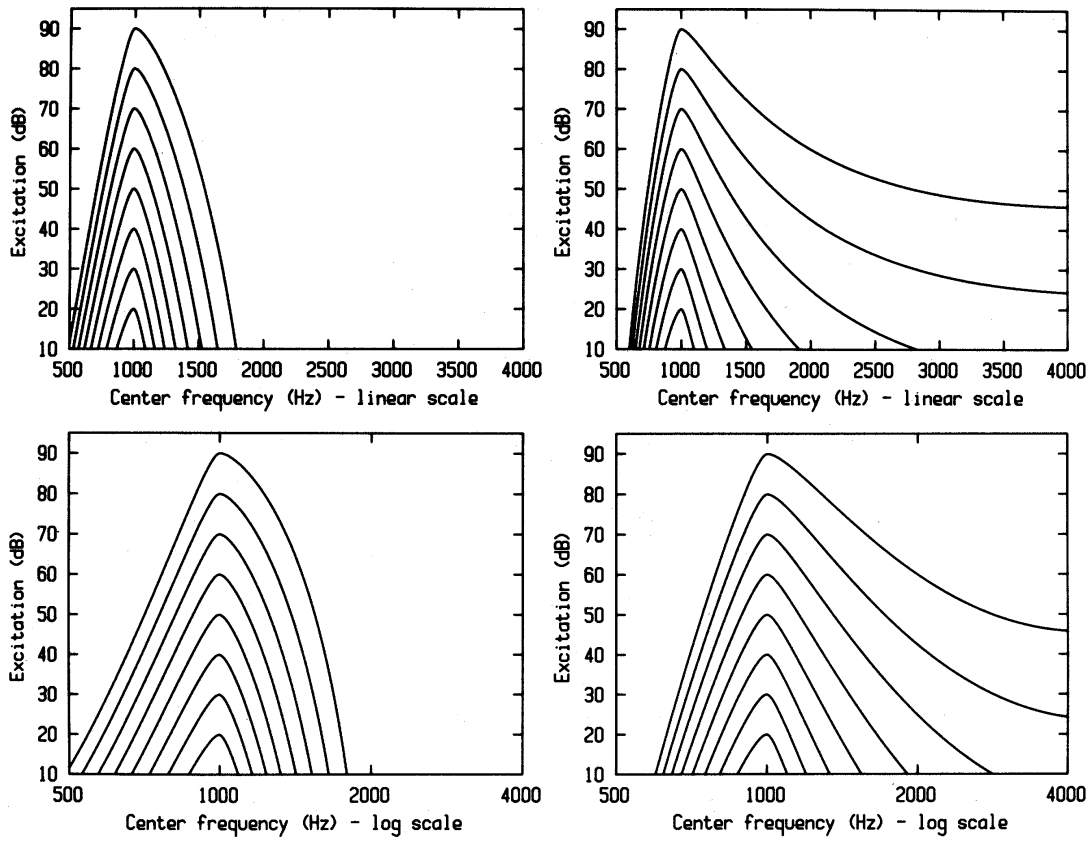


Figure 5.8: Dependence of excitation upon stimulus level

upper: linear scale; lower: logarithmic scale.

left: filter shape varied according to output; right: filter shape varied according to input.

5.3.4.3.1 Physiological mechanism

In [Moore and Glasberg, 1987], it is shown that the response of each filter varies according to the level of the *input* to the filter, rather than its output. This is demonstrated in Figure 5.8 [Moore, 1995], where the left-hand panels show excitation patterns calculated on the assumption that the level at the *output* of each filter determines its shape, and the right-hand panels show excitations patterns calculated on the assumption that the level at the *input* of each filter determine its shape. Only the right-hand panels show the well-known upward spread of excitation in response to the louder signals.

This would be acceptable proof of a real physical mechanism, were it not for the fact that the assumption underlying the right-hand panels is physiologically impossible! To demonstrate this, consider a pure tone stimulus. The assumption suggests that a filter at one point on the basilar membrane changes its response as the level of the input to that filter increases. This is required behaviour, even if the input signal is rejected by the filter in question. Recall from

Chapter 3 that the filter response is partially due to the mechanical properties of the BM, and that the detection of the movement of the BM is entirely due to the hair cells upon on it. Thus, it is impossible for the filter at any given point to "see" its input if it is rejected by the filter, since a rejected frequency component will cause no oscillation of the BM at that point. The auditory system has no knowledge of the signal *before* the filter, since the detection mechanism lies after the filter[3].

This suggests that the auditory system cannot change the shape of each filter in response to the *input* level. Unfortunately, [Moore and Glasberg, 1987] show that the *output* level of a given filter cannot explain its change in shape either.

The answer to this apparent contradiction is intuitively simple. The feedback mechanism within the auditory system responsible for changing the filter shape with stimulus level is (in all probability) driven from the outputs of *all* the filters. It seems likely that the mechanism determines the maximum output level in this manner, and then adjusts *all* the filters accordingly. This is physiologically sensible; the maximum displacement of the BM is detected, and the movement of the BM *as a whole* is damped in response. This explanation gives the required (known) excitation patterns, without suggesting that the auditory system must have some pre-filter knowledge of the input signal – knowledge which cannot be explained from known physiology.

5.3.4.3.2 Implementation

The output of each filter corresponds to the instantaneous displacement of the basilar membrane at that point. Hence, the largest output from the bank of filters is equivalent to the maximum instantaneous displacement of the basilar membrane. However, changing the filter shape in response to this instantaneous value would cause the filter shape to oscillate with the cycles of a sine wave input signal. This behaviour is undesirable, and causes the model to be unstable. For this reason, it is necessary to apply a peak detect and hold algorithm to the output

[3] It is likely that the active feedback mechanism within the cochlea (which causes the excitation pattern to be amplitude dependent) is driven by information from the stereocilia, rather than the inner hair cells themselves. However, the stereocilia are reacting to BM movement, so they are still "post-filter". Thus the argument remains valid whichever detector is employed.

of each filter *before* determining the maximum displacement value. Thus, the model tracks the envelope of each band, rather than the instantaneous amplitude.

The approximate frequency at the output of each filter is known (due to the gammachirp filter having an approximately band-pass response), so the waveform peaks in each band are separated by a known distance. Thus, the envelope can be tracked by "riding the peaks" of the waveform. This is implemented as follows.

hw is defined as half the wavelength (in samples), given by

$$hw = \frac{fs}{2\,fc} \tag{5-5}$$

where *fs* is the sampling frequency, and *fc* is the centre frequency of the gammachirp filter. Then the time-domain envelope in each band is given by

$$fb\_pks(band, t) = \max[\mathrm{abs}[fb\_output(band, t\text{-}hw : t)]] \tag{5-6}$$

where *fb_output* is the time-domain output of the filter, and *band* is an index to each filter (band). A value of one half wavelength is used because the rectified (absolute) waveform contains two peaks per wavelength (cycle), and the algorithm only needs to look back to the previous peak to ride the envelope correctly. This gives a rapid response, whilst maintaining model stability.

Finally, the maximum envelope across all bands is calculated, thus

$$fb\_pks\_mx(t) = \max[fb\_pks(1{:}nbands, t)]] \tag{5-7}$$

This gives the value of the maximum displacement of the basilar membrane, which is used to change the shape of the gammachirp filters.

The value of *c* in equation (5-1) is calculated from the peak value, *fb_pks_mx(t)*, thus,

$$c(t) = 3.29 - 0.059 * [20 \log_{10}(fb\_pks\_mx(t) + ina) + ms] \tag{5-8}$$

The numerical values in this equation were obtained by fitting the filter shapes generated by equation (5-1), to those measured using human subjects (shown in Figure 5.2) for the range of *c* and *fb_pks_mx(t)* covered by the human performance data. *ina* is included to prevent the calculation of log(0) during a silent input signal. The value of *ina* is just below the minimum audible threshold (i.e. *ina*udible), and has a negligible effect on supra-absolute-threshold calculations. *ms* refers to the play-back level, as discussed in Section 5.3.1.

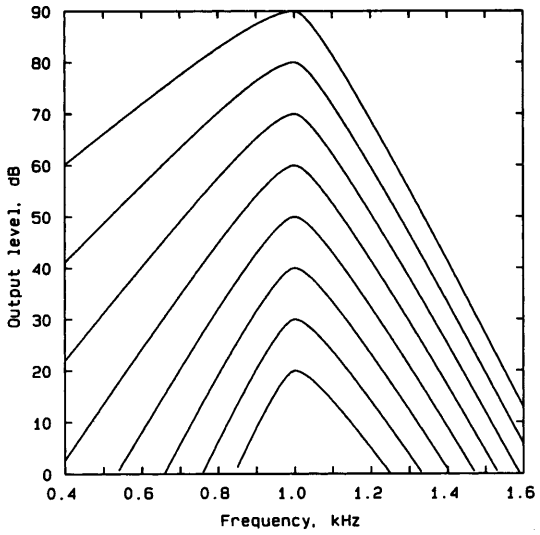


Figure 5.9: Amplitude dependent response of basilar membrane [Moore, 1995]

5.3.5 Hair cell transduction

At each point along the basilar membrane, its movement is transduced by a number of hair cells, as discussed in Chapter 3. In the model, this will be simulated by appropriate processing of the output of each gammachirp filter. Subsequent processing will take place in parallel, in as many bands as there are gammachirp filters (see Section 5.3.4.1).

5.3.5.1 Meddis Hair Cell model

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 [Meddis, 1986], [Meddis, 1988], and [Meddis *et al*, 1990] simulates the random firing of individual cells, and the probability of each cell firing. A MATLAB implementation of this model is included on the accompanying CD-ROM.

A diagram of the model is shown in Figure 5.10. The model tracks transmitters, which are released from the hair cell into the Post Synaptic Afferent Fibre (which leads on to the auditory nerve). The probability of a transmitter being released from the hair cell is proportional to the number of transmitters available within the hair cell, and its instantaneous displacement. In the gap (Synaptic Cleft) between the cell and the fibre, some transmitters will be lost, and others will return to the hair cell where they will be reprocessed and made available for subsequent

re-release. The number of transmitters within the Synaptic Cleft determines the firing probability of the fibre[4].

This model accounts for the hair cell response discussed in Chapter 3, as illustrated in Figure 5.11. At the onset of a sound, there are many transmitters available within the hair cell for release. These transmitters immediately flood the Synaptic Cleft, giving a large peak in the probability of firing. The number of transmitters in the Cleft decreases as the sound persists, and eventually decays to a steady state, where the number of transmitters released into the Cleft is balanced by the number lost or returned to the hair cell. With the correct parameters, the model provides a good match to measured mammalian individual hair cell performance [Meddis *et al*, 1990].

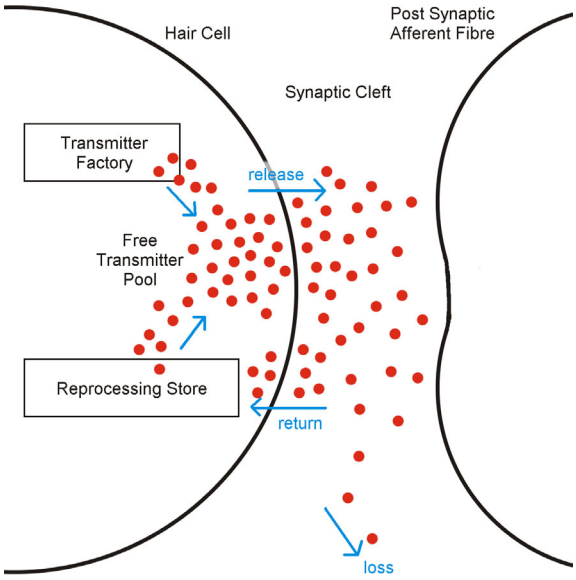


Figure 5.10: Schematic representation of the Meddis Hair Cell model

A single hair cell at each point along the BM could be simulated by processing the output of each Gammachirp filter with the Meddis hair cell model. Unfortunately, the matter is complicated by two issues. Firstly, it is hypothesised that the range of motion of the BM may be compressed by the action of the outer hair cells. As discussed in Chapter 3, the exact compression mechanism is currently unknown. However, it is certain that a single hair cell (and hence the Meddis Hair cell

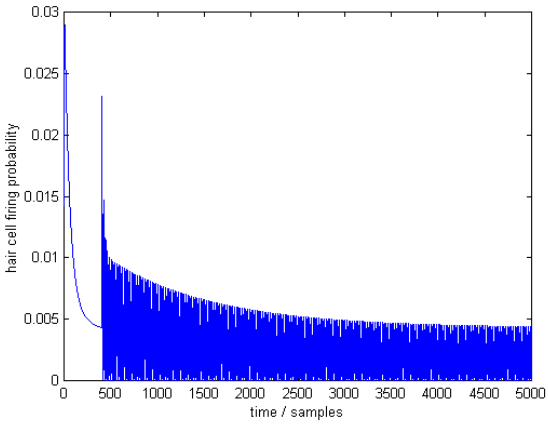


Figure 5.11: Response of Meddis Hair Cell Model

[4] This apparently complex system is *much* simpler than previous models. For example, [Smith and Brachman, 1982] split each hair cell into 512 separate sites, each with its own transmitter factory. However, the simpler Meddis model accurately simulates a larger number of measured phenomena than previous models.

model) will be driven into saturation by the dynamic range of everyday sounds, unless the movement of the BM is compressed.

Secondly, there are at least three categories of inner hair 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 less complex model of hair cell action is sought. Ideally, the model should include the compression of the BM, and yield the same hair cell firing probabilities for a given input as the Meddis model, but without the complex internal processing.

5.3.5.2 Adaptation

[Dau *et al*, 1996a] suggests a model which yields similar results to that of [Meddis, 1986] via a specially designed filter mechanism, taken from [Püschel, 1988]. The output of each gammachirp filter is half wave rectified, then low pass filtered at 1 kHz (see Figure 5.12). 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. For a given stimulus, this circuit simulates the response of the inner hair cells remarkably well. The output is the probability of firing, rather than the neural firing signal itself. However, the latter can be generated by driving a random process from the output of the circuit, if so desired. In the present model, the firing probability is used directly.

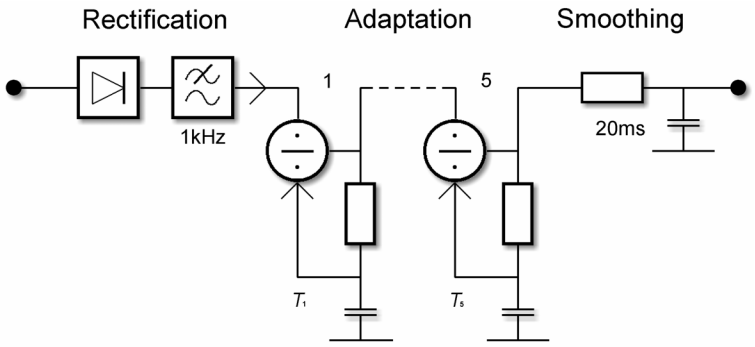


Figure 5.12: Circuit to simulate Hair Cell response

Whereas an individual inner hair cell can exhibit absolute saturation, this is not simulated by the adaptation model. An individual hair cell is said to have reached saturation if an increase in stimulus loudness fails to cause an increase in the firing rate of the cell. However, within

normal listening conditions, the movement of the BM is restricted by the action of the outer hair cells such that the inner hair cells are not driven into saturation. It is only in dead cochlea, where the active feedback mechanism is absent, that the saturation of the inner hair cells is relevant[5]. Hence, by simulating the inner hair cell response *without* saturation, this model is simulating normal hearing, and the compression of the BM is implicitly taken into account.

It has been shown that the overall measured adaptation of the human auditory system is well matched by the adaptation model shown in Figure 5.12 (from [Dau *et al*, 1996a]). The time constants used are 5, 50, 129, 253, and 500 ms.

5.3.5.3 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 after the rectification. 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 [Robinson and Dadson, 1956], processed through the stages of the model.

[Dau *et al*, 1996a] state that a fixed internal threshold does not accurately model the HAS response to signals near the absolute threshold of hearing. However, it will be shown that the novel detection mechanism used in the current model yields accurate near-threshold behaviour. Two types of noise are simulated by fixed thresholds within the model; the internal noise discussed here, and the perceptual noise discussed in Section 5.4.2. For this reason, the internal absolute threshold discussed here is 3 dB lower than the figures of [Robinson and Dadson, 1956] would indicate, since the perceptual noise will raise the threshold by 3 dB.

The advantage of using a fixed threshold instead of internal noise is that the model will always yield the same result with a given input stimulus. If random noise was introduced, many repetitions of the same measurement would be required before a human performance estimate

[5] The inner hair cells may also be driven into absolute saturation in the presence of very loud stimuli, or in elderly subjects where the active feedback mechanism has begun to break down. In the former, hearing loss is the likely result, where as in the latter it has already occurred. This is not accounted for by the present model.

could be calculated. The aim of the model is not to simulate individual responses, which must then be subjected to time consuming repetition and averaging. Rather, the model should simulate averaged human behaviour.

The internal noise does not have a uniform spectrum. Rather, the power of the noise increases at low frequencies (see [Soderquist and Lindsey, 1972]). There are two possible methods of simulating this feature using the current fixed internal threshold approach.

1. The internal threshold values in the lower critical bands may be increased, thus raising the thresholds in line with the internal noise.
2. The incoming audio signal may be filtered, thus simulating the reduced sensitivity to lower frequencies which arises from the internal noise.

The first method takes no account of the upwards spread of masking due to the internal noise, whilst the second method ignores the decreasing effect of internal noise with increasing stimulus level. The second method is implemented within the current model, but the first method can be implemented by setting appropriate fixed internal thresholds, and removing the butterworth high-pass filter described in 5.3.3. The second method was chosen because it can be readily implemented in combination with the middle ear response, simply by creating a filter which is the inverse of the MAF response. In addition, near threshold low frequency artefacts are relatively uncommon in coded audio, so accuracy in this range is less important.

5.3.5.4 Outer Hair Cells

The outer hair cells are believed to respond to signals originating from higher processing centres. Together with the inner hair cells, they are thought to form an active feedback loop, often referred to as the Cochlea Amplifier, as discussed in Chapter 3. This hypothesised mechanism is thought to be responsible for three measurable effects: amplitude dependent spectral masking; compression of dynamic range and in-ear generation of sounds.

The amplitude dependence of spectral masking within the HAS arises directly from the amplitude dependent nature of the excitation pattern upon the basilar membrane. This has been simulated directly by the Amplitude dependent gammachirp filter bank, as discussed in 5.3.4.3. The mechanism employed is reminiscent of the proposed Cochlea Amplifier [Yates, 1995], indicating that the model may match the physiology.

The compression of dynamic range is simulated by the adaptation circuit discussed in Section 5.3.5.2.

The in-ear generation of sounds is not directly accounted for in the present model. This topic is extensively discussed in Appendix A, where audible in-ear generated sounds are shown to be unimportant when assessing coded audio. It is assumed that inaudible in-ear generated sounds, which can be detected using very sensitive measuring equipment, are irrelevant.

5.4 Perceiving a difference

Two performance metrics are important when assessing high quality coded audio. Firstly, given the original audio signal, and the coded version, is there any audible difference between the two? This question can be answered by calculating the probability of perceiving a difference between the two signals. The result gives an indication of the transparency of the coding process. Secondly, if there is an audible difference between the two signals, how annoying is this difference? If a human listener were asked to choose between two coded audio signals, both audibly inferior to the original, which would they prefer? This is a more complex question, which human listeners do not answer consistently. There have been some successful attempts to answer this question (e.g. [Hollier *et al*, 1995]), and these methods can be employed with the present model. This second question is addressed in Chapter 8. The first question is the subject of this section.

So far, we have simulated the passage of sound through the ear canal and cochlea – see Figure 3.1 (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 5.13. The next task is to determine the processing that is necessary to yield an accurate indication of perceived difference.

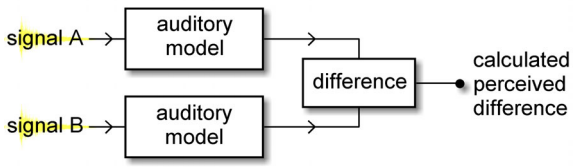


Figure 5.13: General method for calculating the perceived difference between two audio signals

The outputs of the model are n time varying signals, where n is the number of gammachirp filters, as discussed in Section 5.3.4.1. Henceforth, n will be referred to as the number of

bands. Calculating the perceived difference will involve comparing corresponding signals in each band, as shown in Figure 5.13.

This process will be calibrated using known human performance data, in the form of masked thresholds. This same criterion was used to test the performance of the Johnston model in Chapter 4. Masked thresholds are relevant to the assessment of coded audio. If the model can correctly predict the level at which one sound becomes audible in the presence of another, then it should also correctly predict the audibility of coding noise in the presence of the original signal. The similarity between determining the just noticeable difference, and determining the masked threshold is discussed in [Zwicker and Fastl, 1990], where the two situations are shown to yield similar internal differences. The possible problems of using masked threshold data in this manner are discussed in Appendix B.

Full details of the psychoacoustic experiments referred to throughout this section can be found in Appendix C. In a previous version of this model ([Robinson and Hawksford, 1999] and Appendix E), a smaller set of psychoacoustic experiments were used to calibrate the model. The absence of temporal pre-masking data resulted in a model that exhibited no pre-masking. This caused the model to be more sensitive to the onset of sounds than human listeners. For this reason, a new difference detection circuit is outlined below, which has been calibrated using all the psychoacoustic experiments listed in Appendix C.

5.4.1 Possible strategies

5.4.1.1 Calculating the perceived difference by simple subtraction

If the difference calculation in Figure 5.13 is a simple subtraction, then inaudible signal differences lead to large calculated perceived differences (CPDs). Further, if several “just audible” signal differences are processed in this manner, the peak calculated perceived difference for each signal-pair is inconsistent. For example, the at-threshold temporal post-masking 0 ms signal-pair gives a peak CPD of 0.2915, whereas the at-threshold tone masking noise signal-pair gives a peak CPD of 0.0288.

It is apparent that the peak difference between the internal representations is a poor predictor of perceived difference. A possible intuitive solution to this problem is to integrate the difference in some manner, but no time constant can reconcile the differences between the different psychoacoustic experiments.

Comparing the two quoted CPD values, the reader may doubt that any single detection criterion can be appropriate in all circumstances. This may be partially true, but a single detection criterion has been successfully used in several models, and one such approach will now be considered.

5.4.1.2 Calculating the perceived difference using signal detection theory

The model in [Dau *et al*, 1996a] correctly predicts the temporal pre- and post-masking thresholds, and the noise masking tone threshold, in addition to other psychoacoustic phenomena.

This model, in common with many others, assumes that some internal "perceptual noise" must be present, which is responsible for masking differences that would otherwise be audible. To detect the difference signal within this noise, an optimum detector is used. Signal detection theory states that there is no *single* optimum detector; rather, the choice of optimum detector depends upon the task [Green and Swets, 1966]. For example, if the target is a 1 kHz tone, and the masker is gaussian noise, then the optimum detector would average the signal every 1 ms, such that the target signal would reinforce itself, if present. A suitable decision criterion, based upon the output of the averaging process, would decide if the target was present, or absent. The optimum criterion would depend on the amplitudes of the signal and the noise. The mechanism that yields the greatest percentage of correct decisions is defined as the optimum detector.

In order to specify the optimum detector for a given task, the nature of the task must be known. The more complete the a-priori knowledge of the target and masker, the more successful the optimum detector will be. This is because the ability to detect a signal not only depends on the nature of the masking noise, but also on how completely the signal is defined within the detector. This is a fundamental limit of the theory of optimal detection.

Within Dau's model, optimal detection works well. The noise is added by the model itself, so the statistics of the noise are known exactly. The model is given access to a supra-threshold target + masker signal, and to the masker in isolation. Comparison of these two signals yields an internal representation of the target. Where a time-domain representation of the target is known, the optimum detector is convolution. Thus, each stimulus is convolved with the stored target to determine if the target is present within the stimulus. With the correct amount of internal noise, the stimulus ceases to be detected at the measured masking threshold, thus matching human performance. Hence, an optimal detector is used to simulate a sub-optimal detector (the HAS) by the addition of noise.

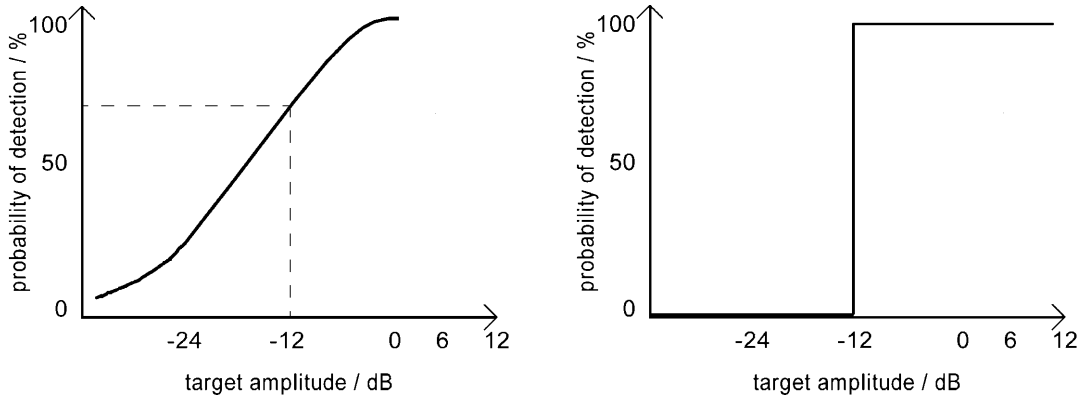


Figure 5.14: Detection thresholds. (a) stochastic (b) deterministic

In this configuration, the model simulates a human listener exactly – that is, repeated threshold determinations are required before calculating the average. However, the internal perceptual noise can be replaced by a fixed internal threshold. This changes the model's performance from stochastic (Figure 5.14(a)) to deterministic (Figure 5.14(b)), whilst maintaining one correct threshold (70.07% in Figure 5.14(a)). The optimal detection process is still applicable, because the target is known.

Unfortunately, this optimal detector is not suitable for audio quality assessment. In this application, the target is an *unknown* audio coding artefact, whereas the optimal detector can only search for a *known* target. Human listeners often undergo training, where they are sensitised to coding artefacts by listening to some highly audible examples. It may be possible to carry out the same process with an optimal detector, thus building an internal library of possible targets (artefacts). However, there is no guarantee that the same artefacts will be present in all subsequently auditioned material. Most importantly, new audio codecs often produce new and unfamiliar coding artefacts. It is not desirable to re-train the model for every new audio codec. Rather, the model should simulate the response of a trained expert human listener.

The difference between the original and coded signals could be *anything*. Where the target is unknown, it is unclear what form the optimal detector should take. For this reason, another approach is used in the current model.

5.4.2 Chosen Strategy

In the current model it is assumed that "perceptual noise" is present within the HAS. This noise limits the accuracy of the detection process, such that some of the information carried upon the auditory nerve is lost in the subsequent neural processing. However, adding noise within the model would create a stochastic detector. Though human responses are stochastic, the model is required to simulate averaged human behaviour. For this reason, the internal noise is replaced by a fixed detection threshold of 0.1 model units. Any differences below this threshold would be lost in internal noise, and hence inaudible.

All the psychoacoustic experiments in Appendix C quote 70.07% detection probabilities. Since this data is used to calibrate the model, the deterministic 0.1 model unit threshold is tied to a stochastic 70.07% probability of detection. The difference detector operates upon the output from the adaptation stage (5.3.5.2). Calibration of the detector is carried out from the data in Appendix C by comparing the masker + target at threshold with the masker in isolation, as processed by the model.

In the present model, no a-priori knowledge of the difference between the two signals is required. Instead, the model switches between two opposite methods that aim to remove the internal noise, in order to reveal any difference between the two signals. Though the current model does not add internal noise, the use of strategies that *would* be appropriate for removing internal noise yields a close agreement with measured data. This indicates that this approach may be used within the human auditory system.

The two noise reduction methods are as follows. The first assumes that there is a degree of temporal uncertainty within the HAS, since the two signals are presented sequentially, rather than simultaneously. For an input signal x, the internal noise of amplitude Δa will be equivalent to a temporal uncertainty of Δt, if $dx/dt \approx \Delta a$. From the temporal masking data in Appendix C, it is calculated that, for $\Delta a = 0.1$ model units, $\Delta t = \pm 5$ ms. This means that any signal difference which can be cancelled out by time-shifting one signal by ± 5 ms is inaudible due to internal noise.

The second noise reduction method assumes that the internal noise has a zero mean. If this is true, then integrating the signal will cause the noise to cancel, whilst any persistent signal

difference will reinforce. This matches the measured phenomena whereby the threshold of detection for a pure tone decreases as the duration of the pure tone increases.

Neither noise reduction method is appropriate in all situations. For example, the first method will not detect a quiet, but persistent signal difference, whereas the second method will not detect amplitude modulation. However, in combination, these two methods will detect the following "just audible" signal differences:

1. noise in the presence of a tone
2. a tone in the presence of noise
3. a short gap within noise
4. a short tone after noise
5. a short tone before noise
6. a 1 dB level increment in a tone

All sounds are either entirely noise-like, entirely tone-like, or some combination of the two. All artefacts consist of the addition of noise, the addition of tones, the subtraction of noise, the subtraction of tones, or the change of amplitude of a signal component. Hence, this list covers most possibilities. All these detection tasks are correctly undertaken by the two detection methods described above. Other (untested) tasks may also be correctly handled by these two detection methods.

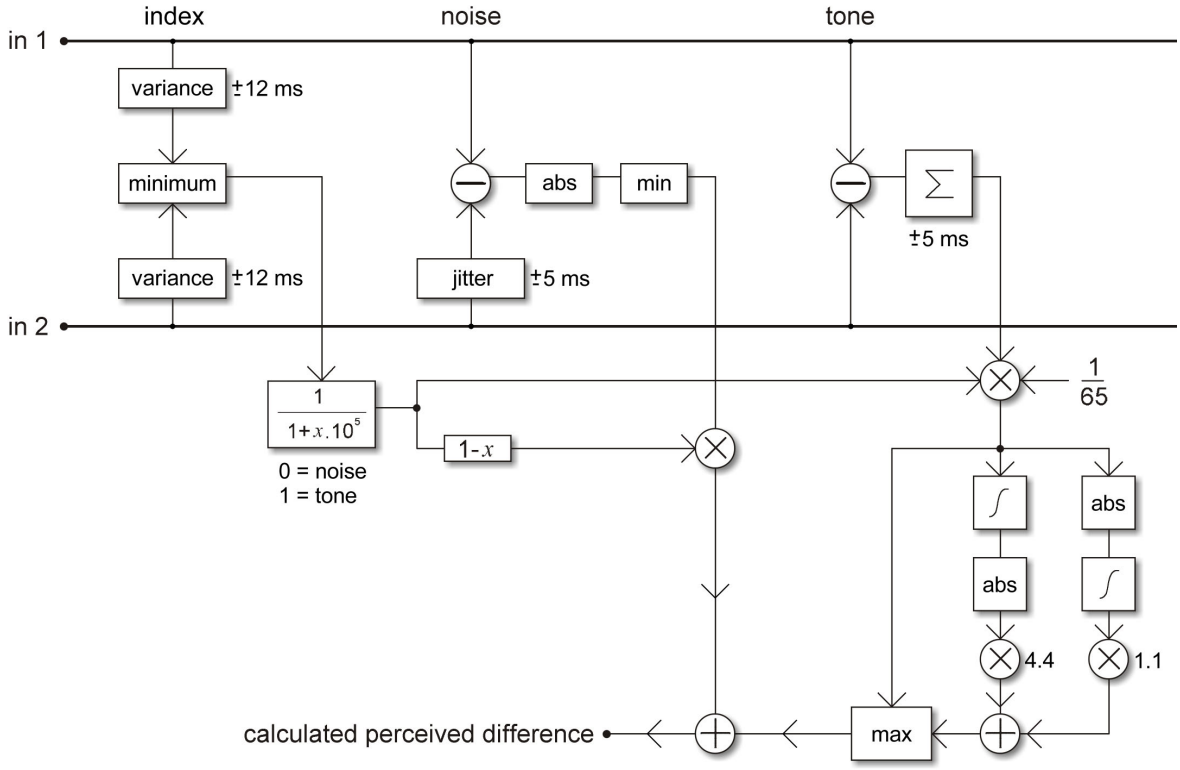


Figure 5.15: Difference detection circuit

The circuits for implementing the two detection methods are shown in Figure 5.15, and are described below.

The first method is appropriate for complex, or time-varying input stimuli, and is denoted "noise" in Figure 5.15. The two signals are compared in the follow manner. For each incoming sample, the current sample of the first input is subtracted from every sample of the second input within a range of ± 5 ms. The minimum absolute difference yields the instantaneous calculated perceived difference value. A CPD greater than 0.1 model units indicates that an audible difference has been detected.

The second method is appropriate for simple or time-invariant input stimuli, and is denoted "tone" in Figure 5.15. The two signals are compared in the following manner. The simple difference is summed over ± 5 ms. The result is integrated with a time constant of 200 ms, and the absolute value is calculated. In a parallel process, the order of these two steps is reversed. The two parallel results are scaled and summed. Finally, the calculated perceived difference is given by this value, or the instantaneous simple difference, if larger.

The model determines which strategy is most appropriate by examining the two input signals, using the circuit denoted "index" in Figure 5.15. The variance of each signal is computed, and the smaller value is used to calculate an index. As the index tends to zero, the "noise" circuit is used, and as the index tends to one, the "tone" circuit is used. Most real-world signals are somewhere between the two. The input signals to the detection circuit are the outputs from the adaptation section. These will be stationary (i.e. the variance will tend to zero) for pure tone inputs above 100 Hz.

In summary, the detection circuit switches between two possible detection strategies, depending on the nature of the input. With a steady-state input, the difference detector integrates the difference over time, thus averaging the internal noise to reveal any persistent signal differences. With a time-varying input, temporal uncertainty within the HAS becomes significant, and only instantaneous differences which cannot be accounted for by temporal uncertainty or internal noise are detected.

5.5 Implementation Speed

The model described in this chapter processes a monophonic audio signal at 0.0001x real time on a P-II 300 MHz PC under Windows 95. This means that the model processes one second of audio in two hours. To compare two audio extracts, each of one second in length, would take over four hours. This rate of processing, whilst typical amongst auditory models[6], is not very useful for the assessment of coded audio.

Most of the computational burden is due to the interpreted nature of the MATLAB environment. If the code was translated into C++ and compiled, significant speed increases would certainly be achieved. Within the MATLAB environment, some of the standard functions are written in C, whilst others are themselves interpreted code. Re-writing the code in order to make maximum use of compiled functions in place of interpreted code offers significant speed benefits, though the computation speed is still prohibitively slow.

[6] In experimental psychoacoustic models, it is usual to process a single critical band. This approach is not suitable for the assessment of coded audio, since the entire audio bandwidth must be assessed.

The following changes were made in order to produce a "fast" version of the model, which would give similar results to the full version, but in a fraction of the time.

1. The code was transferred to a DEC alpha server, running MATLAB 5.1. This offered a 5.5x speed increase over the 300 MHz PC, without necessitating any changes to the code.
2. The amplitude dependent (gammachirp) filter bank is replaced by an amplitude independent (gammatone) filter bank. This filterbank is implemented by the MATLAB "filter" function, which is written in C. The length of each filter within the filterbank is adjusted to be appropriate for the frequency in question, such that the length of the filter *fl* in samples is given by

$$fl = \min(2048, \text{round}((fs/fc) * 20)) \tag{5-9}$$

   where *fs* is the sampling frequency, and *fc* is the centre frequency of the filter.

3. The number of bands is reduced from 96 to 48. A further reduction to 24 bands would be possible, but would cause 80 Hz to be attenuated by 15 dB, due to its position half way between two filters. 48 bands is a good compromise, as shown in Figure 5.5.
4. All constants are pre-calculated once, and stored on disk.
5. The final stage of the adaptation circuit is implemented via MATLAB's "filter" function.
6. The Calculated Perceived Difference is determined once every 25 samples, rather than every sample. The adaptation output is integrated with a time constant of 20 ms, so significant sub-sampling of the signal is possible without losing information. However, to prevent the true peak CPD from being significantly larger than the maximum sampled value, it is unwise to increase the sub-sampling ratio beyond 25:1 for *fs* = 44100 kHz.

These modifications increase the execution speed from 0.0001x real time to 0.02x real time. This means that the fast version of the model processes one second of audio in 45 seconds, as opposed to two hours. This fast version of the model has been validated to ensure that it still matches human performance in many areas.

Amplitude dependent filtering is the most significant feature absent from the fast version of the model. However, the amplitude dependence, number of bands, and CPD sub-sampling can all be specified or overridden by the user via input arguments to the model – see the MATLAB code on the accompanying CD-ROM for further details.

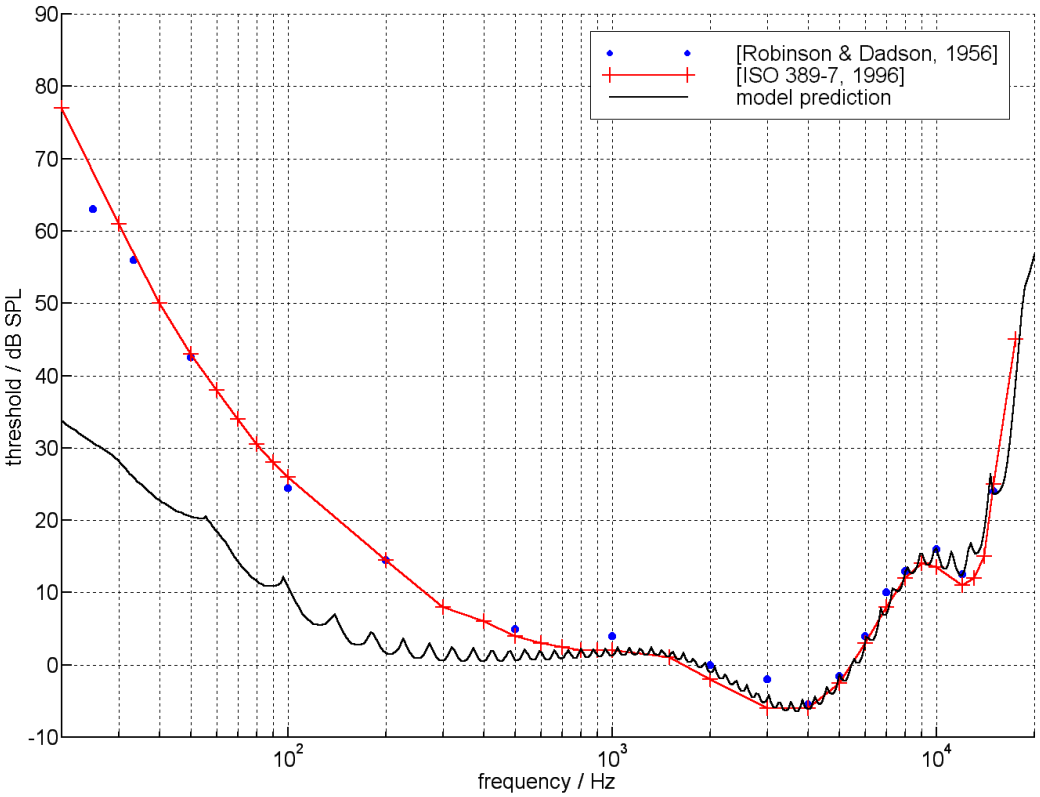


Figure 5.16: Model prediction of the absolute threshold of hearing

The measured responses of human listeners are compared with the response of the model.

5.6 Demonstration

In order to demonstrate the performance of the model, the graphical results from a small selection of psychoacoustic tests are included in this section.

5.6.1 Threshold of audibility

The minimum audible field, or absolute threshold, represents the quietest sound of a given frequency that is audible to a human listener. This has been measured many times. The measurements from [Robinson and Dadson, 1956] and [ISO 389-7, 1996] are compared with the performance of the model in Figure 5.16. The model matches human performance very well above 1 kHz. Below this frequency, the model exhibits a lower threshold than a human listener. This discrepancy is due to the absence of low frequency internal noise within the model, as discussed in Section 5.3.5.3.

The ripple within the model prediction is due to the use of 48 discrete gammachirp filters. Where the test frequency matches the centre of the one the filters, the threshold is at its lowest, and where the test frequency lies half way between two filters, the threshold is at its highest. The ripple can be reduced by increasing the number of gammachirp filters. The choice of 48 filters in the fast version of the model is a compromise between execution speed and accuracy, which the user can easily override.

5.6.2 Spectral Masking

If a narrow band of noise is introduced at 1 kHz, the threshold is increased around this region, as shown in Figure 5.17, which can be found on the next page. This is a demonstration of spectral masking, where any sound below the indicated threshold will be inaudible in the presence of the noise. This graph was produced by comparing the noise in isolation to the noise plus a test tone. The curve indicates the just audible level of the test tone in the presence of the noise.

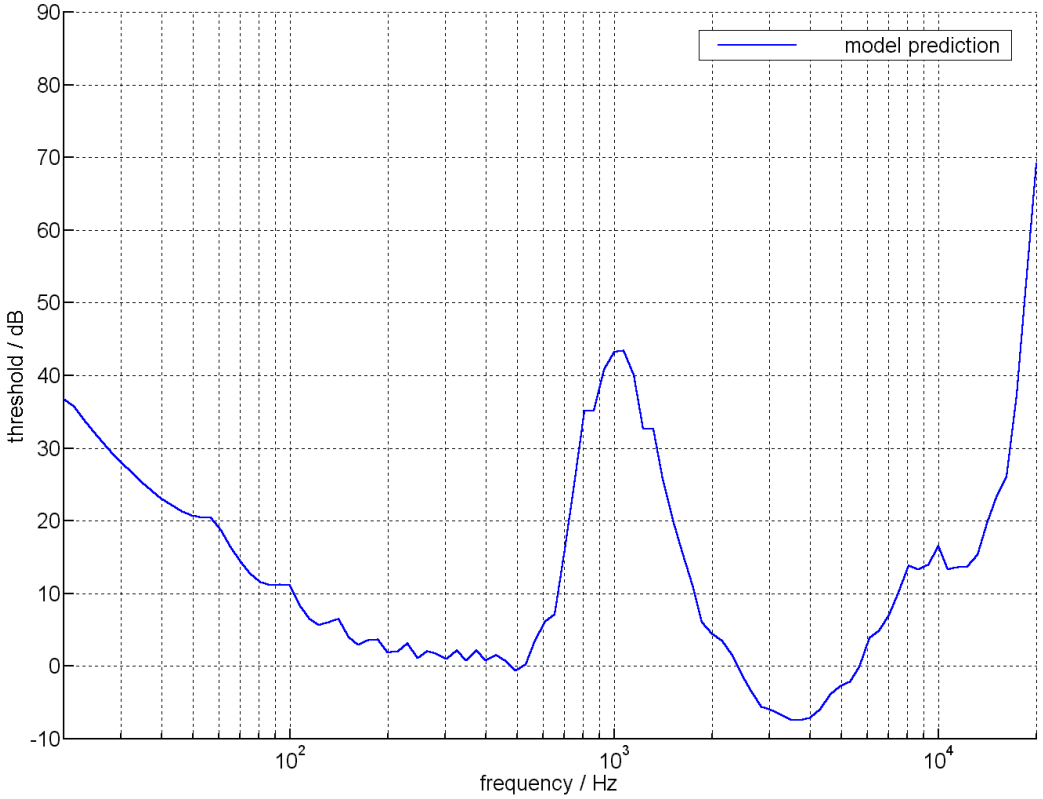


Figure 5.17: Noise masking tone threshold exhibited by the model

To determine this threshold, the model takes the place of a human listener in a simulated psychoacoustic experiment. The model compares the noise in isolation with the noise plus a test tone. If the model judges that the test tone is audible, then the level of the test tone is decreased. If the model judges that the test tone is inaudible, then the level of the test tone is increased. The comparison is then repeated with the test tone at the new level.

The model is deterministic, so there is a constant threshold, below which the signal is always judged to be inaudible, and above which the signal is always judged to be audible. When the level of the test tone converges on this value (to within 0.1 dB), the value is stored, and the frequency of the test tone is increased. Thus, the threshold is determined across the audible frequency range, as shown on this graph.

This simulated psychoacoustic test procedure was developed to demonstrate the model, and to plot Figure 5.16 and Figure 5.17; it is not used in the assessment of coded audio.

5.6.3 Amplitude dependent filter response

A novel feature of this auditory model is the amplitude dependent nature of the filterbank. The response of the filters varies as shown in Figure 5.18.

It can be seen that, as the loudness of the stimulus increases, the attenuation below the centre frequency is reduced.

A full list of the psychoacoustic experiments that the model has correctly simulated can be found in Appendix C.

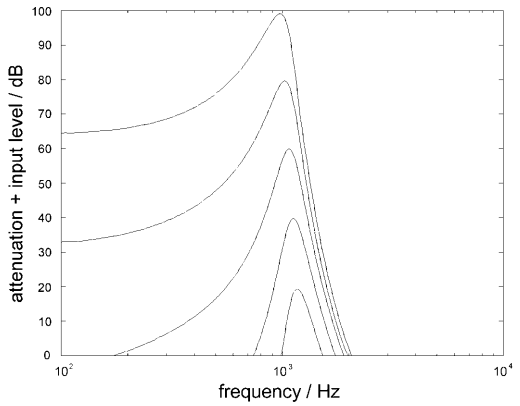


Figure 5.18: Amplitude dependent response of monophonic model

5.7 Conclusion

An auditory perceptual model for the assessment of coded audio signals has been developed. The model simulates the functionality of the physiology found within the human ear. The passage and transformation of sounds from the free field to their neural representation has been accounted for in the model. Finally, the performance of a novel detection network was calibrated using known psychoacoustic data.

6

Spatial Masking

6.1 Overview

The majority of masking experiments concentrate on monaural listening; that is, the test stimulus is applied to one ear only, or an identical stimulus is applied to both ears. Human hearing is binaural, and in real listening environments we rarely experience identical signals at both ears. Where the target signal and masker are separated in space, unmasking of up to 20 dB can occur, compared to the monaural case. In this chapter, we review existing binaural masking data, and describe a psychoacoustic test in which the spatial masking properties of the human auditory system are investigated.[1]

6.2 Introduction

A large amount of quantitative data is available which describes the masking of signals within the human auditory system. The audibility of tones in the presence of broadband noise (e.g. [Bernstein and Raab, 1990]), narrowband noise in the presence of tones (e.g. [Moore *et al*, 1998]), and tones following bursts of noise (e.g. [Zwicker, 1984]) have all been extensively studied. Through such studies, the spectral and temporal masking properties of the human auditory system have been rigorously probed and quantified. However, almost all of these experiments have employed monophonic stimuli in which the target and masker are co-located. Historically, both stimuli were transduced by a single loudspeaker. More recently, headphones

[1] The experiment described herein was carried out jointly with Prof. Woon Seng Gan, Associate Professor of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

have found favour in psychoacoustic experiments, and all stimuli are presented to one ear in isolation, or to both ears simultaneously (diotic presentation). Whatever the method of presentation, the masker and target are perceived as originating from the *same* location, whether or not it is the real location of an external transducer, or a perceived location within the listener's head.

In real listening situations, the signal we wish to hear (e.g. someone talking) rarely originates from the same location as the background noise (e.g. air conditioning). Likewise, in a stereophonic or multi-channel audio system, there is no guarantee that the unwanted noise and distortion will be co-located with the desired signal at any given time. Hence, existing monophonic masking knowledge is insufficient to predict the audibility of coding errors under such circumstances. For this reason, it is necessary to examine the audibility of one signal in the presence of another signal at a separate location.

There have been many experiments to investigate the masking of lateralized sounds (for a review, see [Durlach and Colburn, 1977], section IV, part C). In these experiments, the masker and target are presented via headphones, and are perceived as being located at two separate points within the listener's head. The perceived interaural location of the stimuli is controlled by manipulating the interaural time delay (ITD) and interaural level difference (ILD) for each stimulus. At the extreme, the masker may be applied to one ear, while the target is applied to the other. Alternatively, the target and masker may be applied to both ears at equal loudness, and a delay added to the target signal at one ear only, yielding a phase difference between the two ears. Usually, some combination of ITD and/or ILD is used to lateralise the target and masker at different locations within the listener's head, and the resulting masked threshold is compared to the monophonic case.

To concisely summarise all available data, it can be stated that the target is more audible when it is spatially separated from the masker. Hence, compared to the presentation of the masker and target monaurally (diotic presentation), the amount of masking is reduced. This is referred to as the Binaural Masking Level Difference (BMLD).

Such data indicates that processing non-co-located stimuli using a monaural auditory model may over estimate the amount of masking. To put it another way, a monaural model may predict that something is inaudible, when, due to its perceived location away from the signal, it may be highly audible.

Unfortunately, the masking due to stimuli presented dichotically over headphones tells us little about the masking due to sounds in the free field, i.e. presented via loudspeakers. This is the more usual situation when listening to reproduced audio material, and is the target use for our model. To extend our auditory model to incorporate human binaural hearing, we require data that extends the usual noise masking tone, tone masking noise, and temporal masking experiments to 3-dimensions. With such data, the model can be extended, calibrated, and verified to correctly predict the masking due to real sound sources in real listening rooms.

Data from experiments carried out via loudspeakers is scarce. A summary of the most useful references is given in Table 6.1, which can be found on the next page.

This existing data does not provide enough information on which to base a model of free-field spatial masking. Most importantly, all the experiments study noise masking tone, or noise masking impulses. There is no tone masking noise, or temporal masking data available for spatially separated stimuli. In addition, individual experiments have the following problems.

- [Colburn and Durlach, 1965] and [Carlile and Wardman, 1996] are typical of the many sets of headphone obtained data available, though the latter is interesting in that the stimuli were placed at "virtual locations" via Head Related Transfer filtering. Neither give true spatial masking data.

- [Doll and Hana, 1995] and [Santon 1987] use only 3 different masker locations, and 1 target location. The latter uses a variety of target tone frequencies.

- [Gilkey and Good, 1995] and [Saberi *et al*, 1991] give the most thorough spatial data, testing various target and masker locations. However, the target is a series of band-limited clicks, rather than a pure tone. Such data is useful for validating a working model, but pure tone stimuli are more useful for gaining an understanding of the underlying auditory process at the most basic level.

(this list is continued after Table 6.1.)

Reference	Target			Masker			Notes
	Signal	*Frequency*	*Position*	*Signal*	*Frequency*	*Position*	
Carlile + Wardman 1996	tone	0.6 kHz, 4 kHz	diotic	noise	narrowband broadband	virtual 40°, 90°	**headphone presentation.** HRTF processed masker, sometimes filtered to 1 critical band.
Colburn + Durlach 1965	tone	500 Hz	ILD and ITD varied	noise	broadband	diotic	**headphone presentation.** lateralised stimuli.
Doll + Hana 1995	tone	0.5 kHz, 4 kHz	0°	Notched noise	At least 1 octave around target.	0°, 20°, 40°	**speaker presentation.** Noise Notch width = 0 Hz condition yields free-field masking data
Gilkey + Good 1995	click train	0.387 – 0.938 kHz 1.72 – 4.70 kHz 5.20 – 14.0 kHz	180°, 135°, 90°, 45°, 8°, 0° 100°, 90°, 80°, 45°, 0°	noise	0.238 – 1.33 kHz 1.14 – 7.10 kHz 3.53 – 19.1 kHz	0° 90°	**speaker presentation.** low, medium, and high-frequency data. target and masker location varied.
Kidd *et al* 1998	tonal series	8-part tonal sequence using 1 of 16 frequency ranges within 215 – 6112 Hz	0°, ±30°, ±60°, ±90°	8 Synchronised noise bursts	200 – 6500 Hz	0°, ±30°, ±60°, ±90°	**speaker presentation.** 1 of 6 possible tonal sequences must be identified. target-masker separation reported, rather than actual locations.
Saberi *et al* 1991	click train	0.9 – 9.0 kHz	10° increments horizontally, 30° increments vertically	noise	0.7 - 11 kHz	30° increments horizontally, 0° only	**speaker presentation.** horizontal and vertical plane measurements.
Santon 1987	tone	0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4 kHz	0°	noise	broadband	0°, 45°, 90°	**speaker presentation.** only masker location varied

Table 6.1: Summary of spatial masking experiments from the literature

- [Kidd *et al*, 1998] set an identification task, where the listener is required to identify the masked signal as one of 6 possible tonal sequences. This may or may not yield similar masking thresholds to a pure detection task. A greater problem is that the paper quotes the separation between masker and target, rather than their individual locations. The author believes that averaging the data for target left masker front, target front masker right, and target half-left masker half-right into a single 90° measurement destroys the usefulness of the data for the present work.

Overall, some of these experiments ([Santon 1987], [Doll and Hana, 1995]) provide useful reference points, and others ([Gilkey and Good, 1995], [Saberi *et al*, 1991]) may be used to verify the performance of the completed model. However, the lack of clear, complete information on which to base and calibrate a model means that an investigation into spatial masking must be carried out.

6.3 Spatial Masking Experiment

6.3.1 Aims and parameters of experiment

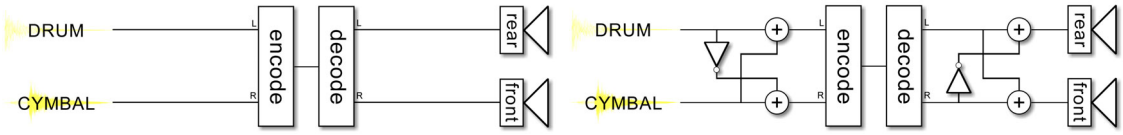
The aim of the current experiment is to investigate the spatial masking properties of the human auditory system. This knowledge will be incorporated into an auditory model, to predict the audibility of one sound in the presence of another, where the two sounds may be located at separate points in space. This auditory model will then be used to predict the perceived sound quality of multi-channel coded audio, replayed in a real 3-dimensional listening environment. The exact details of the spatial masking experiment are set out in the next section. In this section, we define the parameters of the experiment, based on the above stated aim.

The number of spatial locations considered must be large enough to yield adequate psychoacoustic data for the training of an auditory model, but small enough to be practical. Without moving the speakers during the test, the number of stimulus locations is limited by the available number of identical speakers, which is 8. Due to the equipment and time available for the test, all stimuli are located within a horizontal plane encompassing the listener's ears. This allows normal speaker stands to be used in the experimental set-up. Hence, vertical target/masker separations are not considered. As human hearing is assumed (without impairment) to be left/right symmetrical, arranging the speakers in a semi-circular arc to one side of the listener yields the best angular spatial coverage in the chosen horizontal plane. The exact loca-

tions are chosen to match an existing measured set of HRTFs. After the experiment, the at-ear signals caused by each stimulus presentation can be generated by processing the stimuli via the appropriate HRTFs. This capability will be used in the training of the auditory model.

The psychoacoustic tests should extend existing knowledge in the area of masking. The total number of possible tests is constrained by the time allowed for the actual listening (2 weeks in an anechoic chamber), and the duration of each test. The most important tests are those of spectral masking, namely noise masking tone, and tone masking noise. There is time to investigate both phenomena at two different frequencies. For broad-band noise masking a tone, the tone frequencies considered are 1 kHz, where the interaural level difference is the prime localisation cue, and 8 kHz, where the Head Related Transfer Function (HRTF) is the prime localisation cue. For tone masking narrow-band 1kHz centred noise, the tone frequencies considered are 1 kHz and 1.5 kHz, the latter giving an insight into masking of *spectrally* separated stimuli.

Temporal masking is also considered, since the correlation between the masked threshold and the spatial and temporal separation of sounds has never been studied. [Zwicker, 1984] studied temporal masking for masker levels of 80 dB, 60 dB and 40 dB; tone frequencies of 2 kHz and 8 kHz; masker durations of 5 ms, 10 ms, 3 0ms, and 200 ms; and masker/target temporal separations of 0 ms, 2.5 ms, 5 ms, 7.5 ms, 10 ms, 20 ms, 50 ms, and 100 ms. That is, 192 separate masking experiments. To extend this knowledge to 7 different target locations would require 1344 separate masking experiments, each carried out on at least two subjects, at least twice. Each determination of masked threshold takes around 1 minute. Hence, each subject would consent to be tested for 45 hours (plus breaks between testing), and the testers would be present for at least 90 hours. This is probably a severe underestimate of the required time, and is well beyond the 2 weeks available. Hence, we will take a single point on the tests of [Zwicker, 1984] and investigate how the masking due to a particular frequency and temporal delay is dependent on the location of the target tone.



(a) discrete coding of signals (b) matrix coding of signals

Figure 6.1: Coding of surround sound signals

In the discrete case, the front and rear channels are coded via the left and right channels of a stereo codec. In the matrixed case, the front and rear channels are matrixed to the sum and difference of the stereo channels (as in Dolby surround, though the commercial system includes several refinements not implemented here). These stereo channels are coded, decoded, dematrixed and scaled (not shown) to give the front and rear channels.

In addition to these basic psychoacoustic tests, two more experiments are carried out; one to explore the use of audio codecs in multi-channel reproduction systems, and the other to provide human performance data with which to verify the completed model. A major concern when using psychoacoustic based codecs with multi-channel reproduction system is that programme material and coding noise may become spatially separated due to matrixing operations (e.g. Dolby Stereo encoded material) or coding errors. This scenario is discussed extensively in [Gerzon, 1991], and is relevant when transmitting a film with a Dolby surround soundtrack over a digital TV system employing MPEG-1 layer II audio compression. In one experiment, a surround signal is matrixed, coded, decoded, and then de-matrixed, as shown in Figure 6.1. The audibility of coding errors in the de-matrixed signal is investigated, and compared to those present in a signal transmitted discretely via the same codec. In another experiment, the coding noise and programme material are intentionally reproduced at *separate* locations (Figure 6.2), and the relationship between the location and audibility of the coding noise is investigated.

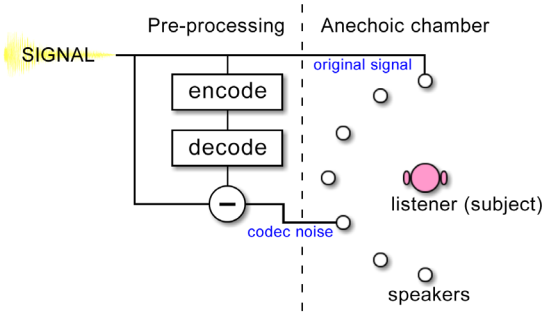


Figure 6.2: Spatially separated codec noise

The original signal is fed to the front loudspeaker. The noise introduced by the codec is isolated by subtracting the original signal from the time-aligned coded version. This codec noise is fed to each of the speakers in turn, and the threshold of audibility in the presence of the original signal is determined. The pre-processing is carried out off-line before the test.

6.3.2 Details of experiment

6.3.2.1 Psychoacoustic phenomena investigated

As discussed in the previous section, the following psychoacoustic tests were performed:

Exp. No.	*Psychoacoustic test*	*Masker*	*Target*
1	**Noise masking tone**	Broadband (white) noise, 35 dB / Hz	1 kHz tone, 80 dB
2	**Noise masking tone**	Broadband (white) noise, 35 dB / Hz	8 kHz tone, 80 dB
3	**Tone masking noise**	1 kHz tone, 80 dB	1 kHz centred 80 Hz wide noise, 80 dB
4	**Tone masking noise**	1.5 kHz tone, 80 dB	1 kHz centred 80 Hz wide noise, 80 dB
5	**Temporal masking**	200 ms white noise, 80 dB	2 kHz tone, 5ms 80 dB (10 ms after noise)
6	**De-matrixed coded audio**	surround sound signal	Codec noise
7	**Spatially separated codec noise**	Music	Codec noise

6.3.2.2 Generation of test stimuli

All stimuli were sampled at CD-quality standard, with a sampling frequency of 44.1 kHz. All synthetic stimuli were generated using an internal floating-point representation of 32-bits, and dithered to 16-bits for output. The start and end of all stimuli were hanning windowed with a rise time of 10 ms, except for the temporal masking experiment, where the masking noise was rectangular windowed, and the target tone was hanning windowed with a rise time of 1 ms. The audio codecs used in experiments six and seven were [Syntrillium, 2000] (mp2) and [FhG, 1997] (mp3). The audio samples for experiment six were taken from [EBU, 1988]. The "Drum" sample consists of 1 bass drum hit from track 29 placed in the rear channel, followed by 1 cymbal hit from track 31-04 placed in the front channel. The "Speech" sample consists of two words from track 50: the word "Christmas" placed in the rear channel, followed by the word "Dinner" placed in the front channel. The audio sample for experiment seven was taken from [Tampera – Feel it], the initial drum hit at 00:01, faded to 70 ms length, and reduced in volume by 10 dB to match the other stimuli.

Historically, wherever noise stimuli were employed in psychacoustic experiments, analogue random noise generators were employed, with appropriate gating and filtering. A consequence of this practice was that each presentation of masker or target noise was different. With the advent of digitally sampled stimuli, the same effect can only be achieved by generating many different samples of noise. In the present experiment, a single noise sample is used throughout each experiment. This changes the signal detection task slightly, as the ear can learn the characteristics of a repeated noise burst in a way that is impossible with a sequence of different noise samples (see Appendix B). However, this disadvantage is offset by two advantages. Firstly, it reduces the complexity of the experiment, and the requirements for storage on the *audiotest* PC. Secondly, it facilitates the subsequent reconstruction of the exact signals heard by the human listener during the test. This will aid the auditory modelling process, but more importantly, will speed up the model testing and verification process significantly. If different noise samples are used, then the statistical spread of the data points is a function of the random nature of the noise, as well as the stochastic process of human hearing and decision. To simulate this, the auditory model must also "listen" to many different samples of noise, and examine the spread to give a probability of detection. However, with a single noise sample, the model only needs to listen to that noise once – it has no need to repeat the process to account for human error, since it can account for the stochastic nature of human hearing internally. Hence, by using a single noise sample, we greatly simplify the validation of the model.

All the targets and maskers used through the spatial masking experiment are included on the accompanying CD-ROM.

6.3.2.3 Experimental set-up

The experiment was carried out at BT Laboratories, Martlesham Heath, Ipswich. The experimental set-up is illustrated in Figure 6.3 - overview, and Figure 6.4 - detail. The listener was seated in an approximately 6m x 5m x 4m anechoic chamber. An anechoic environment is essential for the temporal masking tests, and desirable for all others, so that the stimuli arriving at the listeners ears are uncontaminated by early reflections and reverberation. The walls, ceiling, and floor of the anechoic chamber are lined with foam wedges, as is the inside edge of the motorised door. With the test equipment in place, and the door partially open to allow emergency access, the background noise at the position of the test subject was measured as 20 dB A-weighted, using a B+K SPL meter fitted with an omni-directional microphone capsule.

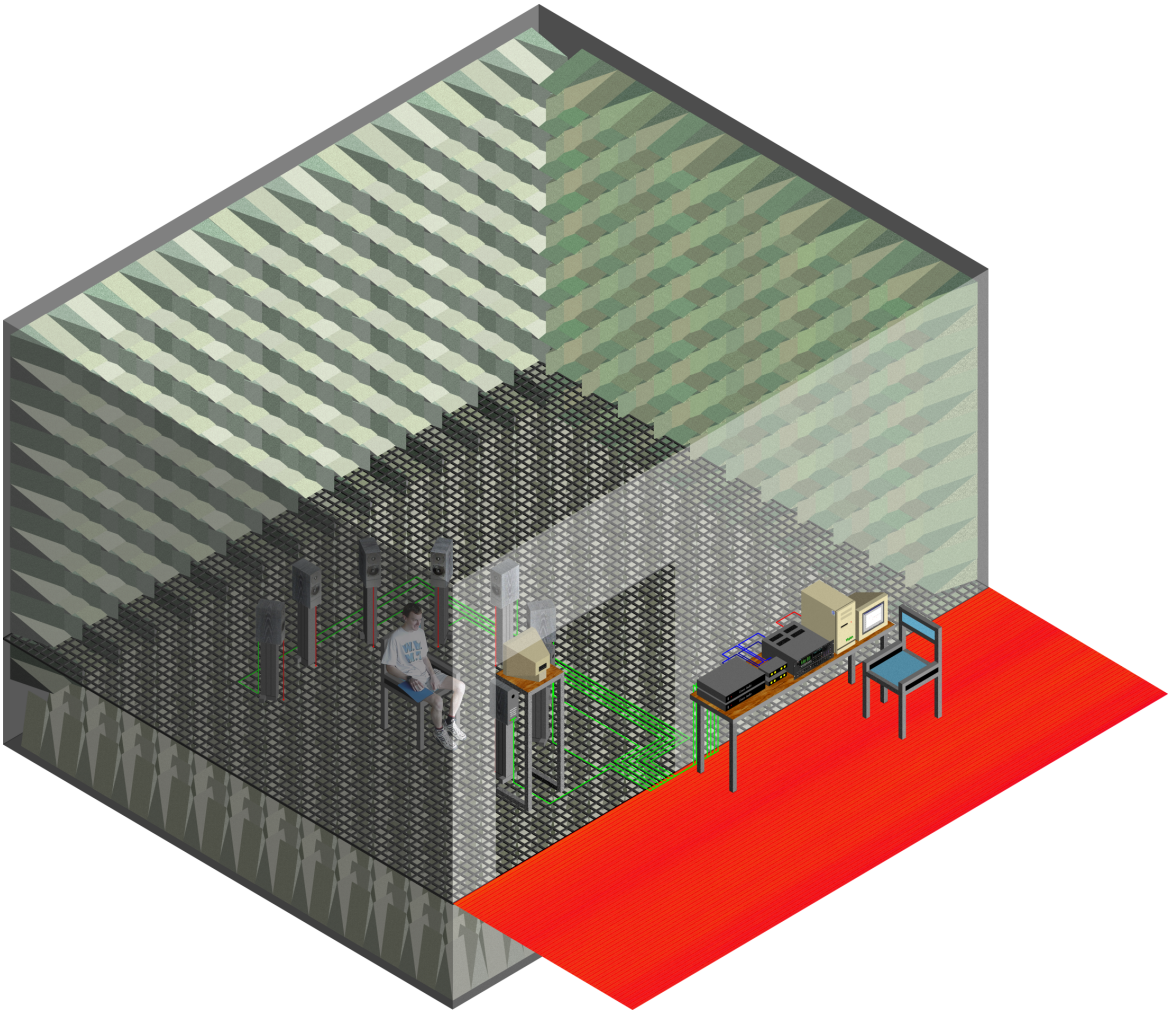


Figure 6.3: Apparatus used throughout the spatial masking experiment - overview

A listener sits within the anechoic chamber, surrounded by loudspeakers. All internal faces of the anechoic chamber are covered by sound absorbing green foam wedges (only the back walls and floor are shown as such in this diagram). A wire mesh acts as a supporting floor within the anechoic chamber. During testing, the square aperture in the front right wall is blocked by a motorised door (not shown). Outside the chamber, the control PC and audio equipment are supported upon a table. Audio connections are shown, as follows: Red = digital, Blue = line-level, Green = speaker.

The experiment was administered via a Pentium II based 300MHz PC system, running the Windows 95 operating system (SR-2.5). Many aspects of the test, including stimulus presentation, the collection and processing of user response data, and the automated training of subjects, were controlled by a custom designed software suite, called "*audiotest*", which is included on the accompanying CD-ROM.. The only items located within the anechoic chamber were a seat for the listener, the eight loudspeakers and stands, a PC monitor display, and a cordless PC mouse. All equipment stands within the chamber were covered with sound absorbing material.

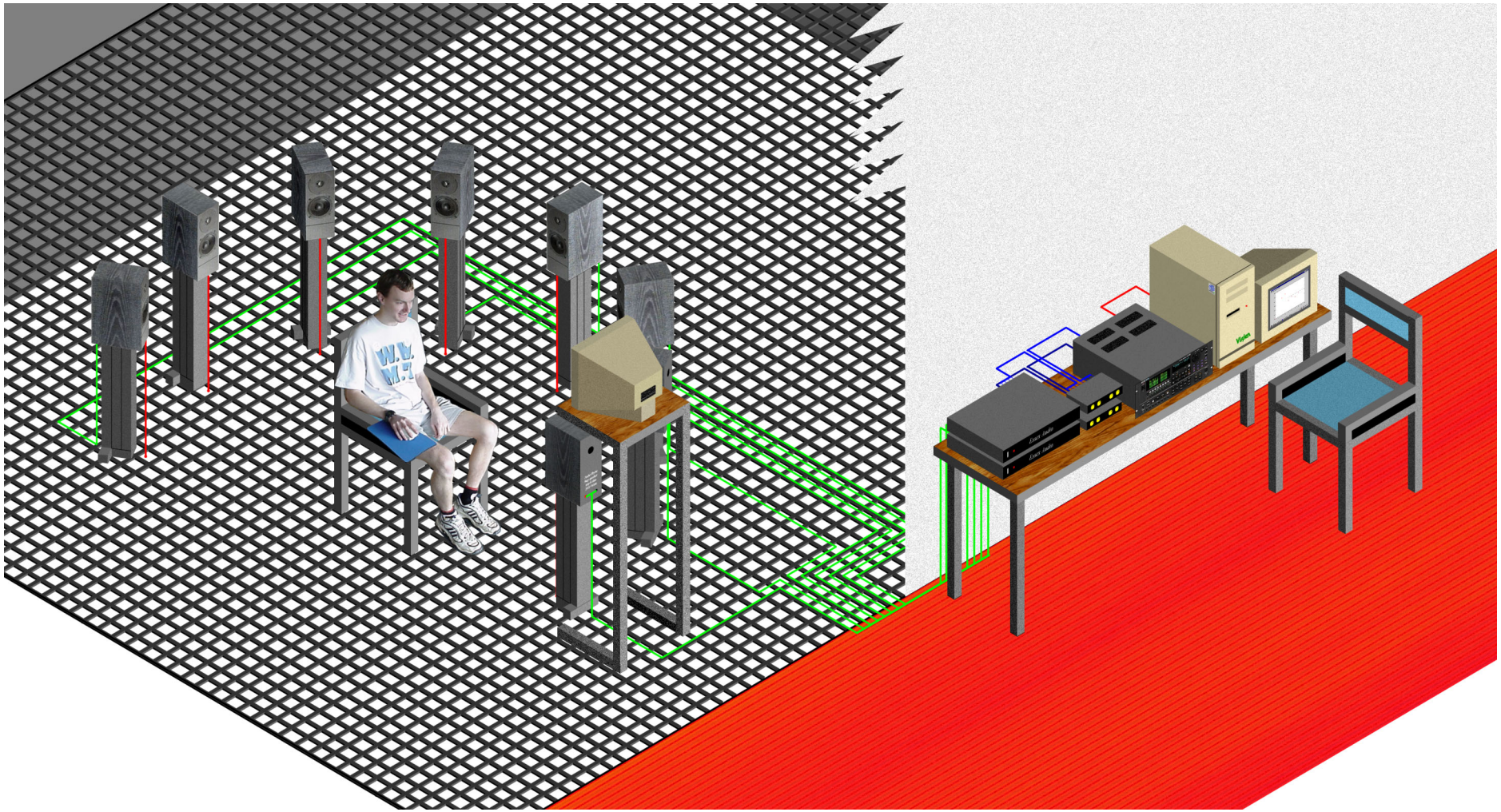


Figure 6.4: Apparatus used throughout the spatial masking experiment – detail

The audio data was sent via a CardD+ digital soundcard to an Akai DR-8 hard disc recorder, which was used for Digital to Analogue conversion and channel switching. The PC soundcard offered two synchronised digital audio output channels (stereo left and right), and these were mapped to any two of eight channels using the DR-8. The eight channels of analogue output from the DR-8 were fed, via passive pre-amps, to a series of LFD PA-1 power amplifiers, and finally to the eight Audio Physic Step loudspeakers within the anechoic chamber. The eighth loudspeaker (behind, and to the right of the listener) was only used in the matrixed audio test – all other tests used the semicircle of seven speakers to the listener's left.

6.3.2.4 Calibration

The playback level was calibrated such that a digital full scale 1 kHz sine wave replayed via one loudspeaker produced a sound pressure level of 90 dB at the position of the listener's head within the anechoic chamber. This calibration was carried out for each channel individually by measuring the SPL with a B+K sound pressure meter, and adjusting the passive stepped attenuator as appropriate. Having calibrated every channel in this manner, eight attenuators were found to be set at identical positions. This confirmed that the eight amplifiers and speakers were matched to better than 1 dB.

6.3.2.5 Subjects

A total of six subjects were used for the listening test. The hearing of each subject was tested using a B+K audiometer, calibrated to ISO R 398 (1964). Four subjects were found to have perfect hearing up to 8 kHz (the upper limit used during these tests), whilst the other two showed some hearing loss above 4 kHz. The two hearing impaired subjects were only used in tests employing 1 kHz targets[2], however their results will be treated with caution.

[2] Each subject's hearing was measured at a convenient time between tests. Unfortunately, one hearing impaired subject had taken part in experiment two before their hearing had been tested. Their data for this experiment was rejected, as noted in Section 6.3.4.

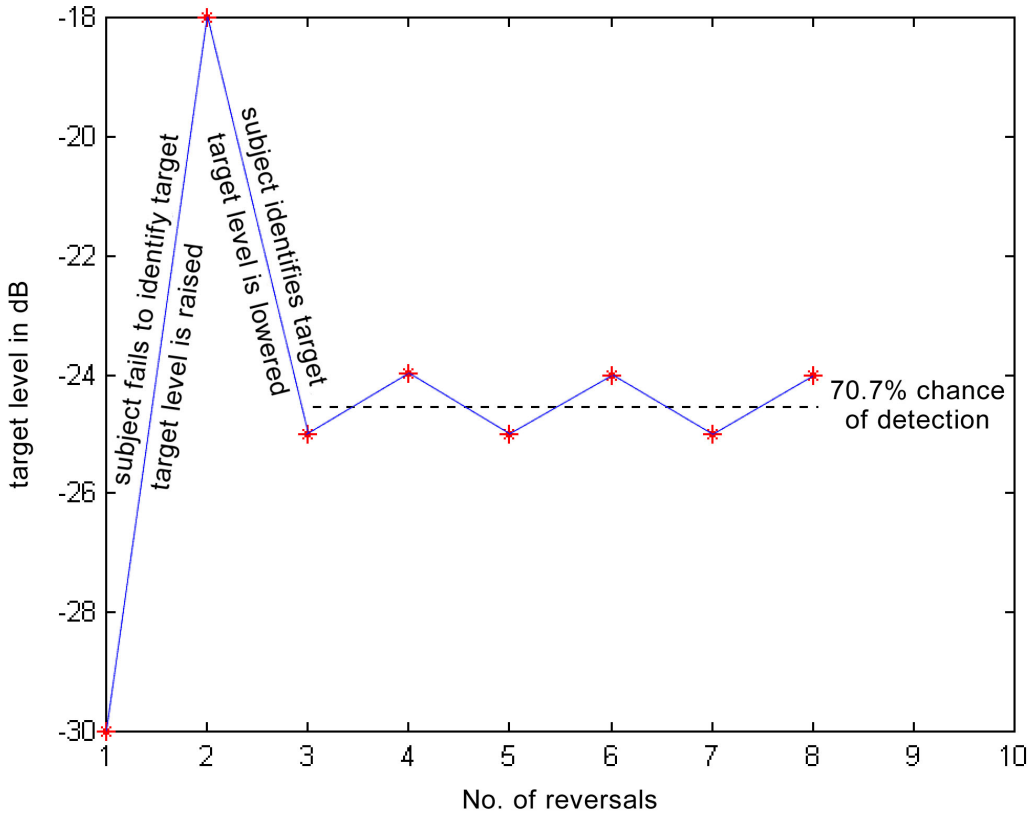


Figure 6.5: Transformed up/down tracking procedure

6.3.2.6 Threshold determination

Each masking threshold was determined by a 3-interval, forced choice task, using a one up two down transformed stair case tracking method. This procedure yields the threshold at which the listener will detect the target 70.7% of the time [Levitt, 1971]. The process is as follows.

For each individual measurement, the subject is played three stimuli, denoted A, B, and C. Two presentations consist of the masker only, whilst the third consists of the masker and target. The order of presentation is randomised, and the subject is required to identify the odd-one-out, thus determining whether A, B, or C contains the target. The subject is *required* to choose one of the three presentations in order to continue with the test, even if this choice is pure guesswork, hence the title "forced choice task." If the subject fails to identify the target signal, the amplitude of the target is raised by 1 dB for the next presentation. If the subject correctly identifies the target signal twice in succession, then the amplitude of the target is reduced by 1 dB for the next presentation. Hence the amplitude of the target should oscillate about the threshold of detection, as shown in Figure 6.5. In practice, mistakes and lucky

guesses by the listener typically cause the amplitude of the target to vary over a greater range than that shown. A reversal (denoted by an asterisk in Figure 6.5) indicates the first incorrect identification following a series of successes (upper asterisks), or the first pair of correct identifications following a series of failures (lower asterisks). The amplitudes at which these reversals occur are averaged to give the final masked threshold. An even number of reversals must be averaged, since an odd number would cause a +ve or –ve bias. Throughout these tests, the final six (out of eight) reversals were averaged to calculate each masked threshold.

The initial amplitude of the target is set such that it should be easily audible. Before the first reversal, whenever the subject correctly identifies the target twice, the amplitude is reduced by 6 dB. After the first reversal, whenever the subject fails to identify the target, the amplitude is increased by 4 dB. After the second reversal, whenever the subject correctly identifies the target twice, the amplitude is reduced by 2 dB. After the third reversal, the amplitude is always changed by 1 dB, and the following six reversals are averaged to calculate each masked threshold. This procedure allows the target amplitude to rapidly approach the masked threshold, and then finely track it. If the target amplitude were changed in 1 dB steps initially, then the dissent to the masked threshold would take considerably longer, and add greatly to listener fatigue. In the case where the listener fails to identify the target initially, then the target amplitude is increased by 6 dB for each failed identification, up to the maximum allowed by the replay system (90 dB peak SPL at the listener's head).

Figure 6.6 gives a top-down overview of the entire spatial masking experiment, from the seven psychoacoustic phenomena under investigation, to the 10000+ stimuli presentations. Due to the reversals and repetitions, each masking threshold is determined 48 times. Figure 6.6 can be found on the next page.

Figure 6.6: Overview of the spatial masking experiment structure

Only the first one or two branches are shown, whilst the arrows to the right indicate further branches.

6.3.2.7 Experimental Procedure

The test procedure for each listener (subject) was as follows. With the subject sat on the chair within the anechoic chamber, the chair was adjusted such that the subject's ears were in-line with the centre of the loudspeaker to their left (speaker 4, at 90°). The subject was given the cordless PC mouse, and instructed to look at the PC monitor at all times during the test. Other than this, the subject's head was not restrained in any way. Before the first test, *audiotest* played the subject the following pre-recorded instructions, replayed via the front loudspeaker:

> "In the following series of listening tests, the task is to select the odd-one-out. Each test features three audio samples, "A", "B", and "C". You should select the one that sounds different from the other two. If you cannot tell any difference between the three samples, please try and guess. If you answer correctly, you will hear this sound [bell]. If you choose incorrectly, you will hear this sound [beep]. Please try the following example."

There followed a simple, interactive example, consisting of two beeps and a beep plus a burst of noise. The subject was not allowed to continue until they had correctly identified the odd-one-out (the presentation containing the burst of noise). If they failed to so, *audiotest* played the following message:

> "That was the wrong answer. You can now see the correct answer highlighted in red. Please try again in this new test, in which the correct answer may have changed."

In practice, no subjects failed this test[3]. When the subject correctly identified the odd-one-out, *audiotest* played the following message:

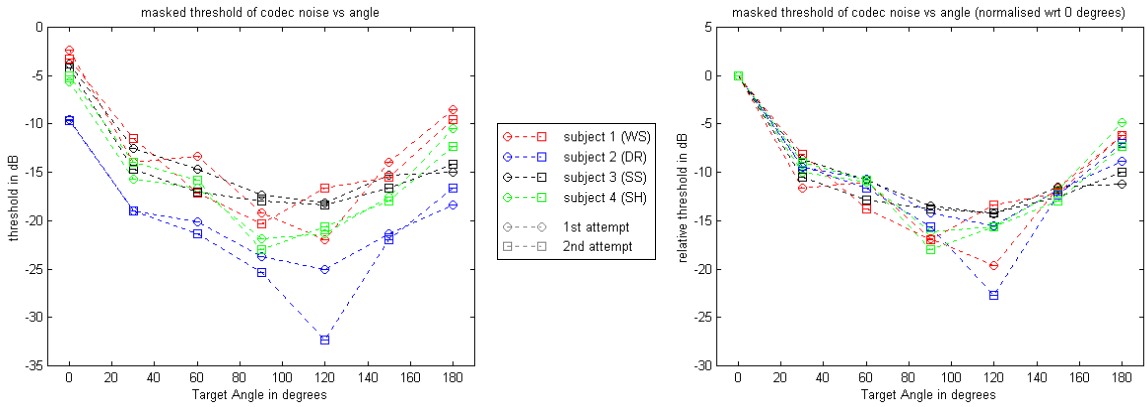
> "Well done, you selected the right answer. You are now ready to start the listening test. Alternatively, you may listen to these instructions again."

[3] During the interactive example, two subjects who became very familiar with the instructions (i.e. bored) intentionally chose the wrong answer. This demonstrated that *audiotest* performed as specified.

If the subject opted to commence the test, the test administrator checked that they had no further questions, then closed the anechoic chamber door.

Unbeknown to the candidates, the first threshold measurement in each experiment (consisting of many stimulus presentations and eight reversals, as illustrated in Figure 6.6) was for training purposes only. After this, a threshold measurement was taken at each target location. Following each threshold measurement, the subject was given 20 seconds rest. Between the fourth and fifth measurements (i.e. at the midpoint of the experiment), the subject was allowed an unlimited amount of time to rest and stretch their legs within the anechoic chamber. After each rest period, *audiotest* informed the subjects verbally of their current position within the test, e.g. "This is test number three out of eight." The candidates were released from the anechoic chamber immediately after the final measurement. Each set of eight measurements (one training set, plus seven locations) lasted 15-20 minutes.

After a break of at least 20 minutes, the subject repeated the experiment. The two sets of threshold measurements were compared, and if any pair of measurements differed by more than 3 dB, the subject was asked to repeat the experiment for a second time, after at least 30 minutes rest. If, after three attempts, no two measurements of the threshold at a particular target location agreed to within 3 dB, then the subject's results were marked for possible rejection.



(a) raw results (b) normalised results

Figure 6.7: Results of experiment seven, showing the effect of normalising each subjects' thresholds to the value at 0°

6.3.3 Results

6.3.3.1 Normalisation of threshold measurements

The raw data from each psychoacoustic experiment consists of two or three threshold determinations for each spatial location from each subject. The raw data from experiment six (codec noise) is shown in Figure 6.7(a). A general trend across subjects is apparent, though the data points from different subjects are widely spaced. One method commonly employed to clarify data from spatial masking experiments is to normalise each subject's results to their threshold at 0°. Thus the variations in threshold with angle are maintained, but the difference in absolute threshold between subjects is removed. Figure 6.7(b) shows the results of this normalisation on the data from experiment six.

6.3.3.2 Graphical results

The numerical results from each experiment are included on the accompanying CD-ROM in both MATLAB 5.1 and Excel 7 format. The results from each experiment are reproduced graphically below. The six reversals which are averaged to give each single threshold measurement are **not** included in the error analysis. Each graph includes error bars of ± 1 standard deviation, calculated from all threshold measurements for a single location by one candidate, or all candidates (as noted on each individual graph). Graph (c) in each Figure includes the data for the most reliable subject in each test, as discussed in Section 6.3.4.2.

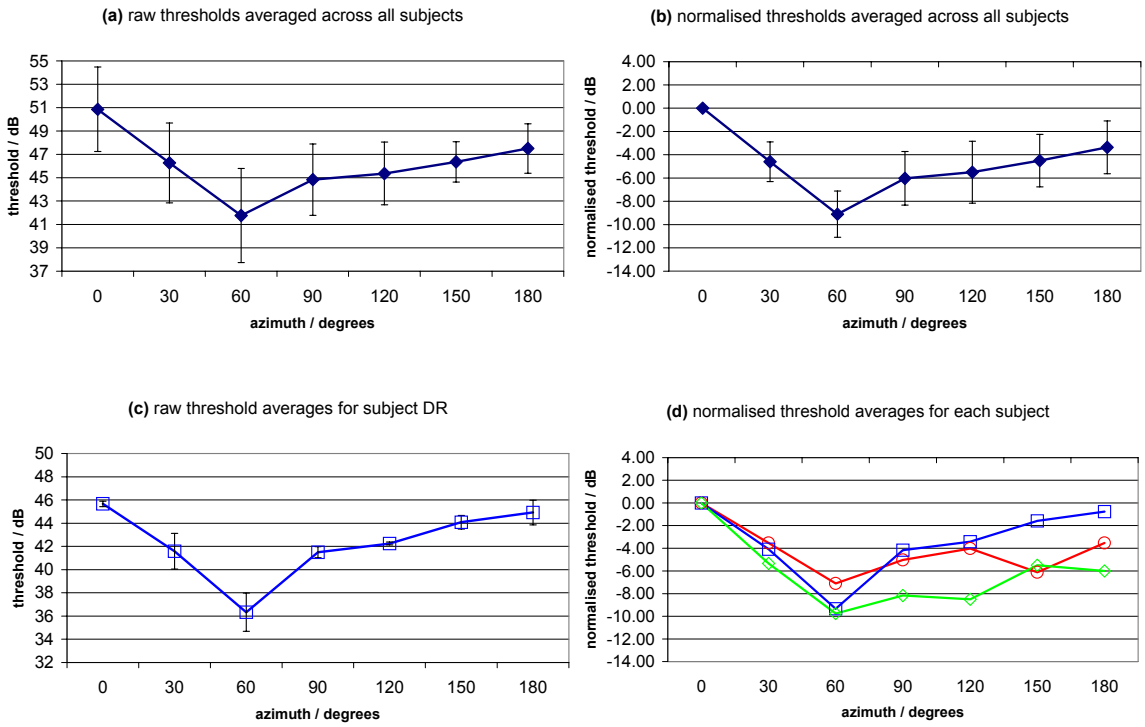


Figure 6.8: Results of experiment 1: Broadband noise masking 1 kHz tone

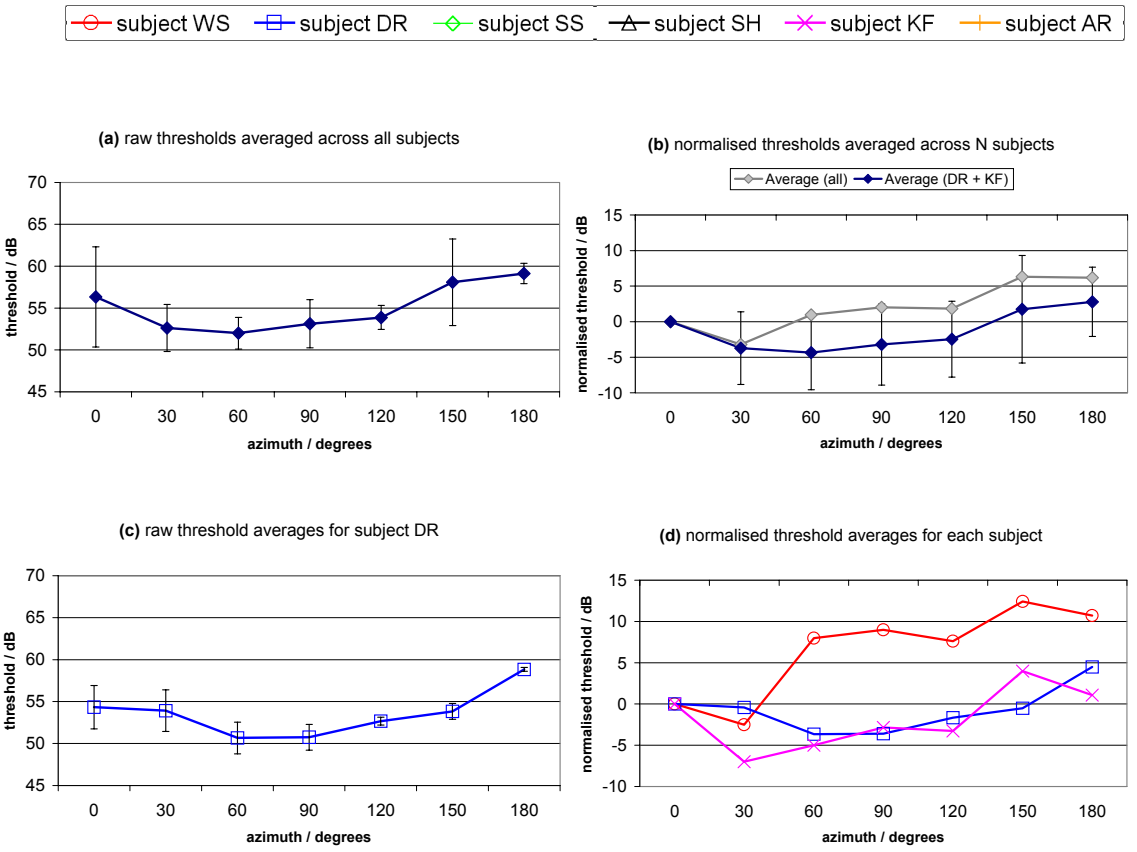


Figure 6.9: Results of experiment 2: Broadband noise masking 8 kHz tone

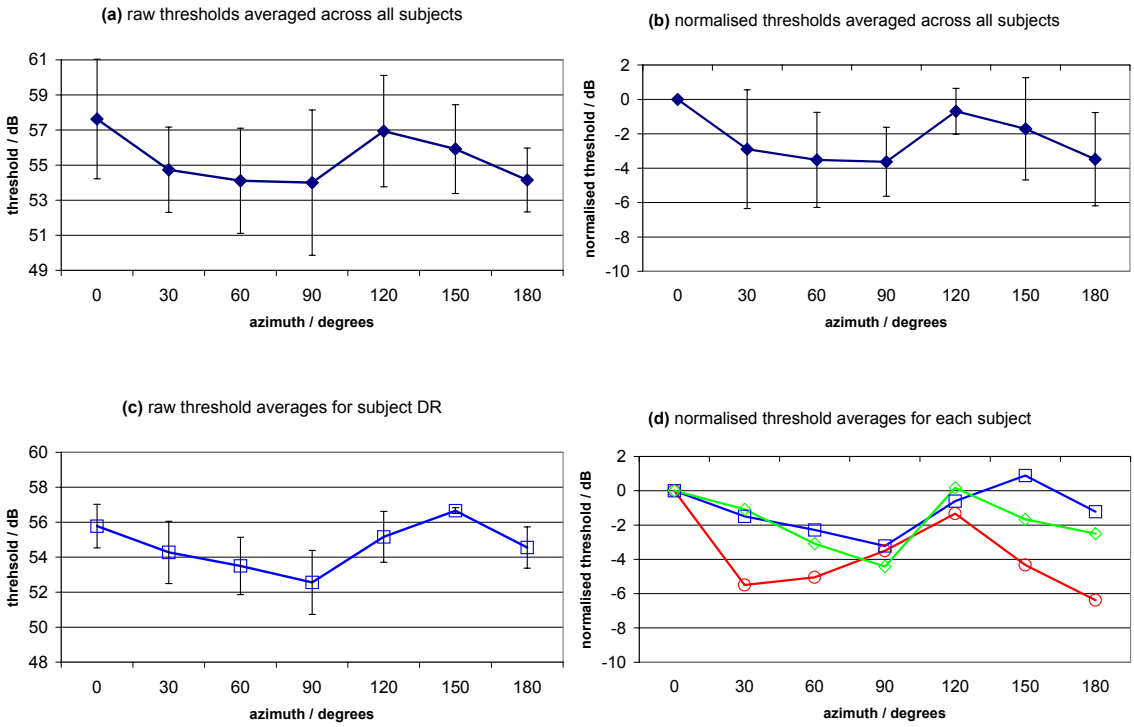


Figure 6.10: Results of experiment 3: 1 kHz tone masking 80Hz wide 1 kHz noise

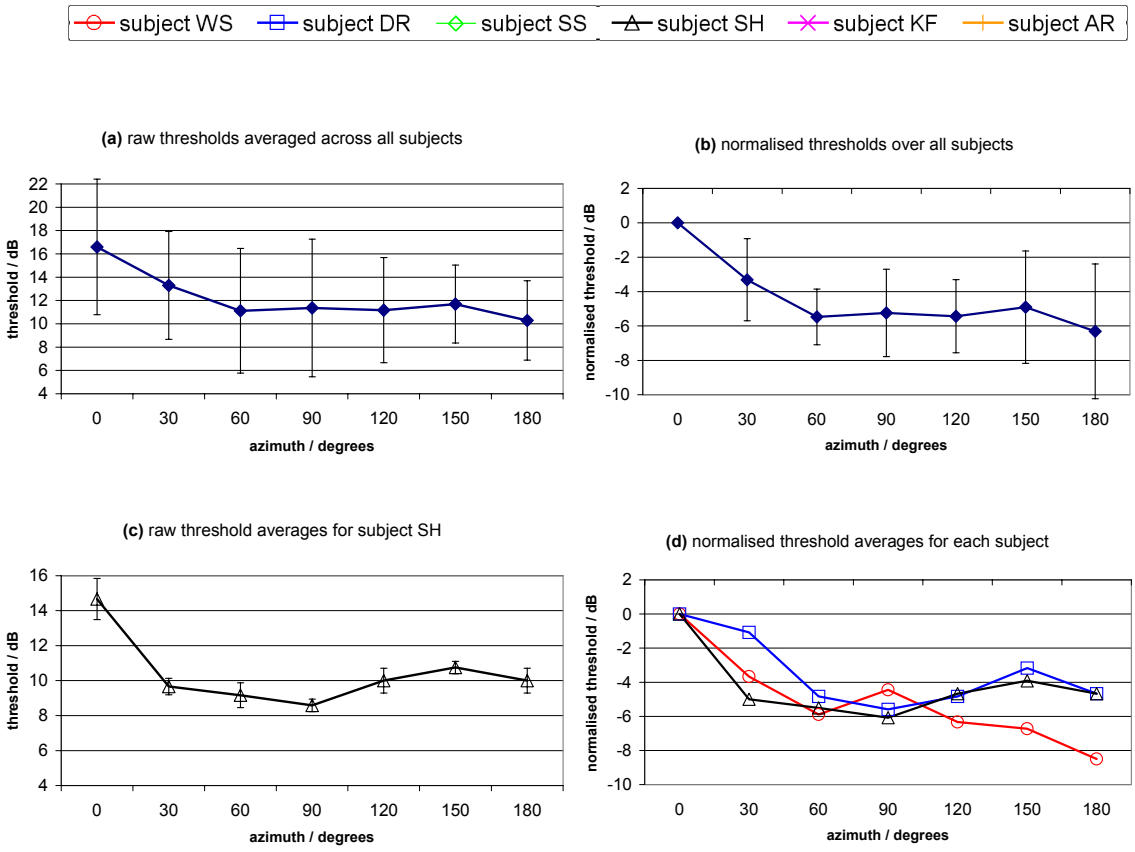


Figure 6.11: Results of experiment 4: 1.5 kHz tone masking 80 Hz wide 1 kHz noise

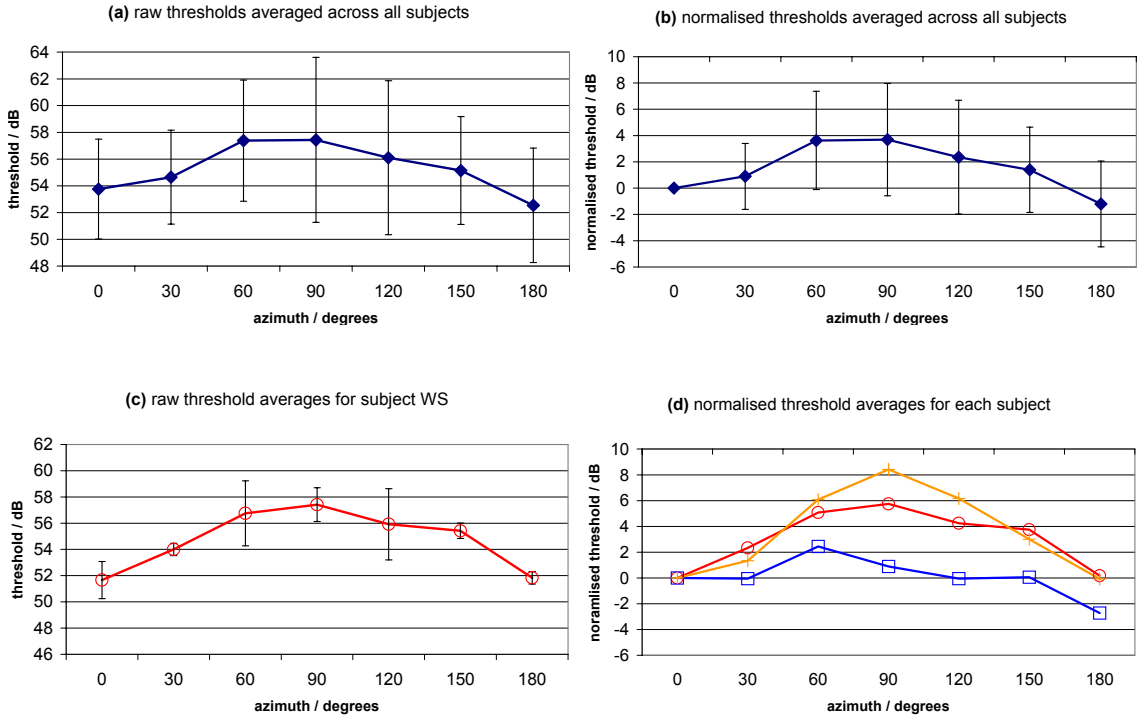


Figure 6.12: Results of experiment 5: Temporal Masking

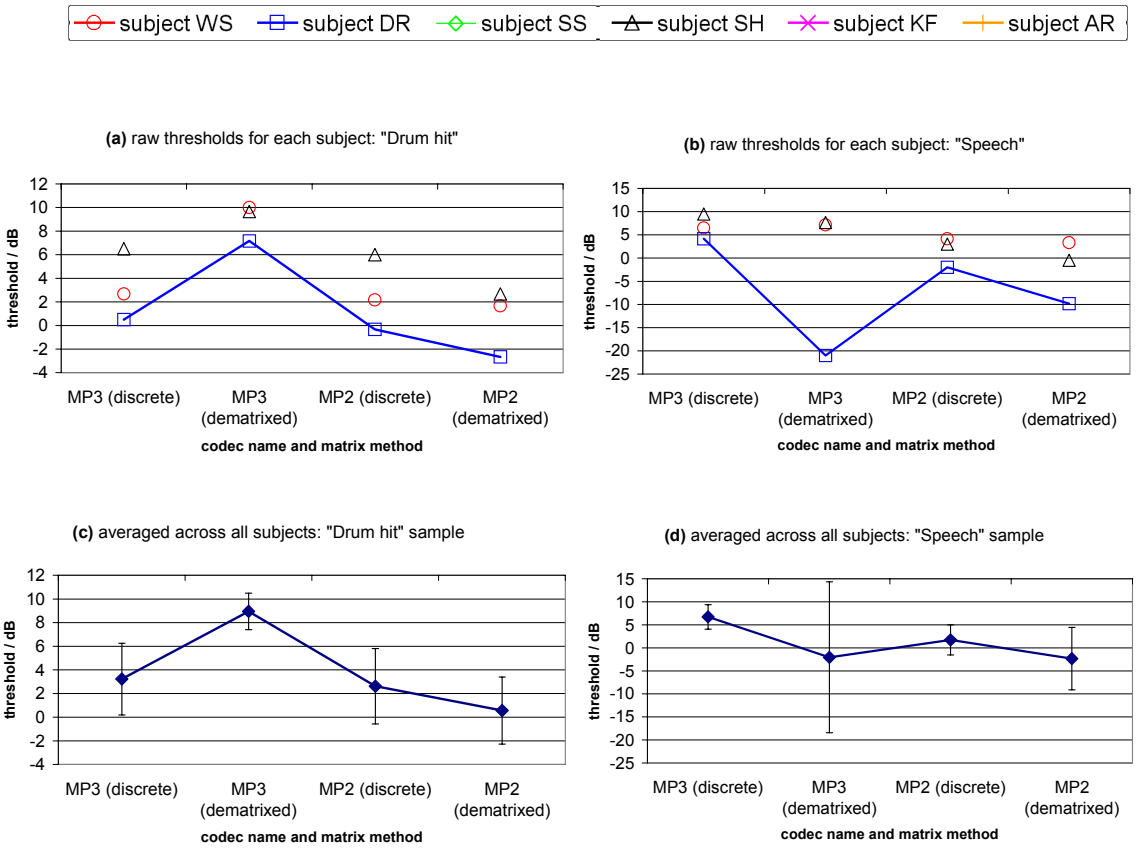


Figure 6.13: Results of experiment 6: De-matrixed coded audio

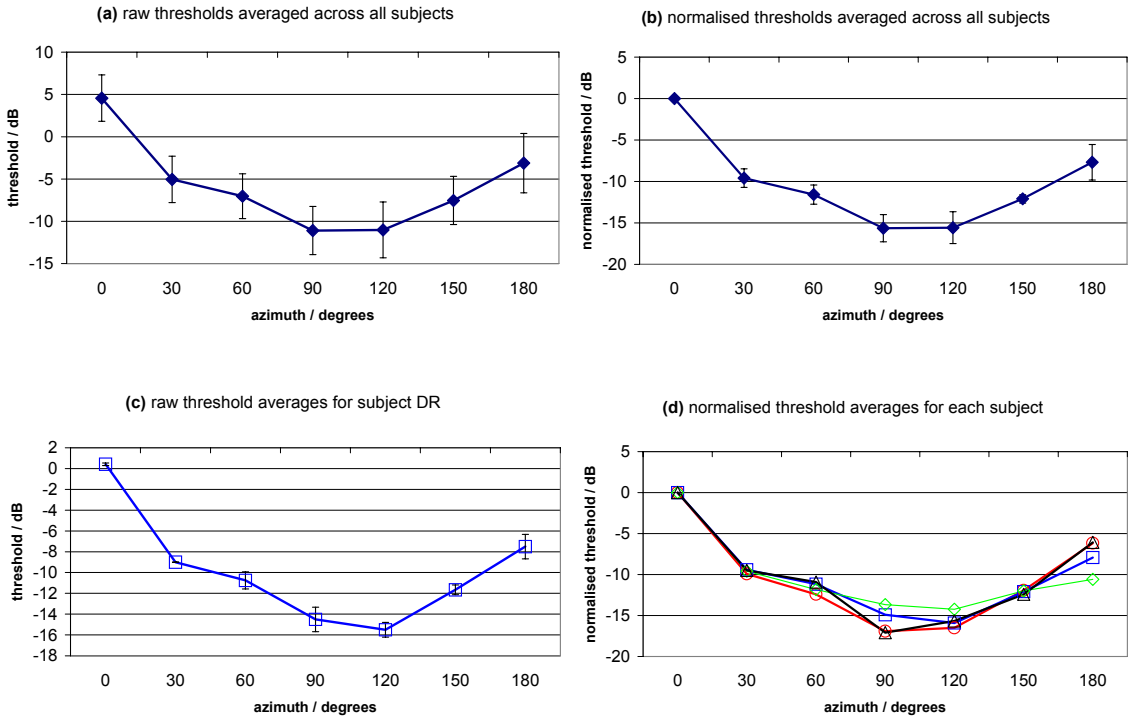


Figure 6.14: Results of experiment 7: Spatially separated codec noise

6.3.4 Discussion

For each experiment, the plot of masked threshold against target azimuth reveals a general trend which is different for each task. In some experiments, the errors are so small that accurate, useable quantitative data may be extracted from the data. In all experiments, the trends are clear enough that qualitative conclusions can be drawn about the variation of masked threshold with target azimuth.

The final experiment yields the greatest variation in threshold with azimuth, possibly because this is the only experiment to employ broadband stimuli as both target and masker. This may explain why previous researchers have concentrated on broadband stimuli; the larger variation in masked threshold with target azimuth gives a clearer picture of the effect of spatial masking. Also, the 3-6 dB statistical variation found within much psychoacoustic data is less significant where the measured effect is of the order of 20 dB. Compare this with the results of experiment five: the inter-subject variation is again of the order of 5 dB, but here the variation in threshold with azimuth is only 6 dB. In this instance, though a trend emerges from the data, the quantita-

tive data is less useful because the variation between subjects is nearly as large as the variation with target azimuth.

In addition to some of the quantitative data being swamped by statistical errors, there are two specific inconsistencies in the data which require investigation, as follows.

Experiment two: subject WS. Referring to Figure 6.9, the results from this subject are significantly different from the others who participated in this test. This experiment tests the subject's ability to hear an 8 kHz tone in the presence of broadband masking noise. The audiogram of subject WS shows a 10 dB impairment at 8 kHz in the left ear. For this reason, the results of subject WS in experiment two will be rejected.

Experiment six: subject DR: "speech" sample: mp3 dematrixed. The masked threshold for subject DR is 27 dB lower than that achieved by the other candidates. Comparison with the other experiments shows that this is drastically greater than can be expected due to inter-subject performance differences. A close examination of this particular sample shows the parameters of the experiment to be unsuitable for this test sample, as described below.

This particular matrixed audio sample ("Speech") was handled very badly by the mp3 encoder. The codec low pass filters the out-of-phase signal component at 1.2 kHz. In the "Speech" sample, this caused most of the out-of-phase component (which is dematrixed to the rear speakers) to be lost, though the in-phase component (which is dematrixed to the front speaker) was reproduced faithfully. In the test, the listener compares the masker only (the original signal) with the masker plus target (the original signal plus the "noise due to the codec"). As the codec removes most of the signal, the "noise due to the codec" is the inverse of the signal, such that, when added to the masker (the original signal), the result is almost silence (see Figure 6.15). This occurs when the target signal is presented at 0 dB relative to the masker (i.e. at the same level as coded), and is clearly audible.

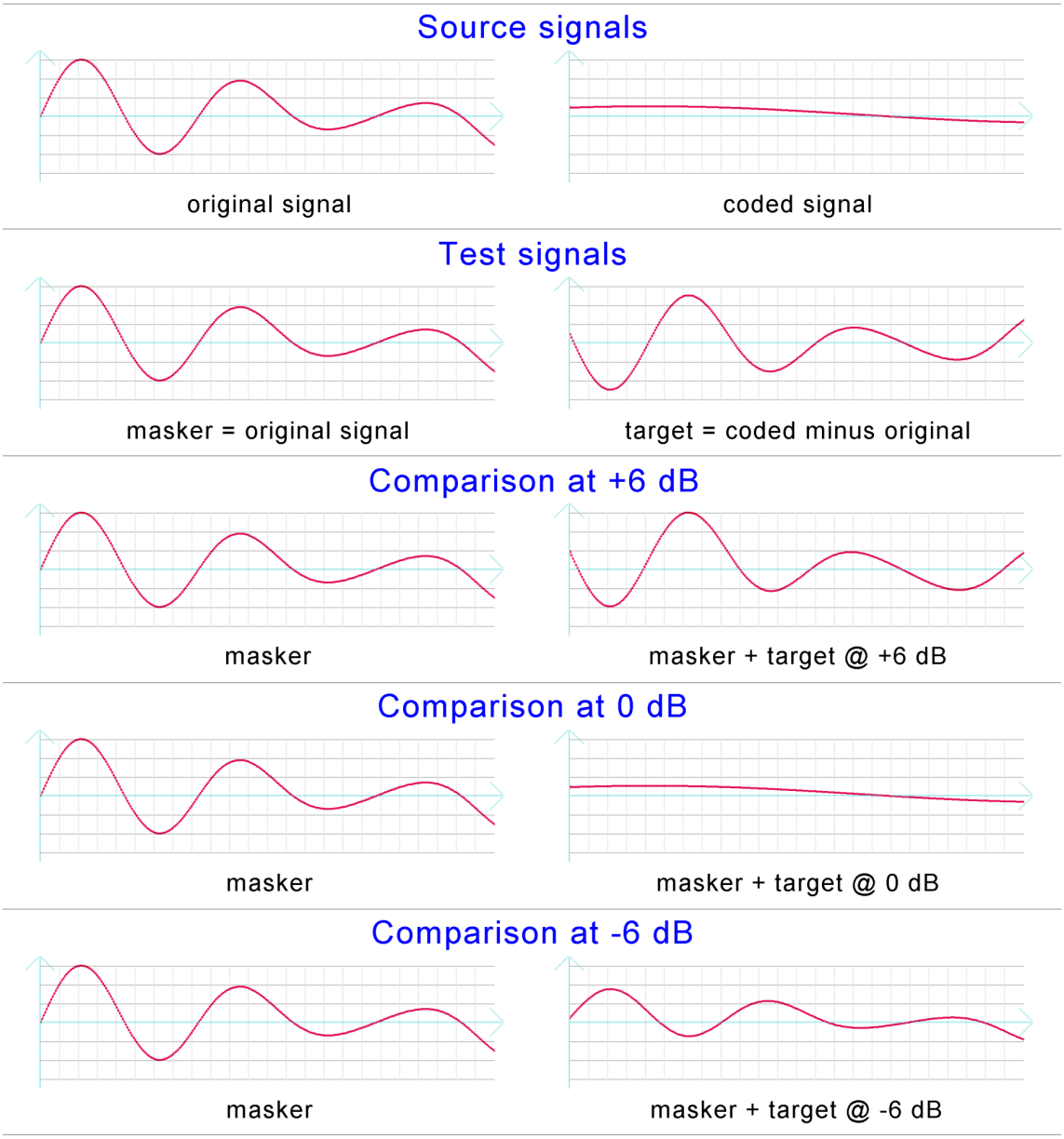


Figure 6.15: Problem signals in experiment six

A poor codec has removed most of the signal, as shown in the first pair of graphs, "Source signals". In the test, the *masker* is the original signal. The *target* is the coding noise, which is isolated by subtracting the original signal from the coded signal. In this instance, the target is approximately the inverse of the masker (as shown in the second pair of graphs, "Test signals"). The listener compares the masker to the masker plus target. With the target at the original level, the masker and target almost cancel (as shown in the fourth pair of graphs, "Comparison at 0 dB"). If the target is increased by 6 dB, then the masker plus target is approximately the inverse of the masker signal (as shown in the third pair of graphs, "Comparison at +6 dB"). If the target is decreased by 6 dB, then the masker plus target signal is a scaled version of the masker itself (as shown in the final pair of graphs, "Comparison at –6 dB").

A problem arises if the target signal is presented at +6 dB relative to the masker. As Figure 6.15 illustrates, this is the same as taking the original signal, and adding twice the inverse: the result is an inverted signal. Few people can detect phase inversion, so this signal is indistinguishable from the original. This yields a local minima in audibility around +6 dB, as shown in Figure 6.16. In experiment six, subjects WS and SH converged on this local minima, whilst subject DR "jumped over" this minima by successive correct answers *before* the first reversal (during which the step size is 6 dB, as discussed in Section 6.3.2.6). Having avoided the false minima, subject DR reached a realistic audible threshold in the usual manner. If subjects WS and SH had not encountered the local minima, it is likely they would have achieved similar thresholds to subject DR.

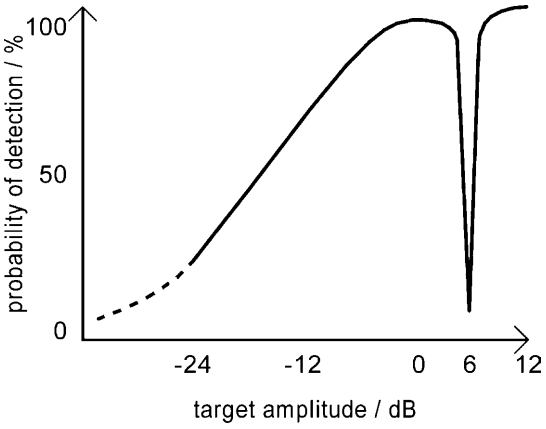


Figure 6.16: False minima in audibility for experiment six at +6 dB

(drawn for illustrative purposes only – not from experimental data).

In conclusion, if the masker and target may be correlated, it is important to test for false minima in audibility *before* carrying out a full psychoacoustic test. If any false minima are found, then the target amplitude range must be bound to exclude them. In experiment six, the target amplitude of sample "Speech" processed via mp3 dematrix should be prevented from exceeding 0 dB.

6.3.4.1 Subject ability

The ability to perform well in psychoacoustic tests is not perfectly correlated to hearing ability as measured by an audiogram. Training improves acuity in any given task, though in this test there was no evidence that thresholds improved with repetition (after the initial training period). This is surprising, since other tests (e.g. [Moore *et al*, 1998]) have indicated that a longer training period may be necessary than was possible in this current series of tests. Adequate training is defined such that further training would not produce an improvement in performance. The conclusion must be either that the current experiment consisted of adequate training, or that so little training was employed that candidates never began to improve. The latter is unlikely, since some performance improvement would be expected during 512 attempts at the

same task (4 guesses x 8 reversals [inc. 2 not used for data] x 8 locations [inc. 1 used for training] x 2 repetitions) if it was going to occur at all!

Though the training period was short, most candidates had the opportunity to overhear another candidate carrying out the same task before them, so this may have helped. Another feature of the experiment which may have compensated for the relatively short training is the requirement that the candidate repeats each measurement to within 3 dB, as discussed in 6.3.2.7. Had the candidate improved significantly (more than 3 dB) between the first and second repetition (possibly due to insufficient training before the first measurement), a further repetition would be required, and the first measurement would be rejected. In practice there is no evidence of this happening: candidates do not improve between the first and second repetition, rather the differences are random in nature. However, the safeguard was there.

Even after subjects are trained adequately in a given task, there remain large differences between the masked thresholds exhibited by different subjects. The "best" subjects can detect masked sounds that are up to 15 dB quieter than those just audible to the "worst" subjects, as shown in Figure 6.17, which can be found on the next page. It is interesting to compare a subject's ability to detect a signal in quiet (as measured by an audiogram) and their ability to detect the same signal in the presence of a masker. The presence of a masker may raise the threshold by tens of dB, e.g. for the 1 kHz tone used in experiment one, the threshold is raised from 5 dB SPL in quiet to 45 dB SPL in the presence of the masker. A listener with poorer hearing may have a threshold in quiet of 20 dB or more. This means that a 45 dB SPL tone in quiet would be clearly audible. However, a 45 dB SPL tone in the presence of a masker is inaudible to our impaired listener – not only has their absolute threshold been shifted upwards by hearing damage, but the masked threshold has also been raised, in this case by 7 dB to 53 dB.

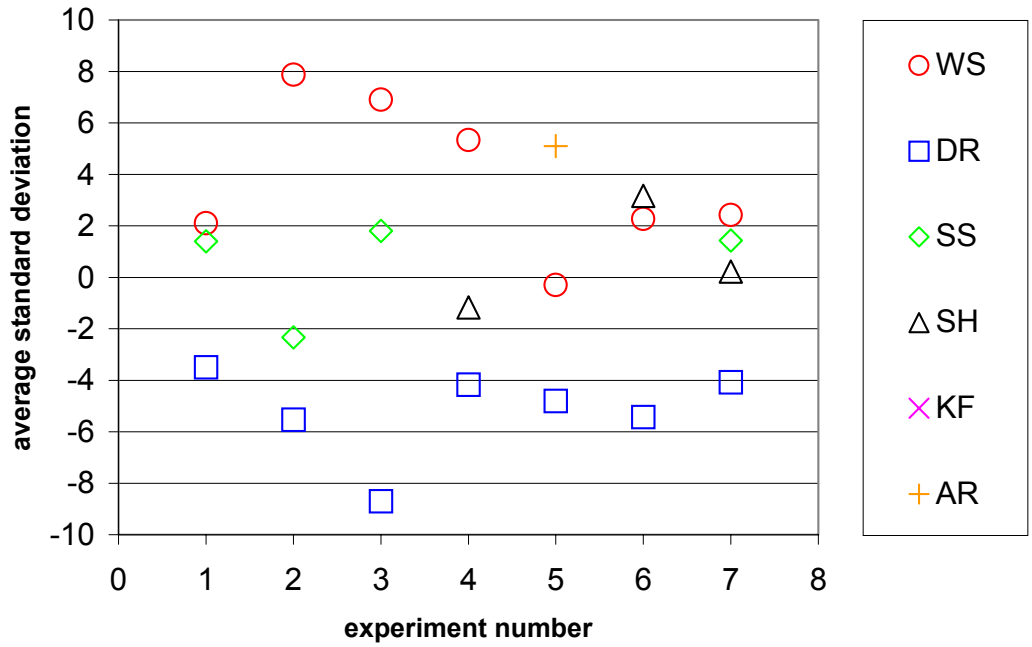
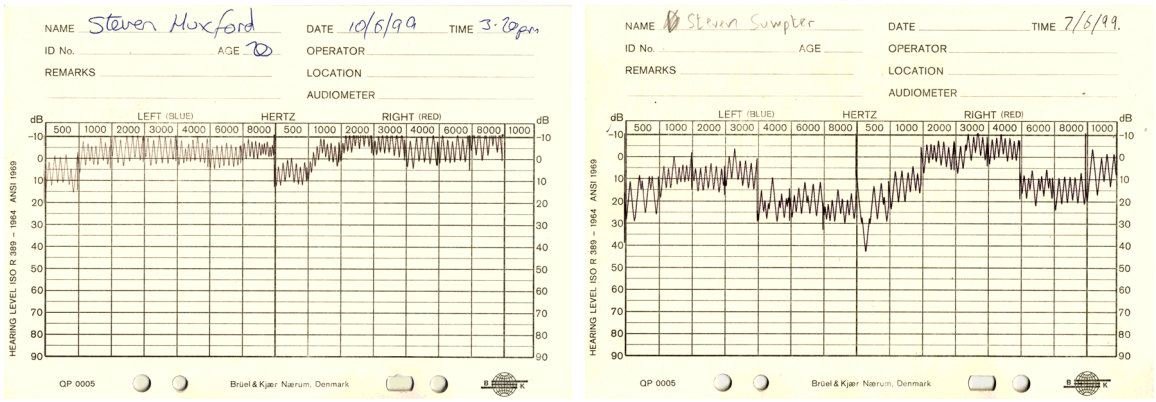


Figure 6.17: Average masked threshold estimate for each candidate, sorted by experiment

Data from each experiment are normalised to the overall averaged threshold for that experiment, thus only inter-subject differences remain.



(a) subject SH – excellent hearing (b) subject SS – impaired hearing

Figure 6.18: Audiograms of two subjects from spatial masking experiment

The audiograms show hearing performance relative to typical human hearing thresholds (normalised to 0 dB). The apparent impairment at 500 Hz is present on all candidates audiograms, and is almost certainly due to the low frequency noise from air conditioning present in the testing room. The anechoic chamber used for the psychoacoustic experiments was quiet, but all other available rooms contained air conditioning, hence there was no way to avoid this problem. As no experiments use stimuli under 1 kHz, having inaccurate 500Hz audiogram data was not thought to be a significant drawback.

It is tempting to correlate threshold in quiet with masked threshold, and to explain all difference between listeners as arising from hearing damage as shown by an audiogram. Reference to each subject's audiogram shows that those with poorer hearing *sometimes* have higher thresholds, as may be expected. In particular, ranking the subjects by threshold for detecting a 1 kHz tone in quiet (from the audiogram) gives the same order as ranking the subjects by thresholds for detecting a 1 kHz tone in the presence of broadband noise (from experiment one). However, for subjects with similar (good) thresholds in quiet, there can still be differences in masked thresholds – e.g. 3-4 dB between DR and SH in experiment four, and 4 dB between DR and SH in experiment seven. Strangely, SS has the worst audiogram overall (Figure 6.18), but his performance matches that of DR (the subject with the lowest thresholds throughout) in experiment three. Finally, subject WS has a fair audiogram, but his thresholds are even higher than his audiogram would suggest. As subject WS was involved in the creation of the experiment, this shows that familiarity with the audio samples and test procedure is not as great an advantage as might be expected.

Figure 6.19 shows the relationship between threshold in quiet and masked threshold for all experiments. The data from each experiment has been normalised to leave only inter-subject differences. After normalisation, the correlation coefficient is +0.4672, which indicates a weak positive correlation.

This shows that poor performance in psychoacoustic tests is not entirely due to measureable hearing impairment. If hearing impaired subjects were the only ones to exhibit a poor correlation between threshold in quiet and masked threshold, then the poor correlation could be attributed to non-linear hearing impairment. However, the masked thresholds of DR and SH are consistently 4 dB apart, even though both subjects exhibit no hearing impairment. This indicates that, beyond the mechanical condition of their cochlea, there are higher auditory processes which vary between subjects. This makes the task of model-

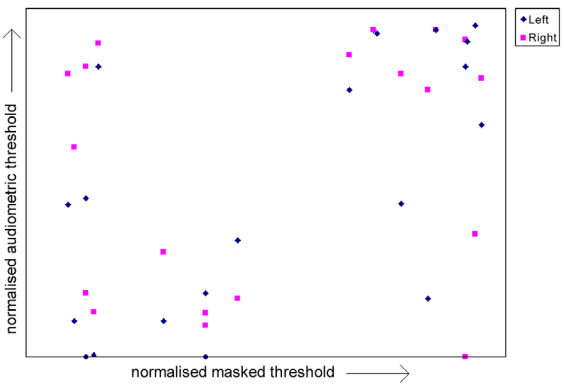


Figure 6.19: Relationship between audiometric threshold in quiet and masked threshold

This graph demonstrates the poor correlation between these two quantities.

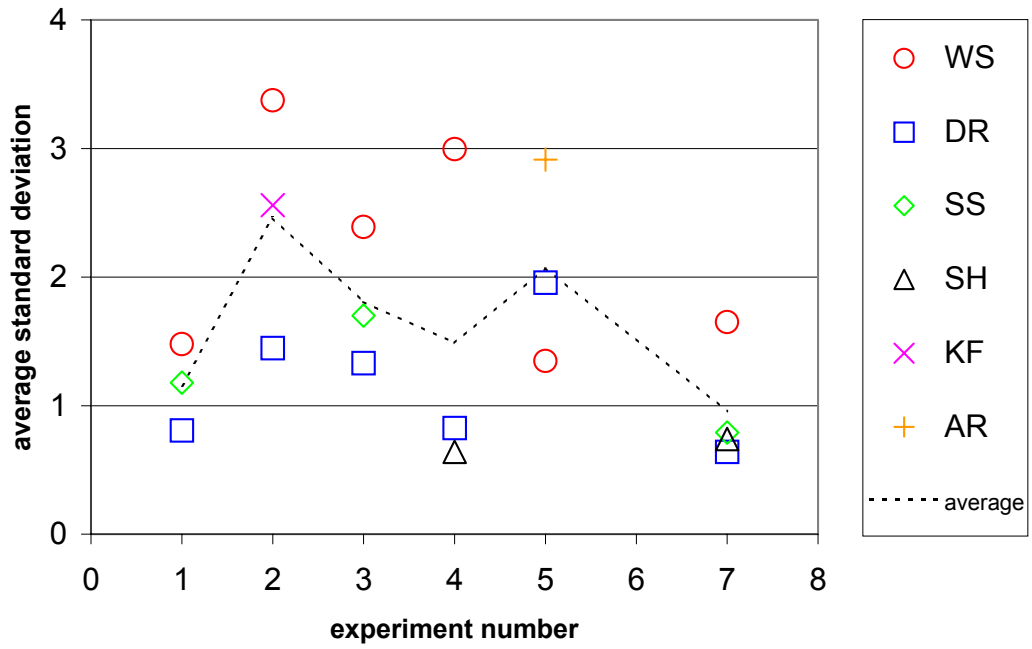


Figure 6.20: Standard deviation of masked threshold estimate for each subject, sorted by experiment

ling an idealised listener even more challenging, and is yet another reason why two listeners can have very different opinions about the perceived quality of a given audio system.

In conclusion, a candidate's performance in masking tasks may be inferred in part from an audiogram, but there seems to be another separate talent or process involved in detecting masked sounds that may be present to a certain amount in any individual. Some people are simply better at hearing masked sounds than others, just as some people are more musical than others – an audiogram cannot tell the whole story.

6.3.4.2 Subject reliability

In the previous section, it was suggested that the subject with the lowest thresholds may be considered the "best" subject, since they can hear sounds that others cannot. However, another measure of subject quality is the consistency of their masked threshold estimates. Figure 6.20 shows how the standard deviation of the masked thresholds varies between subjects for each experiment. Subject WS is the least consistent subject in all but one experiment, whilst subject DR is the most consistent subject in all but two experiments. Comparison with Figure 6.17 shows that assessing subject quality in terms of lowest average masked threshold yields the same ranking as assessing subject quality in terms of smallest variance in masked threshold.

Subject WS is judged the poorest subject by both methods, and subject DR is judged the best. The reader will note that in Figure 6.8 to Figure 6.14 -(c), the data from the most reliable subject in each experiment is plotted separately, and it is this data that will be used to train the auditory model.

Finally, it is important to remember that our directional hearing ability is due largely to the filtering effect of the pinna upon the incoming wavefront. Each human pinna is unique, and so the exact spectral features reaching each individual listener from a given point in space will be different [Moller *et al*, 1995b]. This will give rise to variations between candidates in any experiment that examines spatial masking, such that candidates may not even agree upon which spatial location is the most or least audible in any given task. This is shown in some of the results herein, especially those from experiment three.

6.3.4.3 Role of head movement

A lower threshold value indicates that the candidate found the stimulus to be *more* audible. Examining the results, it is surprising that so many of the 180° sources are easier to hear than those at 0°. The HRTFs for these directions are almost identical at 1 kHz, and in a perfectly symmetrical anechoic set-up there is no other reason for a source at 180° to be more audible than one at 0°. However, if the candidate moves their head, even slightly, then the two source positions cease to be equivalent, and the source at 180° is easier to hear.

The lower thresholds at 180° in the experiments employing 1 kHz stimuli may be due to head movements, a slight asymmetry in the experimental set-up, or the candidates HRTFs at 1 kHz not matching at 0° and 180°. The former is most likely, therefore these results must be taken to represent the response of a candidate who is nominally still, but not necessarily motionless. "Head clamped" responses may have been different. Note that the large difference between thresholds at 0° and 180° in the spatially separated codec noise test are almost certainly due to the HRTF, since a narrow spectral region around 9 kHz is 10 dB higher in the 180° HRTF compared to the same region within the 0° HRTF.

6.3.4.4 Comparison with existing data

Experiments one to five take standard psychoacoustic tests, and extend them by separating the masker and target. However, the first threshold measurement in each experiment was determined with the masker and target co-incident. These co-incident thresholds can be compared

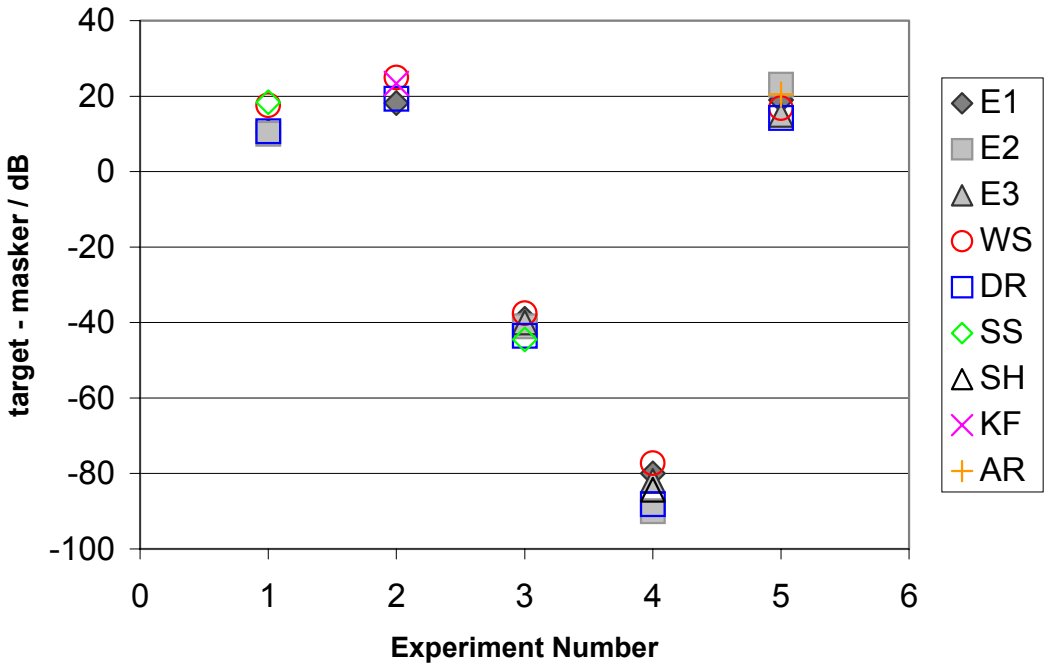


Figure 6.21: Comparison of experimental results with existing data

with those obtained by previous researchers. This will help us to judge the accuracy and validity of the current test.

Appendix D gives full details of the previous research data which is used as a reference, and also the transformation of all the data into a common format of dB SPL for tonal signals, and dB SPL /Hz for noise signals. Figure 6.21 shows up to 3 results from the listerature for each experiment (E1-3) as well as the data from each subject in the present experiment. Both the threshold values, and the variability between subjects, match those from previous experiments very well.

6.3.5 Knowledge gained

The experiments described herein have given us data with which to develop a theory of binaural masking, and to train and verify an auditory model. Beyond this, the results give some interesting new insights into human spatial hearing acuity.

The most startling results were obtained in the temporal masking experiment. Unlike every other task, thresholds are lowest if the target and masker are co-located. This may be because the masker, which occurs first, dominates our auditory attention. If the target is moved away

from it, it may be difficult for us to re-focus on the target location in the time allowed (10 ms), hence the threshold increases with target/masker separation.

The experiments with codec noise showed that separating the noise from the audio could have audible consequences. In experiment seven, when the codec noise was moved 90° away from the audio signal, it became 15 dB more audible. In experiment six, in the comparison of discrete vs matrix transmission of surround signals via psychoaocustic codecs, errors in the matrixed version were more audible in three out of four cases. It is difficult to draw general conclusions from the results of this experiment, since it used specific codec implementations and audio samples. However, it shows that concern over the psychoaocustic coding of matrix audio signals may be valid, but errors are nowhere near as audible as where the codec noise and audio signal are intentionally separated.

6.4 Conclusion

In this chapter, the concept of spatial masking has been introduced. Existing data has been reviewed, and an extensive series of spatial masking experiments have been undertaken. The results from these experiments give new knowledge on human spatial hearing which will be used to develop a binaural auditory model.

6.5 Acknowledgements

The spatial masking experiment described in this chapter was a joint project between the author and Prof. Woon Seng Gan, Associate Professor of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Many thanks to Woon Seng for his hard work and enthusiasm, and for travelling up and down the A12 to Ipswich at twice the legal speed limit in his home country.

The author wishes to thank BT labs for the loan of their facilities and equipment during the experiment. Thanks also to the staff at Martlesham Heath, who were very helpful during our time there.

Finally, the author would like to thank the subjects who kindly gave their time, and spent many hours in the anechoic chamber. They were Woon Seng Gang, Steven Sumpter, Steven Huxford, Kelvin Foo and Andrew Rimell. This experiment would not have been possible without their patience and co-operation.

7

Binaural Auditory Model

7.1 Overview

In this chapter, the binaural processing within the human auditory system is considered. Existing models of binaural interaction are discussed. The monophonic model discussed in Chapter 5 is extended by the addition of a binaural processor. The resulting binaural auditory model is calibrated with measured data, and is shown to correctly predict human performance in free-field masking experiments and horizontal-plane localisation discrimination tasks. Finally, directional accumulators are defined, to judge the overall change in the stereophonic sound stage.

7.2 Introduction

It is often said that two ears are better than one. The presence of two ears allows the auditory system to compare the signals at each ear, thereby deriving two benefits. Firstly, the interaural time and level differences for a given sound source provide strong directional clues. Secondly, if two sources are spatially separate, our two point "sampling" of the sound field allows us to concentrate on one sound in the presence of the other. This ability can be measured in two dimensions as the Binaural Masking Level Difference, or in three dimensions as Spatial Masking.

The principle aim of the present model is to simulate binaural and spatial masking. The localisation of sounds is not a primary requirement of the present model. However, the two mechanisms (binaural release from masking, and localisation) are closely linked, and the present model may be adapted to localisation tasks using techniques described in the literature.

Within this chapter, the term "binaural detection" will be used as a convenient description of any task where binaural processing allows the listener to detect a change which is not audible using monophonic (singled-eared) listening alone. The Binaural Masking Level Difference is one such phenomenon. The spatial masking experiments described in Chapter 6 also count as binaural detection tasks where the masker and target are separate, but not where the masker and target are co-located.

The detection of a change in source location can also be referred to as a binaural detection task. This is only true where the change in location is not detectable due to the signals reaching either ear in isolation. Though the present model is not required to correctly predict the location of a sound source, it should detect any *change* in perceived location. This is relevant in the field of audio quality assessment. For example, the stereo image may be changed in some manner by the coding process, and the model should detect this change.

7.2.1 Webster's Hypothesis

[Webster, 1951] suggested that all categories of binaural detection may be equivalent at some stage of internal binaural processing. It is assumed that the auditory system attempts to calculate the interaural time delay by some form of correlation, convolution, or a functionally equivalent process. It is suggested that the just detectable change in the output of this process will correspond to the threshold condition in *any* binaural detection task. At the simplest level, the just detectable change in the calculated ITD corresponds to the just detectable change in the true ITD. Hence, this mechanism sets the limit on our spatial acuity. However, Webster suggests that spatial masking may be attributable to the same mechanism. Thus, at threshold, the *presence* of the target signal changes the calculated ITD by the same "just detectable" amount.

Webster apparently disproved his own hypothesis. However, in this chapter the hypothesis is revisited using current auditory modelling techniques, with interesting results.

7.2.2 Data

The measurements described in the previous chapter will be the chief source of calibration data for the binaural model. To test Webster's hypothesis, the just noticeable change in incident source angle is also tested. This is measured in [Mills, 1958] as equating to an ITD of 11μs at a frequency of 1 kHz.

7.2.3 Physiology

The monophonic model described in Chapter 5 was based upon the known physiology of the human auditory system, and a similar approach will be attempted here. Unfortunately, the binaural processing within the HAS is carried out within the neural domain. As such, it is less well understood than the processes that transform sound information from an air-born pressure wave into a neural signal. However, some knowledge has been gained from the mapping of nerve fibre interconnections, and the probing of individual cells.

The role of the Superior Olivary Complex in the lateralisation of sounds was discussed in Chapter 3. In summary, it is known that signals from both ears are compared within this region, and it is highly likely that interaural time and level differences are determined therein.

Whilst these are essentially lateralisation processes, they are the only known processes within the auditory system where signals from both ears are compared. Thus, the known physiology gives some weight to Webster's hypothesis, suggesting that these same mechanisms may be responsible for the binaural masking level difference, and the spatial masking phenomena measured in the previous chapter. In addition, models considering only interaural time difference have correctly predicted binaural masking level difference data (e.g. [Colburn, 1977]).

For these reasons, both lateralisation and detection models will be discussed here in relation to the present task.

7.3 Existing models

An excellent review of many models of binaural hearing is presented in [Colburn and Durlach, 1978]. A slightly more recent discussion is found in [Blauert, 1983]. The present review concentrates on models that have contributed in some way to the development of the present model.

7.3.1 Jeffress model

[Jeffress, 1948] proposed a model of binaural interaction which has formed the basis of much subsequent research. His hypothetical neural network is shown in Figure 7.1. The neural signals from the left and right ears are assumed to travel at a finite rate along the nerve fibres. In this way, each nerve fibre can be thought of as a delay line. The "tertiary fibres" are more likely to fire when signals from the left and right ears reach them "almost simultaneously". Thus, it can be seen that the matrix of fibres in Jeffress' model acts as a network of time delays and co-incidence detectors, the purpose of which is to measure the interaural time delay. Interaural intensity differences are incorporated into the model by assuming that louder signals propagate along the nerve fibres at a higher velocity. This is termed the "latency hypothesis", and has received less universal acceptance than the rest of the Jeffress model.

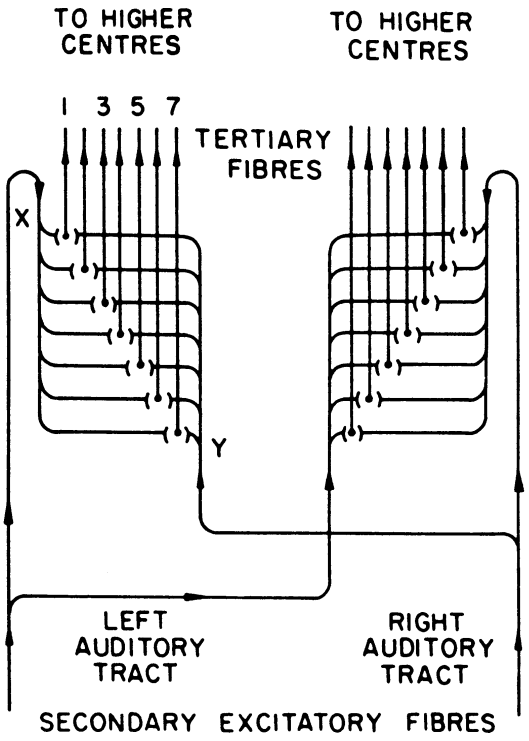


Figure 7.1: Jeffress binaural interaction model [Jeffress, 1948]

This model predicts the lateralisation of sound sources; the tertiary fibre exhibiting the largest output is assumed to correspond to the perceived lateral location of the sound source. The transformation from sound wave to neural signal is poorly specified in this model, as such data was unavailable at the time of the model's conception. However, with certain assumptions, this model can predict much lateralisation data. These specific assumptions are common to most models, and will be discussed in detail later, with reference to a more advanced model.

There are several inconsistencies between the predictions of the Jeffress model and the performance of the human auditory system. The most notable problem is the model's performance near threshold. Since no internal noise is specified, the model exhibits accurate performance *below* threshold, whereas the lateralisation accuracy of human subjects is impaired as the intensity of a sound source approaches threshold. Another problem arises from the complete

trading of time and intensity information at the neural level. This suggests that a single lateralisation judgement is passed to the higher auditory centres. However, as discussed in Appendix E, this is unlikely, as the presentation of a single source with contrary ITD and ILD cues can lead to the perception of two sound sources. This is not possible with any model that combines these two cues before they can be "perceived".

Despite these problems, it may be possible to apply this lateralisation model to the task of detection. This hypothesis is tested in [Webster, 1951]. Webster assumed that the lateralisation model may yield extra information which is not apparent from the individual ears signals, and this extra information may cause the Binaural Masking Level Difference. His hypothesis states that, at threshold, a signal will cause a just detectable change in the output of the lateralisation model. Webster did not have access to models of inner hair cell responses, or to the time domain analysis used throughout the present model. To test his hypothesis, he assumes that the input stimulus passes through a bank of narrow band-pass filters (equivalent to the Basilar Membrane response), and that the output of each filter approximates to a pure tone. The frequency of this tone is equal to the centre frequency of the band-pass filter, and the phase of this tone is determined by the input stimulus. Thus, after processing via the filter, a noise masker and tonal target are both converted to tone-like signals. Where both are presented simultaneously, the resulting phase of the combined tone is calculated via simple algebra. The interaural network detects the phase (time) difference between the two ear signals, and this difference is dependent upon the amplitude of the target. Webster suggests that any stimuli which change the interaural time difference by more than a fixed amount will be audible. The fixed threshold is assumed to be due to internal noise.

Webster demonstrates that this model correctly predicts some binaural detection data. [Jeffress *et al*, 1956] apply Webster's hypothesis to further monophonic and binaural phenomena, with some success. To account for a wide range of data, the value of the just detectable ITD must be of the order of 100 μs. That is, a signal at threshold (where that threshold is due to binaural unmasking) causes a change in the ITD of 100 μs. This is a major problem for the model, since the just detectable ITD in experiments where this quantity is measured directly is around 10 μs [Mills, 1958]. If Webster's hypothesis is correct, it seems strange that the auditory system should be an order of magnitude less accurate when using ITD information to detect the presence of a sound compared to the use of ITD information to locate a sound. However, this

apparent contradiction may be a consequence of Webster's deterministic model; a noisy or stochastic model may solve this problem.

A significant fact emerges from these discussions, and this fact has a parallel in all binaural models. In order to account for the reduced sensitivity of the HAS to larger ITDs, it is necessary to specify the density of tertiary fibres. It is suggested that there are a relatively large number of fibres corresponding to ITDs around zero, and that this number decreases as the ITD increases. This simulates the reduced lateralisation accuracy exhibited for sound sources outside the median plane. It also allows the model to correctly predict the threshold of detecting sounds in the presence of non median plane located maskers.

7.3.2 Cross correlation and other models

The neural matrix at the heart of [Jeffress, 1948] is functionally equivalent to a cross correlation process, providing that the correlation is only computed for a range of samples $\pm t$, where t represents the maximum ITD which can be detected by the binaural processor. [Sayers and Cherry, 1957] pioneered this approach for lateralisation, whilst [McFadden, 1968] adapted a cross correlation model for use in detection tasks. Many others have proposed variations on the cross correlation model, but the details of these models are often incomplete, rendering any quantitative analysis impossible (e.g. [Osman, 1971], where the value of the permitted time and/or phase shifts are not clearly specified).

A complete consideration of all the known models of binaural interaction is not appropriate here. However, given the obvious similarities between many of these models, a review of three areas where there are large variations is presented.

Firstly, the newer models employ increasingly sophisticated pre-processing of the input stimulus, before the cross correlation itself. If the pre-processing within the model matches that within the HAS, then the model is found to simulate human performance in a wider range of tasks [Theiss, 1999]. This is encouraging in respect of the present model, since the peripheral processing matches that within the HAS to a significant degree of accuracy.

Secondly, the outputs of the cross correlation are often combined into a single "correlation coefficient" (an unfortunate name, since this sometimes has little to do with the statistical device of the same name). This can be used in detection experiments, together with two mono-

phonic variables. Alternatively, all the "rightwards" coefficients may be summed, and compared with the sum of all the "leftwards" coefficients. This comparison leads to a prediction of lateralisation. There are yet further alternatives. Most significantly, the outputs of the correlation can be combined in *any* optimum or sub-optimum manner, in order to match human perception. Many models make bold claims to predict almost all binaural data, but require a different weighting of the correlation outputs to accomplish each individual task correctly. This may or may not represent the real processes present within the HAS, but such an approach cannot be applied to the present time-domain model, which must correctly detect just noticeable differences for a wide variety of binaural tasks, without prior knowledge of the task itself.

Thirdly, the inclusion of interaural intensity differences varies from model to model. No recent model incorporates the latency hypothesis of Jeffress, but some still trade time and intensity differences completely. Many lateralisation models incorporate a separate IID detector, and feed both ITD and IID information to a decision device or weighting function (e.g. [Macpherson, 1991]). Some detector models (e.g. [Colburn, 1977]) do not include an IID detector, yet correctly predict most binaural masking data. However, the efficacy of such models at higher frequencies, where the ILD cue is most important, has not been extensively tested.

7.3.3 Colburn's neural model

The most successful detection model is that of [Colburn, 1977], which is based upon hypothesised neural processing. At the heart of the model lies a "binaural displayer" that is similar in concept to the Jeffress model, but includes many physiologically relevant features. The entire Colburn model is very well specified, allowing direct quantitative prediction. For these reasons, the model will be discussed in detail. A diagram of the model is shown in Figure 7.2.

The signals for each ear are processed via a peripheral transducer. The output of this transducer consists of a set of neural signals. Rather than specifying the firing probabilities upon the nerve fibres, the actual neural spikes are transmitted. These spikes are caused by the firing of the inner hair cells. Several fibres carry independent signals from each point upon the Basilar membrane; hence, a true realisation of this model would incorporate many thousands of neural signals. These signals are fed to both monaural and binaural processors. The monaural processor is not specified in detail, but it is assumed to account for signal detection where binaural effects are unimportant. As such, it sets an upper threshold limit, and is assumed to predict the correct masked threshold in all tasks where the BMLD is zero.

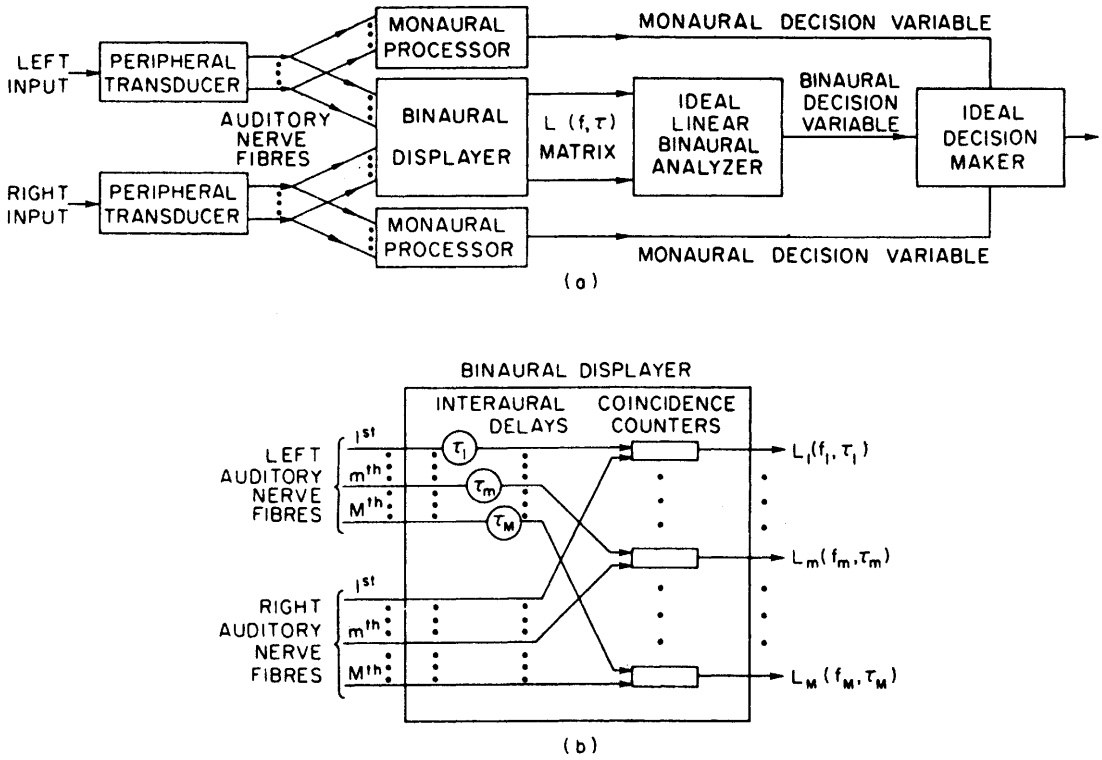


Figure 7.2: Colburn's binaural interaction model [Colburn, 1977]

Auditory nerve signals from both ears are fed into the Binaural Displayer, where they are compared. Initial research [Colburn, 1973] suggested that an ideal binaural processor operating on these neural signals would exceed the performance of the HAS. In order to match human performance, the following two restrictions are imposed. Firstly, the signals on a given fibre can only be compared with those upon *one* other fibre, having the same frequency selectivity, originating from the opposite ear. Secondly, for a given pair of fibres, only firings that occur "almost simultaneously" (with 100 μs) after an internal, interaural delay that is fixed for the pair can be compared.

The illustration in Figure 7.2 is slightly misleading, since the output of the binaural displayer is a two-dimensional matrix $L(f,\tau)$, where one co-ordinate represents frequency (f), and the other represents interaural time delay (τ). Thus, for a noise signal presented to both ears without interaural delay, the largest values within the matrix would lie along the column $L(f,0)$; for a pure tone of frequency f_1, the largest values would lie along the row $L(f_1,\tau)$. The value at each point of the matrix is the total count of the coincidences for that frequency and delay over the stimulus interval. This represents an important restriction of many binaural models. Rather than producing a time varying output, the model commences a count or integration at the start

of the stimulus, and only post-processes the total value that has accrued at the end of the stimulus interval.

As in the Jeffress model, the distribution of the coincidence detectors is non-uniform. [Colburn, 1977] explicitly states the distribution $p(\tau)$, thus:

$$\begin{aligned} p(\tau) &= C && \text{for } |\tau| <= 0.15 \\ &= C\exp\left(-\frac{|\tau|-0.15}{0.6}\right) && \text{for } 0.15 < |\tau| <= 2.2 \\ &= 0.0333C\exp\left(-\frac{|\tau|-2.2}{2.3}\right) && \text{for } |\tau| > 2.2 \end{aligned} \qquad (7\text{-}1)$$

where τ is the interaural time delay corresponding to the fibre pair, C is chosen to give $p(\tau)$ unit area, and all time units are in ms. A plot of $p(\tau)$ is shown in Figure 7.3. ITDs above 1-2 ms are of little importance within the HAS because larger ITDs do not occur for real world sound sources. This limit is due to the distance between ears, which is equivalent to a delay of around 1 ms.

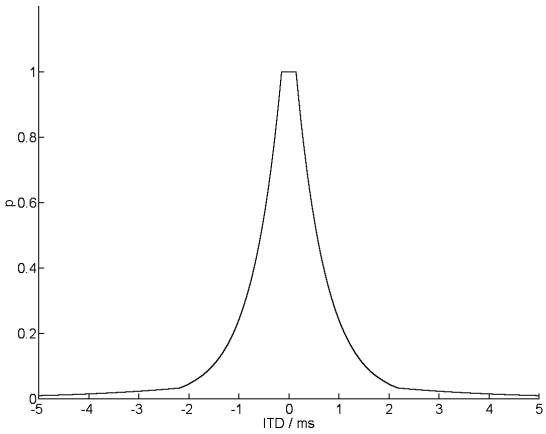


Figure 7.3: $p(\tau)$ distribution of coincidence counters with ITD in Colburn's model

The output matrix of the Binaural Displayer is fed into an Ideal Linear Binaural Analyser. This calculates a weighted average of all the matrix components, which is then judged by an ideal decision maker. The weighting coefficients within the analyser are chosen independently for each task. For example, where the task is the detection of a 1 kHz tone within broadband noise, the fibres corresponding to a frequency around 1 kHz are likely to contain the most pertinent information. The ideal decision maker will determine whether the binaural decision variable is above or below a pre-set threshold, and flag the presence or absence of the target appropriately.

The model is stochastic, due to the peripheral transducer, which simulates hair cells firing in a pseudo random manner. Stochastic or deterministic stimuli may be used. Within a masking experiment, the threshold is defined as the amplitude of the target (or phase of the target, depending on the nature of the experiment) at which the target is correctly detected in 75% of the stimulus presentations. Rather than attempting a time-domain simulation (which would have been impractical in 1977), Colburn creates analytical solutions. This limits the model, since it can only be applied to sources for which formulaic representations of the output of the peripheral transducer are known. There is no fundamental reason why a time-domain implementation of the model could not process any arbitrary signal, by simulating the signal paths through the model. However, the required number of neural fibres makes this task computationally burdensome, even at the present time.

A model suitable for time-domain implementation may be formulated as follows. The model will be based upon the model of Colburn, with some significant alterations. These alterations may or may not cause the performance of the model to disintegrate beyond use, and this is one important reason for testing and evaluating this approach.

Firstly, the peripheral transducer may be replaced by the monophonic model described in Chapter 5. The decision device of the monophonic model is not needed, nor is the final 20 ms low pass filter of the hair cell response, since this removes the fine structure of the hair cell response which required by the interaural time delay calculation.

Secondly, the model can be converted from stochastic to deterministic. This may worsen performance near threshold, but it will allow the use of one pair, rather than many pairs of neural fibres per frequency. As discussed with respect to the monophonic model, this compromise removes the repetition inherent with a time-domain stochastic model.

Thirdly, the process of summing the coincidence occurrences over the entire stimulus presentation is not appropriate if the stimulus is a long, time varying signal (e.g. music). It is suggested that a temporal integration window may be appropriate.

Finally, the manner in which the integrated signal is post-processed cannot be task dependent. Whilst it is reasonable to weight the contributors to the decision variable in a task dependent manner, this approach is not applicable to an audio quality assessment model, where the nature of the possible defect is not always known. It is suggested that the premise of Webster, namely

that any target at threshold produces a just noticeable change in ITD, may be used to develop a simple universal detector.

A new binaural interaction model will now be described, which is based upon the Jefrress and Colburn models, modified in these four respects.

7.4 Binaural model

In the following sections, each stage of the model is described in detail.

7.4.1 Pre-processor

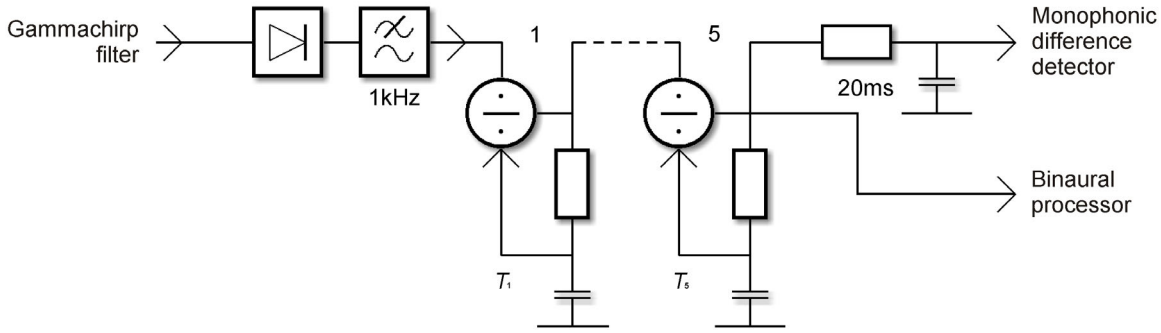


Figure 7.4: Adaptation circuit from monophonic model

showing input from gammachirp filter bank, and outputs to monophonic difference detector and binaural processor

The monophonic model described in Chapter 5 is used as a pre-processor for the binaural model. All procedures and comments pertinent to the monophonic model are also relevant here. Two monophonic models are employed, one for each ear. In addition to acting as pre-processors to the binaural model, the monophonic models are also used exactly as before in order to provide detection of monophonic differences. The physiological based sections of the monophonic model are analogous to the peripheral transducer within Colburn's model. The difference perception module performs the same role as the Monaural Processor, decision variable, and decision maker within Colburn's model.

The final low pass filter within the physiological section of the monophonic model forms part of the monophonic detection process. As such, it represents a physiological component within the higher processing centres, rather than a process that affects the signal upon the auditory nerve. The input to the binaural model is taken from immediately before this filter, as shown in

Figure 7.4. This signal represents the inner hair cell firing *probability*, which is a deterministic variable. This is in contrast to Colburn's model, which uses stochastic hair cell firing events. Since the monophonic model is deterministic, then the binaural model will also be deterministic, so long as no noise sources are added within the binaural processor.

7.4.2 Element by Element Vector Multiplication

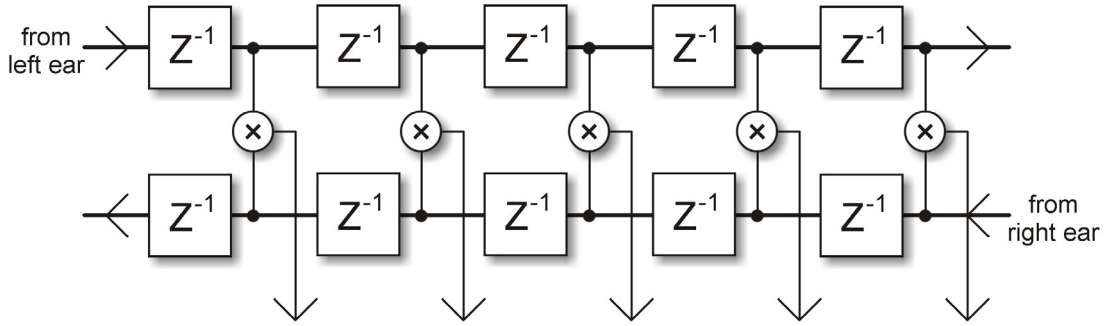


Figure 7.5: Element by Element Vector Multiplier

The signals from each monophonic processor are fed into the network shown in Figure 7.5, known as the Element by Element Vector Multiplier (EEVM). This provides a finite length partial correlation, carried out on a sample by sample basis. The process is carried out in MATLAB by calculating the vector dot product between the two rows of samples. If the operation and function of this network is clear, the reader may wish to pass over the following paragraph.

An appropriate way to visualise the EEVM is to imagine the time versus amplitude signals from each ear travelling towards each other from opposite directions. As the signals overlap, the instantaneous amplitudes of the left and right signals at each point of overlap are multiplied. The result of this multiplication is a time varying series, or two dimensional matrix, where one dimension is time. The other dimension, which corresponds to left/right location in this imaginary view, is an indication of interaural time delay. Signals that reach the right-hand ear first give rise to an internal peak on the left-hand side, and the opposite is also true. This mirroring occurs because the signal from the right-hand ear is travelling from right to left in this picture, and will travel a long way leftwards before meeting the delayed signal coming from the left-hand ear, thus giving rise to a peak.

In the symmetrical network shown, *both* ear signals slide past each other on a sample by sample basis, i.e. there are delays in both signal paths. One unwanted result is that a given sample from the left ear will be multiplied by every second sample from the right ear. For example, the first sample from the left ear will be multiplied with samples two, four, six, eight, and so on from the right ear. This approach is less complex to implement than the multiplication of every sample, since this latter approach would require half sample delays, or else would generate two output samples for every input sample. Skipping every other sample in this manner does not cause significant problems (e.g. aliasing), because the signal is low-pass filtered within the preceding monophonic model.

The total EEVM length is chosen to be 5 ms, so contra-lateral signals separated by more than 5 ms are never compared. If the sampling frequency is given by *fs*, then the EEVM is carried out over *N* samples, where

$$N = \text{round}(0.005\, fs) \tag{7-2}$$

7.4.3 Windowed summation

In the models of both Colburn and Jeffress, the coincidence detectors only respond to signals that occur "almost simultaneously". The response of the coincidence detector could be characterised by a near-rectangular time-domain window function, centred on the time delay in question. The width of this window is approximately 100 μs [Colburn, 1977]. Values of 50-100 μs have been used to represent "almost simultaneously" in most quantitative binaural models.

A similar effect can be achieved via the EEVM by grouping the outputs into 100 μs blocks, and summing the outputs in each block. This process is illustrated in Figure 7.6.

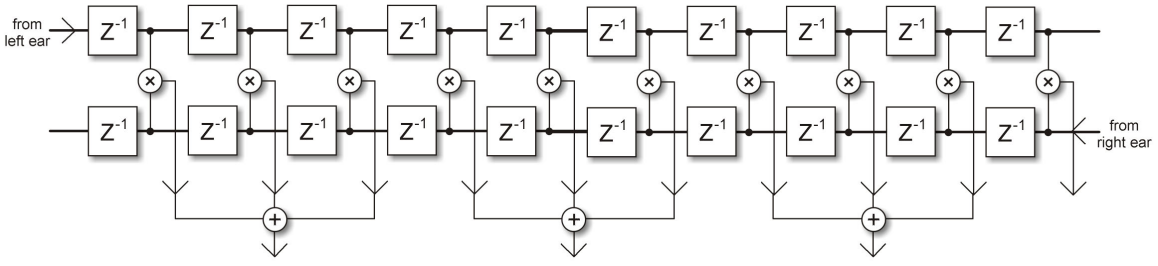


Figure 7.6:Rectangular windowing of EEVM output by summation

There are three problems with this approach. They occur because Colburn describes a continuous time stochastic model, whereas the EEVM forms part of a sampled time deterministic model.

Firstly, at typical audio sampling frequencies (e.g. 44.1 kHz), blocks of 100 μs do not sit evenly within whole samples. In Figure 7.6, each block length must be an integer number of samples. At 44.1 kHz, the block lengths must be multiples of 22.676 μs. Such a restriction severely limits the possibility of tuning the model to match human perception.

Secondly, the coincidence detectors are not spaced at regular 100 μs intervals, even though each coincidence detector has a 100 μs range, this does not mean that the coincident detectors are spaced 100 μs apart. There will almost certainly be some overlap between the detectors within the HAS. Not only is the neat division implied by Figure 7.6 unlikely to arise in neural processing, but also the stochastic nature of neural processing requires redundancy of information, so that the signal can be detected within the internal noise. Overlapping detectors would yield the required redundancy.

Finally, the distribution of coincidence detectors is not uniform for all values of interaural time delay. A function describing the distribution was introduced in Section 7.3.3. If this distribution is to be matched by non-overlapping constant width rectangular window functions, this implies there will be gaps between the window functions in the regions corresponding to larger interaural delays. However, these gaps would give rise to regions where the HAS is unable to locate a sound source. Such arbitrary gaps do not exist within the HAS – rather, the decrease in density of coincident detectors, combined with the internal noise, reduces the information available at higher ITDs. In effect, the signal to noise ratio of the ITD determination process becomes poorer as the ITD increases, so the spatial accuracy decreases.

This loss of resolution for high ITDs can be simulated within a deterministic model by varying the window *width* (rather than detector density) in accordance with the distribution function described by equation (7-1). However, if a rectangular window is employed, then the window transitions must occur at integer sample locations. After rounding to integer sample boundaries, the specified distribution function would be severely compromised, and the performance of the HAS would be poorly matched.

In effect, the spacing of the coincidence detectors (and the statistics of the internal noise) limit the accuracy of the internal ITD calculation. Using the same technique discussed in Chapter 5, the stochastic detection of a signal in noise can be replaced by the deterministic imposition of a fixed threshold. Where the signal exceeds the threshold, this is functionally equivalent to being detected within the internal noise. In the binaural detector, it is not the presence or absence of the signal that must be detected, but the location of its peak. Hence, the threshold value defines the minimum detectable change in internal peak location.

Using this criterion with a rectangular EEVM window would lead to undesirable results. The window function is effectively sampling in the time *delay* domain, hence a rectangular window would lead to aliasing within this domain. As the peak moves *within* the window, no change is detectable, but as the peak moves from one window to the next, the change is detected. This would cause the minimum audible angle to oscillate in the time delay domain, dipping at window boundaries where the process is most sensitive. Such an effect is not found within the HAS. A rectangular EEVM window within a deterministic detector causes problems at window transitions, and the position and spacing of these transitions is vital. A rectangular window within a stochastic model is effectively smoothed by the internal noise, and the problems at window transitions are also smoothed by this noise.

All these factors suggest that a different window shape is required within a deterministic model to yield the same effect as the rectangular window within the stochastic model. The solution is to use a series of overlapping gaussian windows. A gaussian distribution is given by the following equation:

$$f(x) = \frac{1}{(2\pi\sigma)^{1/2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7-3}$$

where μ is the mean, and σ is the standard deviation.

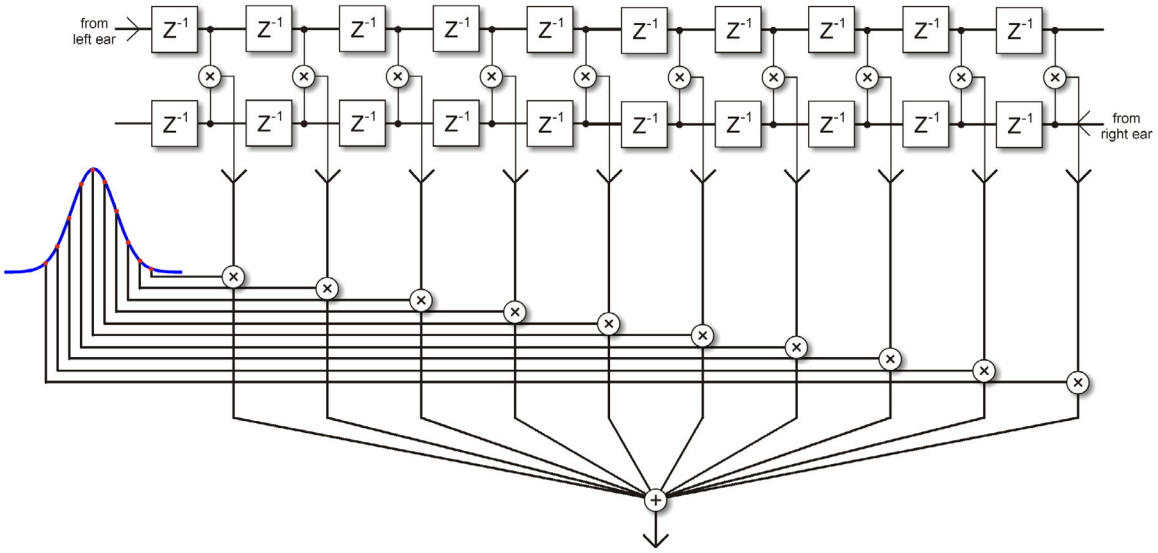


Figure 7.7: Gaussian windowing of the EEVM output to generate one channel for further processing

The manner in which a single gaussian window acts upon the output of the EEVM is illustrated in Figure 7.7. A group of time-varying EEVM outputs are multiplied by the gaussian window function, and summed into a single time-varying signal.

A series of gaussian windows are required. *n* is defined as an index into the time-varying EEVM outputs, such that *n* can be any value between 1 and *N* (defined in (7-1)). *m* is defined as an index into the time-varying outputs of the gaussian window summation, such that *m* can be any value between 1 and the number of gaussian windows *M*. Both *M* and *N* are chosen to be odd, so that the central window and EEVM output can be easily identified. The series of window functions can be described by the following equations:

$$gaus\_window(n,m) = \frac{1}{\left(2\pi\sigma(m)\right)^{1/2}} e^{-\frac{\left(t(n)-\tau(m)\right)^2}{2\sigma(m)^2}} \tag{7-4}$$

$$\sigma(m) = \frac{0.05}{p(\tau(m))} \tag{7-5}$$

where $p(\tau)$ is defined in equation (7-1). Since the outputs of the EEVM are spaced according to the sample rate *fs*,

$$t(n) = \frac{n}{fs} \tag{7-6}$$

The mean of each gaussian function is given by $\tau(m)$, which represents the interaural time delay at the centre of the gaussian window. $\tau(m)$ cannot be defined via a single formula. In the MATLAB implementation of the model, it is calculated via the following criteria:

$$\tau\left(\frac{M}{2}\right) = t\left(\frac{N}{2}\right) \tag{7-7}$$

This equation specifies that the central gaussian window is centred on the central EEVM output. Whilst this is not essential, it acts as a reference point, and aids visual inspection of the output from binaural processor. Since M and N are both odd, and MATLAB only accepts positive integer indexing into arrays, this equation is correct if both indices are rounded upwards. The means of the other gaussian windows are specified iteratively, working outwards from the central window, thus:

$$\begin{aligned} \tau(m) &= \tau(m-1) + \sigma(m-1) \qquad \text{for } M > m > M/2 \\ \tau(m) &= \tau(m+1) + \sigma(m+1) \qquad \text{for } M < m < M/2 \end{aligned} \tag{7-8}$$

A set of gaussian window functions generated using these criteria is illustrated in Figure 7.8. The height of each window is normalised to give each window unity area.

This set of gaussian windows act in the time-delay domain, but the information within this domain is itself changing on a sample by sample basis. The auditory system does not react instantaneously to this information, and a suitable process is required to simulate this fact.

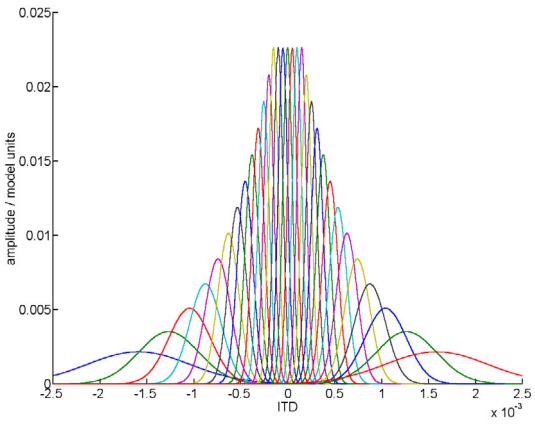


Figure 7.8: Gaussian window bank to sum outputs of EEVM into channels

7.4.4 Time domain windowing of binaural processor output

As discussed in 7.3.3, Colburn's model counts the number of coincidences that occur upon each fibre pair over the entire stimulus interval. This approach is not appropriate for time varying stimuli, and another is sought.

Perceived lateralisation does not depend on the instantaneous output of the EEVM. As discussed in Chapter 3, the HAS exhibits binaural sluggishness, and does not respond immediately to a change in the binaural aspect of a stimulus. Most models simulate this property via a leaky integrator, though the choice of time constant varies. A leaky integrator implies a single sided exponential window function, as shown in Figure 7.9. This is an unfortunate choice, since it causes the model to respond immediately to changes in the binaural aspect of the stimulus, though it does cause the lateralisation perception to persevere beyond the offset of the stimulus.

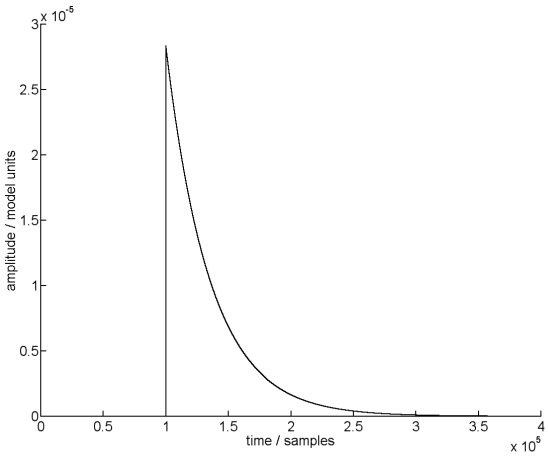


Figure 7.9: Single sided exponential response given by a leaky integrator

A leaky integrator was tested with the present model. It was found that erroneous lateralisation judgements occurred at the onset of sounds. Though a leaky integrator is computationally efficient, a more accurate solution is sought.

Recently, [Culling and Summerfield, 1998] have tried to define the shape and duration of the binaural temporal integration window. From experimental data, it appears that a 100 ms asymmetrical gaussian window is appropriate[1].

1 Culling and Summerfield warn against the blind adoption of this window for binaural auditory models such as the one developed herein because there is evidence that the auditory system can occasionally reset the binaural temporal integration process during brief periods of silence [Yost, 1985]. However, the criteria which cause the reset during some periods of silence, but not others, has yet to be determined. As the temporal window fits almost all known data, it is the best option currently available, and is used within the present model.

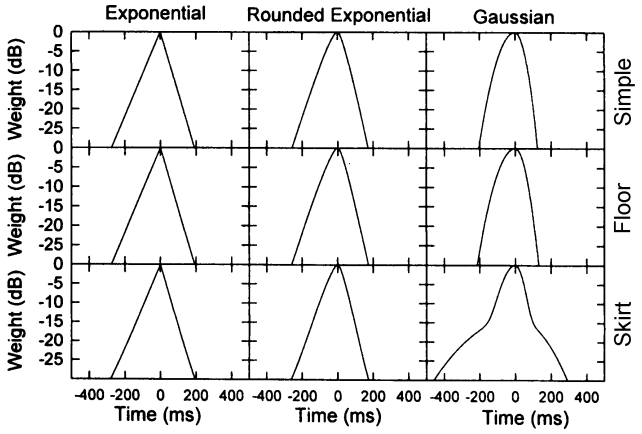


Figure 7.10: Possible binaural temporal windows [Culling and Summerfield, 1998]

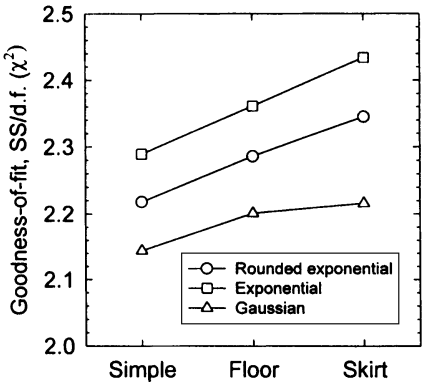


Figure 7.11: Goodness-of-fit for each of the window functions in Figure 7.10

A computationally efficient implementation of an asymmetric gaussian window function is not immediately obvious. An FIR filter would require many thousands of taps. For example, at a sampling frequency *fs* = 44.1 kHz, the suggested window function would require 20000 taps to achieve accuracy down to –25 dB. Such a filter is required for each output channel of the previous window function (there being 31 binaural channels in the present implementation). Also, recall that the EEVM will be performed on each of the 48 channels of the monophonic model output. Thus, the temporal window will be applied 1488 times. With 20000 taps, this equates to 29760000 multiply and accumulate operations per input sample. The exact computation time for this operation has not been calculated, but it would certainly require several hours of processing for 1 second of input using current hardware.

The computationally efficient leaky integrator is too inaccurate, while the accurate gaussian window is too computationally burdensome. For this reason, a compromise is sought. Figure 7.10 shows the window functions considered by Culling and Summerfield, and Figure 7.11 shows the mean goodness-of-fit [Snedecor and Cochran, 1989] for each of these window functions. A lower χ^2 value indicates a better fit. The simple window functions out-perform those with more parameters (called floor and skirt), and the gaussian window functions provide the best overall fit. However, the most striking feature is that all the windows are so similar in their goodness-of-fit. Whilst a single sided exponential response (the leaky integrator) is a poor choice, the double sided exponential window function is almost as good as the gaussian. If an efficient double sided exponential can be implemented, then it would represent a suitable compromise.

A simple IIR filter can only generate a single sided exponential response. However, by using two such filters, and operating one of the filters in the reverse time domain, it is possible to generate a two-sided exponential filter. Moreover, since each side of the response is dependent upon a separate filter, an asymmetric window can be generated, and the time constants of each slope can be adjusted independently.

From the data of Culling and Summerfield, it appears that the ideal window may be slightly longer at lower intensities, and possibly becomes shorter at higher frequencies. However, current data does not make either dependence clear, and the best approach is to generate a single window function that can be modified to be frequency and intensity dependent if reliable data becomes available. The appropriate window shape is calculated from the data in [Culling and Summerfield] by averaging the best-fit asymmetric exponential window responses, as shown in Table 7.1.

level / dB	freq. / Hz	P_r / ms	P_f / ms
40	1000	37.7	10.6
40	500	39.9	27.8
40	250	59.7	34.3
40	125	28.8	39.8
50	500	33.4	16.4
30	500	42.9	29.5
20	500	47.4	25.9
	Average:	41.4	26.3

Table 7.1: Time constants of exponential window from [Culling and Summerfield, 1998]

Thus, P_r = 41.4 ms and P_f = 26.3 ms. P is the time constant of the exponential, such that

$$f(t) = e^{-\frac{t}{P}} \tag{7-9}$$

In psychoacoustic parlance, "forward" refers to the response before the stimulus, and "backward" or "reverse" refers to the response after the stimulus. Hence, P_r is the time constant of the IIR filter acting in the forwards time domain, which gives the exponential decay after the stimulus. Conversely, P_f is the time constant of the IIR filter acting in the reverse time domain, which gives the exponential response before the stimulus. To compute the IIR filter response in the reverse time domain, the input stimulus is reversed, processed via the filter, reversed a second time, and added to the output from the forward-time filter. The overall response is shown in Figure 7.12.

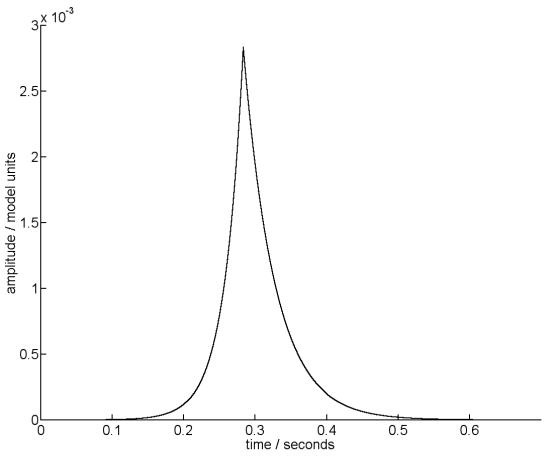


Figure 7.12: Asymmetric exponential window function

Having smoothed the output of the EEVM in both the time-delay and time domains, three further stages are required to complete the calculation of the interaural time delay.

7.4.5 Subtraction of the mean

At this stage in the binaural processor, the signal peak in the time-delay domain is an indication of the ITD, but it is swamped by changes in the overall amplitude. The amplitude is irrelevant to the ITD, so this information should be removed. This step is not essential to the model, but it does simplify subsequent processing: before this stage, the maximum binaural signal amplitude is a function of the input signal amplitude. After this stage, the maximum binaural signal amplitude is an indication of the accuracy of the ITD estimate.

This transformation is achieved by subtracting the instantaneous mean of the binaural channels from each binaural channel.

7.4.6 Oversampling

The contents of the channels of binaural processing at this stage represent the information that is available to the auditory system. However, the channel containing the largest value does *not* represent the ITD, since the channels are spaced approximately 50 μs apart, but the auditory system can detect a 10 μs change in ITD. At first sight, this is a severe mismatch between the model and human perception. However, by comparing the relative signal amplitude in the channels either side of the maximum value, an ITD estimate may be calculated that is not rounded to a single channel. From sampling theory, it is known that a linear interpolation is not the best way to determine the true location of an inter-sample peak. Greater accuracy may be achieved via optimum oversampling. This approach has been suggested by [Theiss, 1999], where the data is oversampled, and smoothed using a gaussian window.

In the MATLAB implementation, the in-built `resample` routine provides a fast and efficient solution, and is used in the present model. The channels are oversampled by a factor of one hundred, which yields a resolution of 1 μs. This is ten times greater than the measured accuracy of the human auditory system, since matching the accuracy at this stage would prevent the model being tuned in the detection process that follows[2]. The use of oversampling in this manner is discussed further in Appendix F.

7.4.7 Peak Determination and confidence value

After oversampling, the channel containing the peak value represents the calculated ITD. This ITD infers the lateral location of the sound source.

There are three problems with this approach. Firstly, the binaural model will always calculate an ITD value, even during silent intervals. This ITD value is meaningless, and should be ignored. Secondly, the accuracy of the ITD calculation is independent of the stimulus intensity. This is in contrast to the performance of the HAS, where the minimum audible angle reduces with stimulus intensity. The just noticeable phase difference measurements from [Zwislocki and Feldman, 1956] are re-plotted in Figure 7.13 to show just detectable ITD as a function of

[2] A 10 μs change in ITD can only be detected under controlled listening conditions. If less critical listening is simulated, an oversampling ratio of ten may be adequate, thus increasing computational efficiency.

stimulus intensity. As the stimulus intensity is reduced from 70 dB to 30 dB, the just detectable ITD change is doubled. The model should simulate this. Finally, the ITD value calculated for a diffuse sound source is of questionable value, and it should be treated accordingly. Unfortunately, no quantitative data is available to define this phenomenon, but some appropriate mechanism must be incorporated into the model to prevent undesirable results from the analysis of diffuse sound fields.

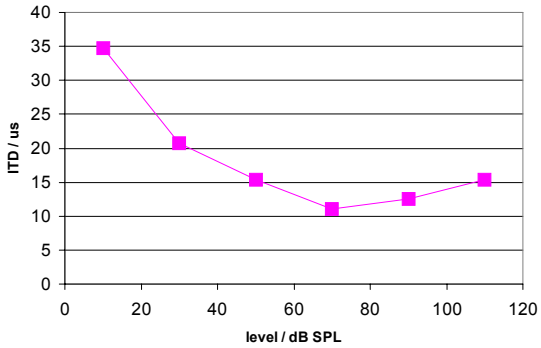


Figure 7.13: variation in minimum detectable ITD with source intensity

To solve these three problems, a confidence value is attached to each calculated ITD. This confidence value is used to weight the difference detection process, as described in the next section. A confidence value of zero is attached to ITD estimates generated during silent intervals, and a confidence value of one is attached to ITD estimates for sounds over 70 dB. The confidence value is calculated from the height of the oversampled peak in the time-delay domain. Thus, low confidence values are also attached to ITD estimates for diffuse sources, which yield lower peaks.

Defining a transform from peak amplitude to confidence value is not trivial, since the confidence value is simulating three separate psychoacoustic phenomena. To simulate all three effects correctly, three separate mechanisms would be required, but without appropriate data this is impossible at the present time. Hence, a single confidence value is used as a compromise.

The data in Figure 7.13 from [Zwislocki and Feldman, 1956] is used to calibrate the confidence value above absolute threshold, as explained in Appendix G. The transform from peak amplitude to confidence value consists of two psychometric functions. Psychometric functions define the probability of detection near threshold in a variety of psychoacoustic tasks. Two such functions are required because two break-points are found in the data: a slow increase in threshold with decreasing intensity, and an abrupt cut-off at the absolute threshold of hearing. This is discussed further in Appendix G.

The confidence value is given by the following equation:

$$cv(t) = \frac{1}{1+e^{-m_2(pa-T_2)}} \cdot \left(\frac{1}{4} + \left(\frac{3}{4} \cdot \frac{1}{1+e^{-m_1(pa-T_1)}}\right)\right) \tag{7-10}$$

where *pa* is the oversampled peak amplitude, $T_1 = 1$, $m_1 = 5$ (to reduce accuracy at lower intensities) and $T_2 = 0.05$, $m_2 = 100$ (to reduce the confidence value to zero at absolute threshold).

In the binaural processor, a confidence factor associated with each lateralisation judgement is calculated using the above equation.

7.4.8 Difference detection

The final stage in the binaural model is the difference detection unit. As discussed earlier, this should correctly detect the just noticeable change in ITD, and the just noticeable presence of the target in a binaural or spatial masking experiment. Webster suggests that the presence of a target sound at threshold may cause the same shift in the internal binaural correlation pattern as a minimum audible change in angle. However, his analysis shows that the former causes ten times the shift of the latter (100 μs compared to 10 μs), so the unification of these two phenomena seems unlikely.

However, Webster did not have access to time-domain analysis, or a sophisticated model of basilar membrane movement and hair cell response. To study his hypothesis further, a lateralisation task and a detection task are simulated via the model. The lateralisation task is the detection of the minimum audible angle. This is measured in [Mills, 1958] as equating to an ITD of 11μs at a frequency of 1 kHz. The detection task is the detection of a 1 kHz tone in the presence of broadband noise. The measurement of this phenomenon is described in the previous chapter. The data for subject 2, and a target angle of 30° is used here.

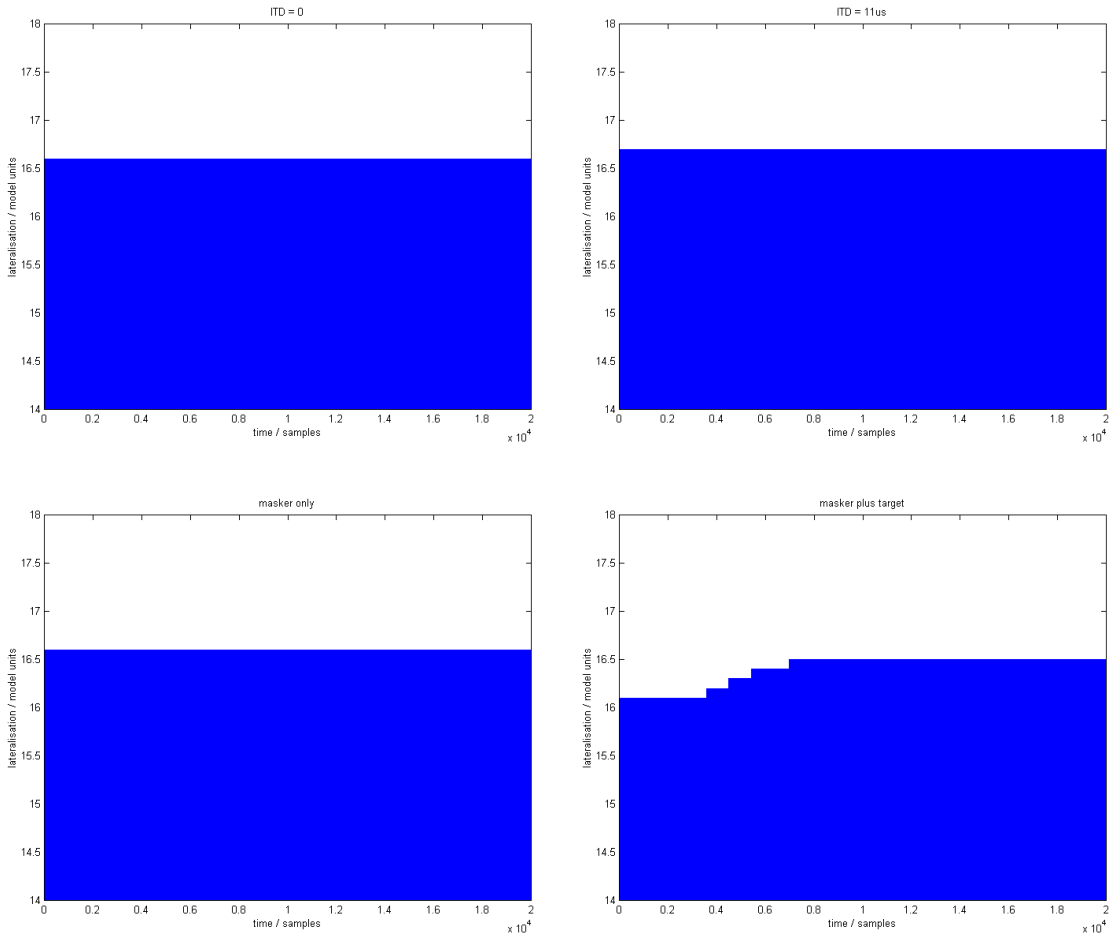


Figure 7.14: Internal signals at threshold for ITD task and masking task

(a) ITD=0; (b) ITD=11 μs; (c) masker only; (d) masker plus target

The results of processing the threshold stimuli for these two tasks through the auditory model are shown in Figure 7.14. For the lateralisation task, the calculated ITD is shown for a real ITD of 0 μs and 11 μs. For the detection task, the calculated ITD is shown for the masker only, and the masker plus target. Thus, the difference in calculated ITD due to the just noticable difference in these two tasks can be calculated.

Two things are apparent: firstly, the disparity between the two difference signals is not as large as suggested by Webster. ITD0 = 16.6, ITD11 =16.7; masker only = 16.6; masker plud target = 16.1. Hence the difference due to the masker is five times larger than that due to the ITD.

Secondly, the larger difference occurs over a shorter period of time: for the noise masking tone stimulus, the difference is only apparent near the start of the stimulus. Hence, integration of the difference signal may bring the two results into line.

Unfortunately, an integration time constant of around 500 ms is required to harmonise the two experiments, and such a large time constant will impair the performance of the model in many other tasks. However, a time constant of 100 ms will set the ITD threshold at 22 μs, and a time constant of 200 ms will reduce it to 18 μs. When the experimental error, and variance between individual subjects is considered, the difference between this value, and the experimental value of 11 μs does not seem so large. It is much closer than Webster's own value of 100 μs, and suggests that there may be a significant degree of truth in his hypothesis.

Most importantly, the difference between an 11 μs ITD threshold and an 18 μs ITD threshold is small in the intended use of this model. It represents a lateralisation accuracy error of one half of one degree. Considering no other time-domain model has successfully unified these two tasks, this is a considerable achievement. For this reason, a single difference detector will be use to detect changes in lateralisation and the presence of binaurally unmasked stimuli.

The difference detector simply subtracts one ITD estimate from the other, and multiplies the difference by the confidence values. The resulting signal is integrated using a 100 μs time constant. A binaural difference is perceived whenever this integrated value is larger than 0.3.

7.4.9 Directional accumulators

Thus, the model is calibrated to detect any just noticeable change in the location of a sound source. This change will not be detected by the monophonic model, but will be audible to a human listener. Where such a change occurs, it is visible on the binaural difference surface, pin-pointed in both time and frequency.

This accurate information is useful for identifying and solving problems in audio codecs. However, it would be useful to generate an indication of the overall effect of the codec upon the "stereo image". Such information, if it can be reliably calculated, would also be useful in other areas of audio quality assessment. The model should detect changes in the stereo image, such as a decrease (or increase) in the width of the stereo sound image, or a general shift of the image to the left or to the right.

Previous models (e.g. [Macpherson, 1991] and [Theiss and Hawksford, 1999]) can detect these phenomena, but only for a small class of artificial signals (e.g. an impulse stimulus, or white noise). These models are not suitable for assessing the stereo image degradation due to an audio codec, because the codec will perform differently with the artificial test signals compared to real music signals. In order to judge the effect of an audio codec upon a music signal, the model must reliably detected changes in the sound stage of such a signal.

The present binaural model can carry out this task. The sign of the binaural difference represents the direction in which the perceived source location has moved, thus any leftwards or rightwards shift in the stereo image can be detected in this way.

A running count is kept of the direction of movement, as judged *before* the integration of the difference signal. Thus, to compare the stereo image of two signals, the oversampled peak position (Section 7.4.7) should be compared on a sample by sample, and band by band basis. For every leftwards shift of the peak, the "left" counter is incremented. For every rightwards shift of the peak, the "right" counter is incremented. This will yield an indication of any overall shift in the image position.

For every shift of the peak towards the centre location, the "in" counter is incremented. Finally, for every shift of the peak away from the centre, the "out" counter is incremented. These final two measurements will determine if the width of the stereo image has been narrowed, or if the central image has been dispersed. These are useful measurements, because the stereo image often collapses toward the centre when intensity stereo coding is used, even though the monophonic signal energy is maintained.

Note that, even if a sound source moves rightwards, some "leftwards" movement will be detected, due to the nature of correlation-like processes (the EEVM) upon real-world periodic signals. However, the dominant movement should still be rightwards, and it is likely that the auditory system uses the dominant cue to judge source direction and movement, having learnt to ignore the minority of signals pointing in the opposite direction. If there is no dominant directional cue, then this indicates that a human listener would be unable to judge the true direction of source movement. This sometimes occurs when auditioning a continuous, high frequency pure tone stimulus in a reverberant environment, and it is appropriate if the model is also confused by such stimuli.

In summary, the binaural difference surface will indicate the time and frequency of any audible difference, and the value of the directional accumulators *left*, *right*, *in*, and *out*, will indicate any general change in the perceived sound stage. The use of these variables in the assessment of a real music signal is demonstrated in Chapter 8.

7.5 Conclusion

In this chapter, existing models of binaural interaction have been reviewed, and a novel time-domain binaural model has been developed. The model has been calibrated with measured data, and, using Webster's hypothesis, is shown to correctly predict human performance in free-field masking experiments and horizontal-plane localisation discrimination tasks. Finally, directional accumulators have been defined which allow the model to judge overall changes in the stereophonic sound stage.

8

Audio quality assessment

8.1 Overview

In this chapter, the inputs and outputs of the auditory model are discussed, and two codec assessments are performed. The model inputs consist of two "ear signals". Three methods of generating these ear signals are discussed. The model outputs consist of three time-varying frequency dependent difference signals: left, right, and binaural. Methods of processing these difference signals to reflect human perception are described. Two codec assessment tests are performed via the model, and the perception of human listeners is compared with the model's predictions. Finally, the change in the stereophonic sound stage of a real music signal is assessed by the binaural model.

8.2 Generating ear signals

The input to the monophonic model is passed through an HRTF filter and a middle ear filter, as described in Chapter 5. These filters simulate the auditory pathway from the free field to the cochlea. This simple approach is only suitable for a single monophonic source, located at the angular position matching the HRTF measurement. In order to simulate a listener auditioning a multi-channel audio system in a real listening room, a more sophisticated transformation is required.

Three transformation methods generate ear signals compatible with the binaural model. Each has its own advantages and disadvantages, which are discussed in the following sections.

8.2.1 Anechoic HRTF crosstalk processing

Consider the typical stereo listening arrangement illustrated in Figure 8.1. Two loudspeakers are placed in front of the listener, at ±30°. The direct sound pathways from each speaker to each ear are shown. All effects of the listening room are neglected.

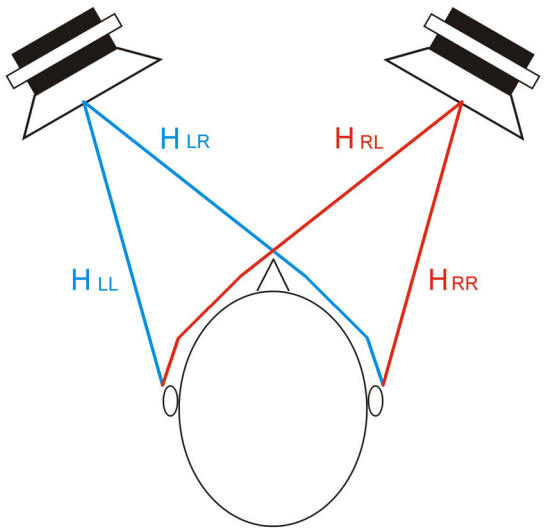


Figure 8.1: Typical stereo layout, showing paths from both speakers to both ears

Due to the (almost) symmetrical nature of the human head, the pathway from the left speaker to the left ear is identical to that from the right speaker to the right ear. Conversely, the pathway from the left speaker to the right ear is identical to that from the right speaker to the left ear. Thus, it can be shown that only two monophonic HRTF measurements are required to simulate this listening situation; one measured at 30° relative to due front, and the other measured at 330° (or –30°, depending on the nomenclature). Both measurements are taken at the entrance to the same ear for a source at the respective angular positions. If the signals emanating from the speakers are S_L and S_R, then the signals reaching each ear, E_L and E_R are:

$$E_L = S_L \otimes H_{30} + S_R \otimes H_{330} \tag{8-1}$$

$$E_R = S_R \otimes H_{30} + S_L \otimes H_{330} \tag{8-2}$$

where H_{30} is the 30° HRTF impulse response, H_{330} is the 330° HRTF impulse response, and $\otimes$ denotes convolution.

This transformation can be extended for use in other speaker layouts. The approach is used in the calibration of the binaural model in Chapter 7, to simulate the ear signals from the spatial masking experiments described in Chapter 6. Appropriate HRTFs are employed for each speaker position. This transformation is ideal for this specific case, since the listening environment is anechoic. It is also appropriate wherever a room independent measurement is required.

This approach can be improved by including the impulse response of the loudspeaker within the calculation. The above equations assume an ideal speaker, where the output equals the input. The response of the speaker can be included by redefining S_L and S_R thus:

$$S_L = x_L \otimes T \tag{8-3}$$

$$S_R = x_R \otimes T \tag{8-4}$$

where x_L and x_R are the left and right channels of the input signal, and T is the transfer function of the loudspeakers. Separate left and right speaker transfer functions can be included if the speakers are not identical.

This approach is appropriate where the listening room is anechoic, or the properties and dimensions of the listening room are not known. However, if the model is required to assess the perceived audio quality of signals replayed within a real listening room, then a more accurate transformation is possible.

8.2.2 Measured speaker/room response

It is possible to account for the response of the speakers, listening room, and listener HRTFs via a single pair of measurements. This requires the use of a dummy head containing microphone capsules within the ear canals, or a real human listener with miniature microphones inserted into their ear canals. The dummy head or real listener is placed at the normal listening location, and the impulse response from each speaker to each microphone capsule is determined.

The impulse responses should be measured in such a way that the entire room response is captured. This is the opposite of normal practice when measuring speaker responses, where the impulse response is windowed to remove all indirect sound. Capturing the entire room response represents a challenge, due to both the duration of the response, and the poor signal to noise ratio as the response decays. A typical maximum length sequence will repeat before the impulse response has decayed sufficiently, so it will be necessary to use a longer MLS, or to use a click stimulus, averaged over many measurements. This is possible with a dummy head, but the slight head movements of a human listener may render the averaging process useless, unless a head clamp is employed.

After the impulse responses have been captured, the ear signals for any audio stimulus replayed over the speakers can be generated by convolution. This technique is implemented in [Theiss, 1999].

The disadvantage of this method is that only measured speaker/room/listener combinations can be simulated. If the desired room or speaker is unavailable, then it is impossible to carry out a measurement. Further, in order to change the simulated placement of the speakers within the room, it is necessary to take further impulse measurements. If the exact parameters of the listening test are known, then this method is ideal. However, if the choice of speaker, or location of speaker and listener are variable, then this method may be too time consuming and cumbersome.

8.2.3 Simulated room response

Where the parameters of the listening environment are variable, or where the listening environment is hypothetical, *simulation* of the sound field within the listening room represents a useful alternative to direct measurement. The response of the speaker, room, and listener are still required, but each may be specified, measured, or simulated individually, and then combined as appropriate. The path of the audio signal from the speakers to the listener's ear may be calculated by the following method.

The response of the speaker may be measured or simulated. The speaker is usually approximated as a point source, but a more accurate simulation would include the directional characteristic of the speaker. This property can be quantified by measuring the on and off axis response for a variety of frequencies.

As the sound wave propagates within the listening room, it will be absorbed by the air. Higher frequencies are absorbed most rapidly, whilst frequencies below 1 kHz are not appreciably attenuated over the distances encountered within normal listening environments. A 1 kHz tone, propagating through air at room temperature is attenuated by 3 dB over a distance of 100 m. The attenuation of other frequencies over any distance can be calculated from the atmospheric absorption coefficients in [Putland, 1994]. This attenuation is in addition to the inverse square law, which states that the intensity of sound from any point source reduces as the square of the distance from the source.

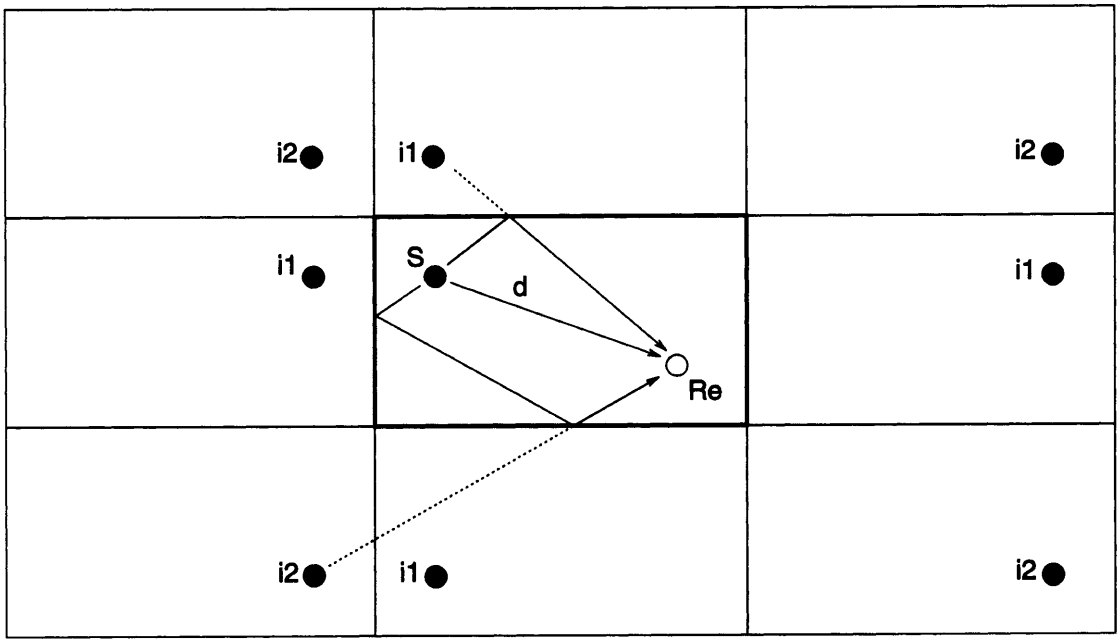


Figure 8.2: Simulation of room reflections using the image model

The boundaries of the listening room will reflect the audio signal. The image model [Allen and Berkley (1979)] uses the laws of geometric optics to model the sound paths within a shoebox shaped room. Thus, the reflections of the room boundary effectively create an image as far behind the boundary as the source is in front, as illustrated in Figure 8.2 [Rimell, 1996]. An extension of the theory [Borish (1984)] facilitates the calculation of the sound paths in a hypothetical room of any shape.

The image model assumes that the reflective surfaces in a room transform only the amplitude and direction of the signal, but not its spectrum. This is substantially untrue, and a frequency dependent extension of the model described in [Rimell (1996)] is more accurate. This model accounts for the different absorption properties of glass, plaster, curtains and other materials found within the listening environment.

Finally, the direct and reflected sound waves reach the listener. The delay and frequency dependent amplitude of the sound waves are accounted for by the above models, as are the angles of incidence. An appropriate pair of HRTF impulse responses are convolved with each incoming sound wave, and the signals for each ear are summed. Thus, the impulse responses that were measured directly in the previous section may be calculated for any arbitrary listening environment using this model.

The disadvantage of this response is that, theoretically, all possible HRTFs must be measured, and made available for the final convolution, since reflected sound can come from any direction. In practice, a limited set of HRTF measurements is used. Each incident sound wave is processed by the closest measured HRTF response, or an interpolated HRTF generated from the three nearest responses[1]. The error between the desired and measured HRTF is insignificant if a sufficiently dense set of HRTF measurements is employed. This technique is implemented in [Rimell, 1996].

8.3 Interpreting the difference surfaces

The outputs of the binaural model are three time-varying frequency dependent difference surfaces. The monophonic left and right difference surfaces indicate noticeable changes in the sound which are detectable at each ear in isolation. The binaural difference surface indicates noticeable changes in the sound which are detectable due to comparison of the signals at each ear. Both monophonic and binaural signals include information about additional signal components (e.g. coding noise or distortion) and subtractive differences (e.g. drop outs). The binaural signal also contains information about any change in the perceived location of a source.

Any value below 0.1 represents an inaudible difference; a value of 0.1 represents a “just noticeable difference”, and values above this are increasingly audible. The difference surfaces are very useful in that they pinpoint audible problems in both time and frequency. However, the surfaces are too complex to yield a simple indication of overall perceived sound quality.

There is another layer of processing within the human auditory system, the mechanisms of which are unknown at present. This processing, which is probably a conscious process, receives all the detectable differences between the two signals as an input, and calculates a single value to quantify the extent to which these differences are “objectionable” or “annoying”. If the reader has ever experienced a subjective test (or for that matter, if the reader has ever been required to give an opinion upon anything) they will be aware that the process is not mecha-

[1] Some researchers have used direct interpolation of the raw impulse responses. This is to be discouraged, since differing time delays are confused, rather than interpolated, by this technique. Interpolation of minimum phase HRTFs is more successful, provided the correct sub-sample delays are added to the resulting impulse response.

nistic, or automatic. Where a numerical value must be assigned to a subjective property, there are elements of instinct, judgement, thought, and learned response.

In the PEAQ model [Theide *et al*, 2000], a neural network is employed to transform several measures of distortion into a single value. About half of these measures of distortion are drawn from linear measurements that are not considered by the present model, while the other half arise from an auditory model comparable to the monophonic model of Chapter 5. The neural net is trained using a large corpus of subjective test results, and achieves a respectable level of accuracy, as shown in Figure 8.3. It is apparent that even this internationally recognised perceptual measurement system has trouble predicting the human perception of poor quality audio samples. This is shown by the spread of values below –3.0, where the correlation between predicted and actual diffgrade breaks down.

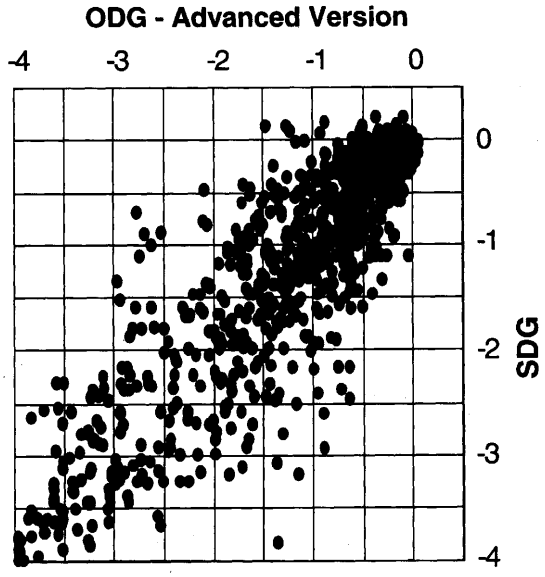


Figure 8.3: Correlation between "objective diffgrade" calculated by PEAQ, and "subjective diffgrade" judge by human listeners

It is not possible to employ a neural network with the present model at this time, because insufficient training data is available[2]. For this reason, a more direct approach is sought, which can be demonstrated to yield promising results without extensive calibration. Such an approach is described in the next section.

[2] In order to gather suitable calibration data, a subjective test was planned, and audio extracts were prepared. However, the BS.1116 recommendations [ITU-R BS.1116, 1997] suggest that inexperienced and untrained listeners are undesirable. A previous test [Goh, 2000] showed that inexperienced listeners can generate random results. The approximate time scale required to train a listener, and perform a complete set of tests is 2-4 days. It was not possible to find listeners who were willing to commit this amount of time.

8.3.1 Error Entropy

If the total area under the absolute audible difference surface is summed, this gives an indication of the magnitude of the audible difference over the entire audio extract. The main problem with this approach is that a very loud difference of short duration will have the same contribution to the total difference as a very quiet difference that persists throughout the extract. The quiet difference may be barely noticeable, whereas the loud difference may be so annoying that the listener's enjoyment is destroyed. It is inappropriate for both these coding errors to yield a similar difference score.

Another issue when summing the audible difference surface is that 5 seconds of perfect quality audio will not cancel (or even halve the effect of) 5 seconds of degraded audio. This indicates that any mean difference measure will be misleading.

It is suggested in [Hollier *et al*, 1994] and [Hollier *et al*, 1995] that a measure of the distribution of the errors over the audible difference surface can be used to give an indication of perceived quality degradation. This measure is taken from video coding, where the entropy (or activity) within a video frame is used to predict the bitrate required to code that frame. For the audible difference, a low entropy value indicates that the difference is concentrated into small regions, and a high entropy value indicates that the difference is spread evenly throughout the audible difference surface. If the audible difference surface is $e(i,j)$, then the audible difference area is given by:

$$E_a = \sum_{i=1}^{n} \sum_{j=1}^{m} | e(i,j) | \tag{8-5}$$

where there are n bands and m samples. The entropy is given by

$$E_e = \sum_{i=1}^{n} \sum_{j=1}^{m} a(i,j).\ln(a(i,j)) \tag{8-6}$$

where

$$a(i,j) = \frac{| e(i,j) |}{E_a} \tag{8-7}$$

Three possible uses of the entropy measure are given in [Hollier, 1994], though only one is demonstrated. In this solution, the error area and the error entropy are used to calculate a subjective score, thus:

$$Y = a + b.\log(E_a) + c.E_e \tag{8-8}$$

where the constants a, b, and c are chosen to match any desired subjective scale. The chosen scale is usually the mean opinion score, or diffgrade (discussed in Chapter 2). The examples in [Hollier, 1994] suggest that *b* should be approximately 1/10 the value of *c*.

8.4 Audio Quality Assessment

In Chapter 4, the results of published subjective listening tests were compared against the predictions of the Johnston model. Several inconsistencies between human perception and the predictions of the Johnston model were discovered. However, the comparison was thwarted because the codecs employed in the published listening tests were different from the codecs used to generate the samples tested by the Johnston model.

A fair performance comparison is only possible if the model "hears" the same audio as the human listeners. To achieve this goal, the exact same audio extracts must be used in both the subjective test and the models assessment.[3] It is difficult to gain access to the audio extracts used in official listening tests, and a full-scale subjective test was not carried out as part of the present work for the reasons given previously. However, the author has taken part in two subjective tests carried out over the internet. This has the disadvantage that the test conditions are difficult to control, but the advantage that many interested and experienced volunteers are available to take part. As a participant, the audio extracts and full results are available to the

[3] If the same CD track and same audio codecs are used for both tests, but the coded audio extracts are generated independently, then a fair comparison may not be possible. This is due to the inaccuracies of transferring digital audio from audio CD to hard disk. Though the numerical sample values can be transferred exactly, an indeterminate offset of several samples is usually introduced at the start of any track. Where an audio codec operates upon discrete blocks of data, a temporal offset will move the block boundaries, and may affect the quality of the coded audio. If the original audio is exchanged as a data file on a CD-ROM format disc, this problem is avoided, since the start of the file is known exactly.

author. The two tests focused on high quality (200 kbs and over) and medium quality (128 kbps) audio coding, as follows.

The audio extract employed in the high quality listening test consists of a solo hi-hat. The sharp transients of this signal make it very difficult to encode, especially by transform codecs such as MPEG-1 layer III, which temporally smear transient signals. The subjective results of this test are shown in Figure 8.4. For further details, see Appendix H. This extract is particularly testing for the model, because a previous version of the monophonic detector (Appendix J) performed poorly at the onset of sounds, where the temporal smearing is greatest.

The audio extract in the medium quality listening test is less demanding of the audio codecs. The subjective results of this test are shown in Figure 8.5. For further details, see Appendix I. Predicting the results of this test is a difficult task for the model, because different audio formats add noise in different ways, using different psychoacoustic models. It is a challenge for the model to predict the audibility of six *different* types of noise in a manner that matches human perception.

Figure 8.4 and Figure 8.5 can be found on the next page.

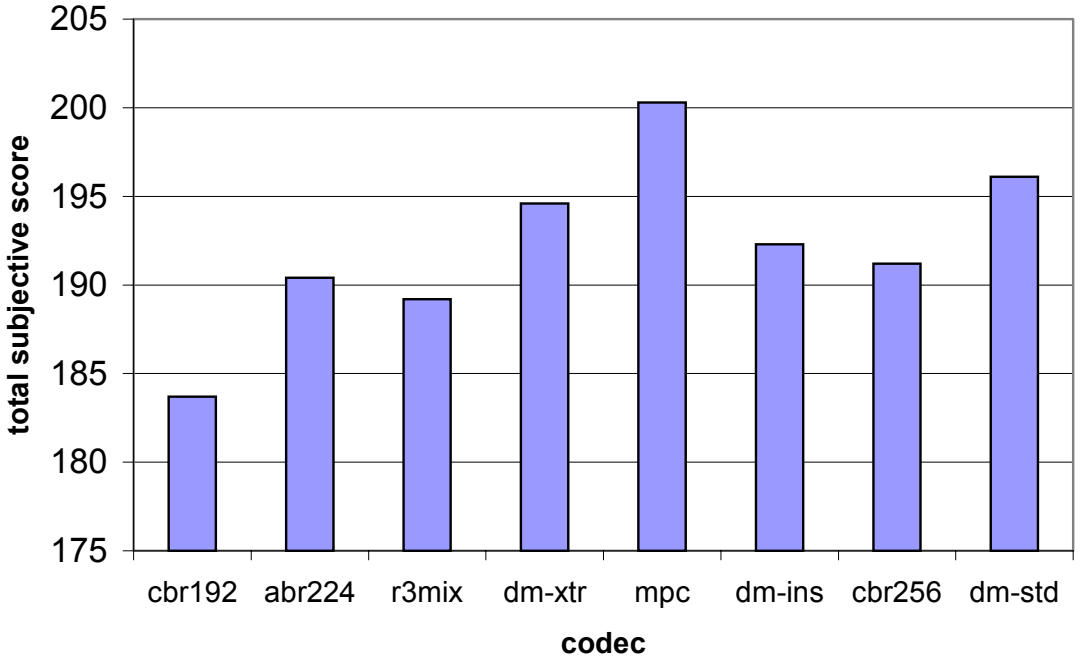


Figure 8.4: Subjective results of the high quality audio test

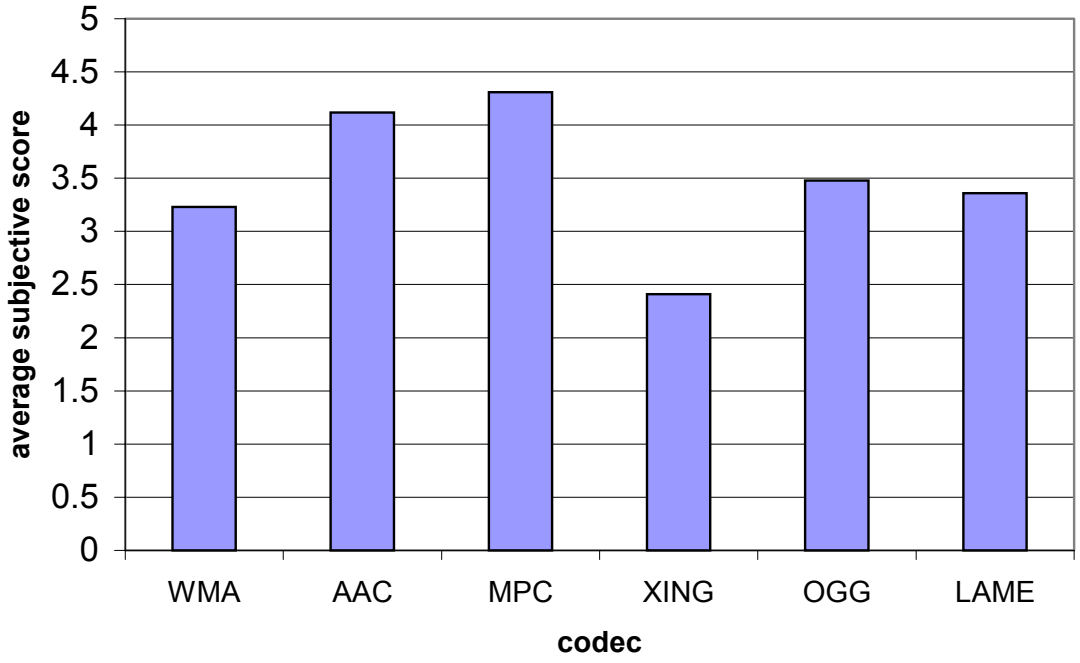


Figure 8.5: Subjective results of the medium quality audio test

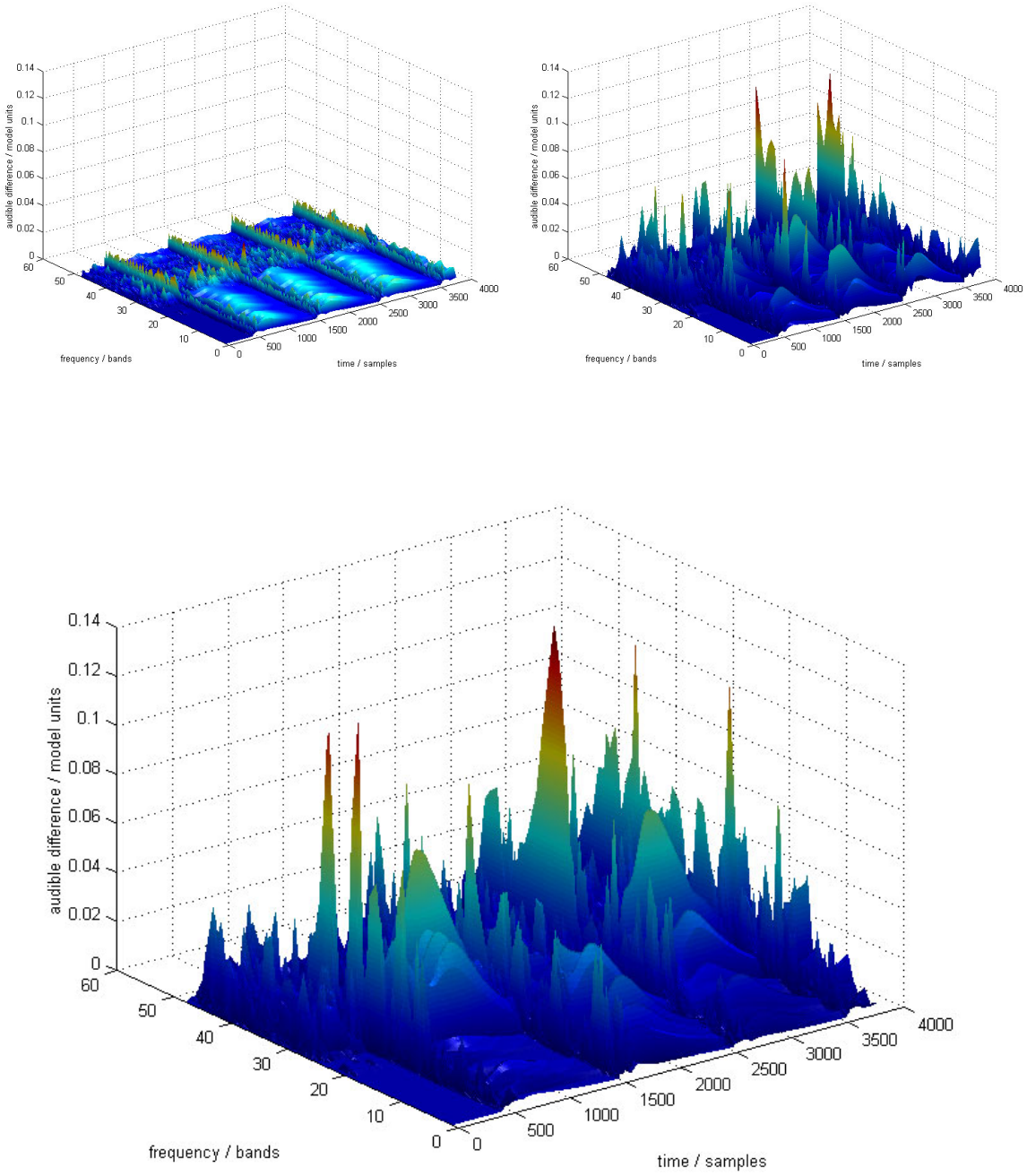


Figure 8.6: Audible difference surfaces for extracts MPC (top left), DM-X (top right) and CBR192 (main). MPC is the only transparent extract on test.

For an explanation, please see Section 8.4.1 on the next page.

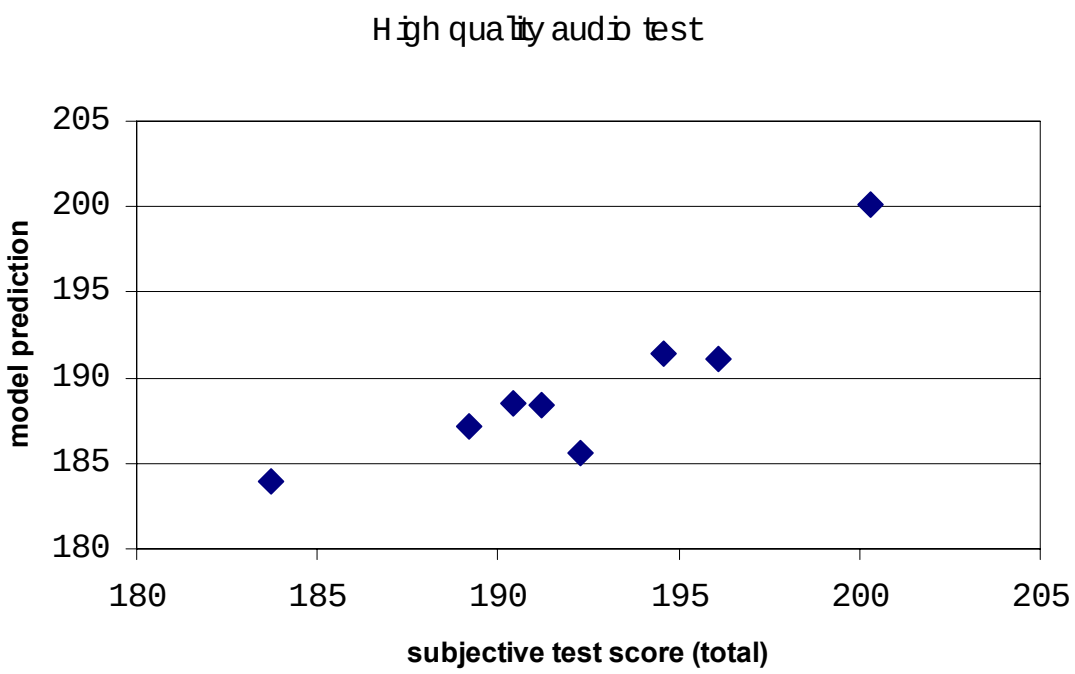


Figure 8.7: Model predictions plotted against subjective scores for high quality audio test

8.4.1 High Quality Audio Test

Each audio sample from this test is processed via the model. Difference surfaces representing the audible difference between each coded extract and the original are generated. Three of the difference surfaces are illustrated in Figure 8.6. Each audible difference surface is processed as described in Section 8.3.1. The constants *a*, *b*, and *c* in equation (8-7) are chosen to match the model output to the scores from the subjective test. This is achieved by setting $b = c / 10$ and performing a linear regression. The resulting value represents the model's prediction of the subjective quality of the codec.

The relationship between the model predictions and the results of the subjective test is shown in Figure 8.7. Judging by this graph, it is possible that the correlation is not linear. However, the range of data in this high quality test is so small that the apparent curvature may be due to random differences. For a linear fit, $R^2 = 0.8$. If the worst data point is removed, $R^2 = 0.9$. This indicates that the model is predicting the human perception of audio quality within this test to a high degree of accuracy.

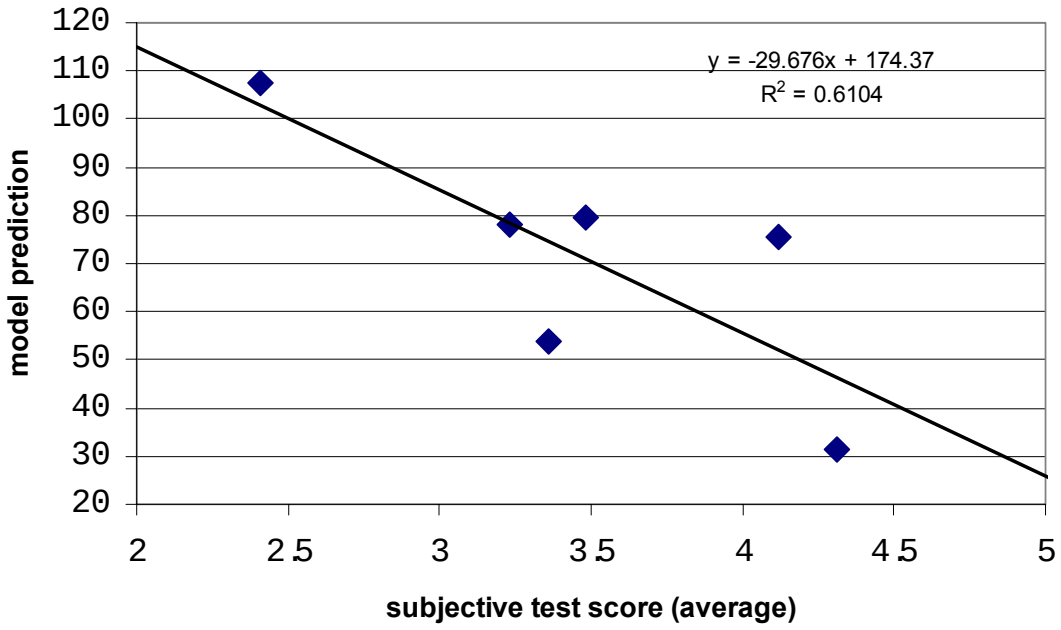


Figure 8.8: Model predictions plotted against subjective scores for medium quality audio test

The success of the model in predicting human perception of this transient stimulus demonstrates that the problems encountered in the previous monophonic difference detector (Appendix J) have been solved.

8.4.2 Medium Quality Audio Test

The same procedure was carried out as for the previous test. Unfortunately, this test revealed a severe problem with the model. The present monophonic difference detector, which has been shown to perform well in a series of psychoacoustic tests (Chapter 5) and in the assessment of transient information (previous section), fails to correctly predict the perceived differences within real music signals.

This failure was discovered too late to engineer a solution, though the reasons for the failure and possible remedies are discussed in Section 8.5.

The previous version of the monophonic difference detector, described in Appendix J, is known to be over sensitive at the onset of sounds. However, this detector has predicted human perception with some success in previous tests (also in Appendix J). For this reason, the audio

extracts from the medium quality audio test are compared via the old monophonic difference detector, yielding the correlation shown in Figure 8.8. This detector integrates the difference, thus a higher peak difference indicates poorer subjective quality. This negative correlation between peak difference and subjective quality is visible in Figure 8.8.

A linear fit provides an R^2 value of 0.6, which indicates that the model prediction and the subjective score show a "moderate" strength of correlation. It would be misleading to state that the model predicts human perception accurately, since the statistical analysis of the results in Appendix I suggests that the scores of 4.12 and 4.31 (right hand points in the graph) are significantly better than the others, whereas the model prediction does not show this.

However, the performance of the old monophonic difference detector is still promising. For example, the original calibration (Appendix J) set the threshold of audibility at a value of 25 model units. A linear regression of the data from the medium quality audio test is shown in Figure 8.8. The line of best fit intercepts the MOS value of 5.0 (transparent) at a value of 26.0 model units. This match is encouraging.

The new monophonic difference detector does not correctly predict human perception in this task. The reasons for this failure and possible remedies are discussed in the next section.

8.5 Difference detection error

Two separate monophonic difference detectors have been developed for use with the present model. The first is described in [Robinson and Hawksford, 1999] and Appendix J. The second is described in Chapter 5. The first detector correctly matched human perception in four psychoacoustic tests, and achieved some success in predicting the perceived quality of audio signals. However, the detector is overly sensitive to the onset of signals. This is a severe problem, because any (inaudible) softening or smearing at the onset of a sound generates a large calculated perceived difference. This restricts the applicability of this model to codec assessment, since some audio codecs smear the sound in the temporal domain by design (for example, WMA, as described in Appendix J).

8.5.1 Development of the new detector

To solve this problem, a new detector was developed, as described in Chapter 5. This detector is tuned to match human performance in seven psychoacoustic tests, including pre-masking.

The adaptation model [Dau *et al*, 1996a] that precedes the monophonic difference detection circuit responds vigorously at the onset of sounds. However, this feature in itself is not a fault, since the inner hair cells are known to respond in this manner. Thus, it is apparent that some of the information from the inner hair cells must be lost in internal processing. This is accounted for by internal perceptual noise.

During the development of the difference detector, it became apparent that no amount of internal noise could account for the measured pre-masking response, if the output of the adaptation model was correct. It was considered that the adaptation circuit might be in error. The adaptation model is a signal processing approximation to the hair cell response. However, accurate simulations of hair cell response (e.g. [Meddis, 1986a]) show the same vigorous activity at the onset of sounds. The design principle of the model is to match the function of the human physiology as closely as possible, and any reasonable reading of current physiological knowledge suggests that the signal at this point in the auditory system contains vigorous activity at the onset of sounds. Hence, the adaptation circuit was maintained, and another solution sought.

Internal amplitude noise, or internal uncertainty within the amplitude domain, could not account for pre-masking. However, internal uncertainty within the temporal domain *could* account for pre-masking. Further, a form of internal time-domain noise (jitter) could account for the difference in internal difference between tone masking noise and noise masking tone thresholds. This hypothesis is discussed in Chapter 5. Briefly, the "just detectable" internal difference is proportional to the time-domain variation of the hair cell firing probability. This was accounted for in the old monophonic detector by a variance measure, but it can be accounted for equally well by internal time-domain noise.

Time domain noise is physiologically possible. The signals upon the nerves are represented by pulse-density modulation, so time-domain noise is *more* likely than amplitude domain noise. The thresholds of three measured masking phenomena (temporal pre-masking, temporal post-masking, and noise masking tone) are accounted for by time domain noise. For these reasons, the mechanism is incorporated into the monophonic difference detector described in Chapter 5.

This difference detector matches human performance in seven psychoacoustic tests, and predicts human performance in an audio quality assessment task that includes a transient stimulus. Unfortunately, this detector fails in the audio quality assessment of a typical music signal. The variation within the hair cell firing probability is such that temporal jitter can be used to ac-

count for almost any difference between two signals. In other words, the jitter causes the model to be deaf to many differences between two music signals.

8.5.2 Possible solution

This problem can be solved by applying some of the knowledge gained during the development of the binaural model. The solution has not been tested, because it involves calibration of a new monophonic detector, which is not possible in the remaining time-scale of the project. However, the root of the problem does not lie within the monophonic detector, but in the module which directly precedes it.

The adaptation model of [Dau *et al*, 1996a] simulates temporal post-masking well, due to the hair cell response, but also due to the final 20 ms smoothing filter. This smoothing filter is an integral part of the model, and it probably simulates the monophonic temporal window, though this is not stated explicitly. There has been constant debate over recent years (e.g. [Oxenham, 2001]) as to whether temporal masking is due to the response of the hair cells, or a higher level temporal smearing process. [Dau *et al*, 1996a] avoid this debate by including the smoothing filter without explanation. Dau states that temporal pre-masking is due mainly to the Basilar Membrane filter response. It is shown in [Dau *et al*, 1996b] that the adaptation model underestimates temporal pre-masking thresholds, and this has been confirmed in the present work. Dau states that this may be due in part to the extremely high sensitivity of the adaptation model at the onset of signals, and this too has been confirmed by the present work. Finally, Dau states that this may be solved by limiting the strong onset response, but makes no suggestions as to how this may be achieved without changing the performance of the model in all other scenarios.

The jitter proposed in Chapter 5 solves this problem for simple psychoacoustic experiments, but not for audio quality assessment. However, the problem may be solved in an entirely different manner. Rather than ascribing temporal masking to the hair cells, let us assume that the hair cells only account for the non-linear *change* in temporal masking threshold with stimulus intensity and duration. Thus, the actual mechanism responsible for temporal masking must lie elsewhere. If [Oxenham, 2001] is correct, then there may be an internal temporal integration window. However, the experience gained in the development of the binaural model suggests that temporal windows within the auditory system are not single sided integrators, but double sided windows. If this is true for the monophonic temporal window, then the integrator at the

output of the adaptation model from [Dau *et al*, 1996a] is inappropriate. Rather, the output of the 5-stage AGC mechanism should be smoothed by a double sided window. If this window is chosen to be asymmetric (c.f. the binaural temporal integration window), then it would correctly account for pre- and post-masking.

With this approach, no internal jitter is required, so the pre-masking can be accounted for *without* compromising the sensitivity of the model to differences within real music signals. Reverse time domain filtering may be employed in order to implement this temporal window, as is used within the binaural processor of Chapter 7.

8.5.3 Discussion

It is disappointing that the model cannot assess the perceived audio quality of a real music extract using the monophonic detector described in Chapter 5. To overcome this shortcoming, it is necessary to switch to the older detector, which offers reasonable results.

However, the "fault" lies in a section of the model which was developed elsewhere. A novel solution has been suggested which arises from another part of the present work, and is supported up by the latest research [Oxenham, 2001].

Finally, it is worth noting that, just as the binaural temporal window may be a simplification of a more complex process, so the monophonic temporal integration may also approximate the underlying mechanism. This possibility has not been probed at the present time. It is reminiscent of the knowledge that was available at the turn of the last century about the nature of light. Some researchers argued that light must be a wave, others that light must be a particle. Without knowing the true answer, it was possible to explain the behaviour of light in most circumstances simply by assuming that it was a wave. Likewise, though there may or may not be simple monophonic and binaural temporal integration windows within the auditory system, assuming that such windows exist accounts for the vast majority of current knowledge, and will lead to a model that is capable of predicting human perception in a wide variety of tasks. The task of testing this hypothesis with the assessment of coded audio falls under the heading of further work.

8.6 Binaural quality assessment

The binaural accuracy of an audio system will have a relatively small effect on the overall perceived quality. This assumption will fail if the perceived location of the source is changed dramatically, but such a dramatic change will also register as a monophonic difference. Hence, the main application of the binaural difference measure is in detecting small problems in otherwise transparently coded audio extracts.

As described in Chapter 7, the binaural model can judge the general effect of an audio codec upon the stereo image. This property will be demonstrated.

8.6.1 Inter channel time delay

Two mechanisms may cause the stereo image to be shifted left or right. In the next section, inter channel amplitude difference is investigated. In this section, the effects of inter channel time delay are discussed.

Where a positive time delay is introduced into one of the stereo channels, the stereo image will shift towards the opposite channel. This also occurs for a person listening off-axis, such that the signal from one loudspeaker reaches them before that from the other. In this situation, the stereo image is “pulled towards” the nearer loudspeaker.

In order to test the binaural model’s ability to detect such a change, integer sample time delays are added to one channel of a digital stereo recording. The signal consists of a live, acoustic recording that has a wide sound stage and an expansive acoustic. The backing instruments sound distant and diffuse, while the lead solo woodwind instrument is located just left of centre. This is exactly the kind of signal which previous models have been unable to localise, due to its highly tonal nature, and the lack of abrupt signal onsets.

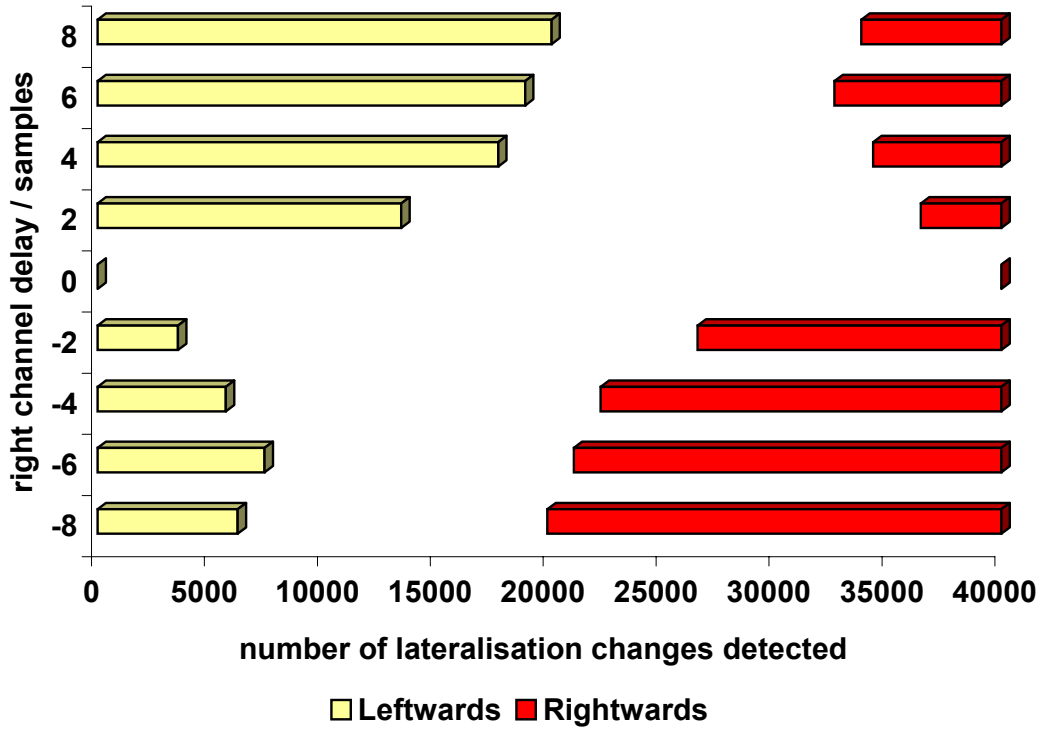


Figure 8.9: Qualitative prediction of image shift as a function of inter-channel delay

The original signal, and copies of the signal with 2, 4, 6, and 8 inter-channel delays are processed via the model. The anechoic HRTF method (Section 8.2.1) is used to generate ear signal from the channels. The monophonic model is used to generate the hair cell input to the binaural model, which in turn calculates the oversampled peak. The position of the oversampled peak for each time/frequency point of each signal is compared with the corresponding time/frequency point of the original signal. The directional accumulator values are computed for each pair. The counts of "leftwards" and "rightwards" changes in lateralisation are shown in Figure 8.9 for each version of the signal.

It can be seen that the relative number of leftwards and rightwards changes correctly predict the direction of the image shift for all the test signals, and that the number of "leftwards" changes increases as the inter-channel delay increases.

8.6.2 Inter channel intensity difference

In most modern studio recordings, the location of the sound source is determined by the inter channel intensity difference. Any change in the overall level of one of the channels will similarly shift the image location in one direction. This phenomenon is not directly incorporated into the binaural model, which only processes interaural time delay. However, when a signal containing an inter channel intensity difference is replayed over stereo speakers, the interaction of the signals from both speakers at each ear give rise to interaural time differences. The theory behind this statement can be found in [Bauer, 1961] and [Bernfield, 1973]. The result is that processing a signal containing an interaural intensity difference via the anechoic HRTF method generates ear signals that contain the correct interaural time delay.

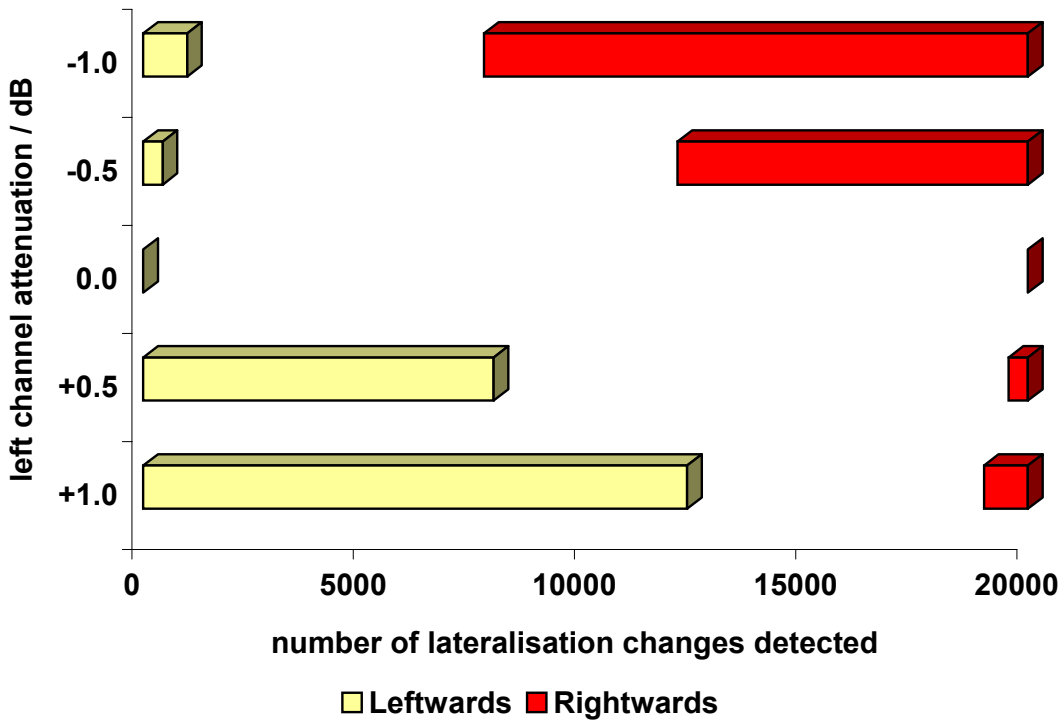


Figure 8.10: Qualitative prediction of image shift as a function of left channel attenuation

This conversion to ILD is expected from theory, but will the model detect it in practice? To test this, one channel of a stereo music recording is attenuated by various amounts. 0.5 dB inter-channel intensity difference is enough to cause a detectable change in source location[4]. The musical extract used in the previous test is also used here. The level changes applied to the right-hand channel are 0 dB, -0.5 dB, and –1.0 dB. The test methodology is identical to that used in the previous section. The number of leftwards and rightwards lateralisation shifts for each stimulus are shown in Figure 8.10, which can be found on the previous page.

Yet again, the larger directional accumulator variable for each signal indicates the perceived change in location. Also, the value of the variable increases as the amplitude difference is increased.

The change in location detected in this test appears more significant than that detected in the previous section. This matches human perception, where the change in location due to an inter channel intensity difference reproduced over loudspeakers is clearer than that generated by an inter channel time delay.

Thus, the model correctly predicts perceived changes in overall image location for a real music signal. It may be possible to calibrate the qualitative values to some quantitative indication of the change in perceived location, though this has not been attempted. Further study is required to judge the stability of the qualitative variables over different types of signal.

8.6.3 Stereo image width

Some audio codecs change the perceived width of the stereo image. Quantifying the perceived width of an image is a difficult task for a human listener. It is easy to hear differences, but quantifying those differences is more challenging.

[4] The minimum detectable monophonic change in level is 1 dB. This is yet another phenomenon that is not detected by the monophonic model, but is audible to a human listener.

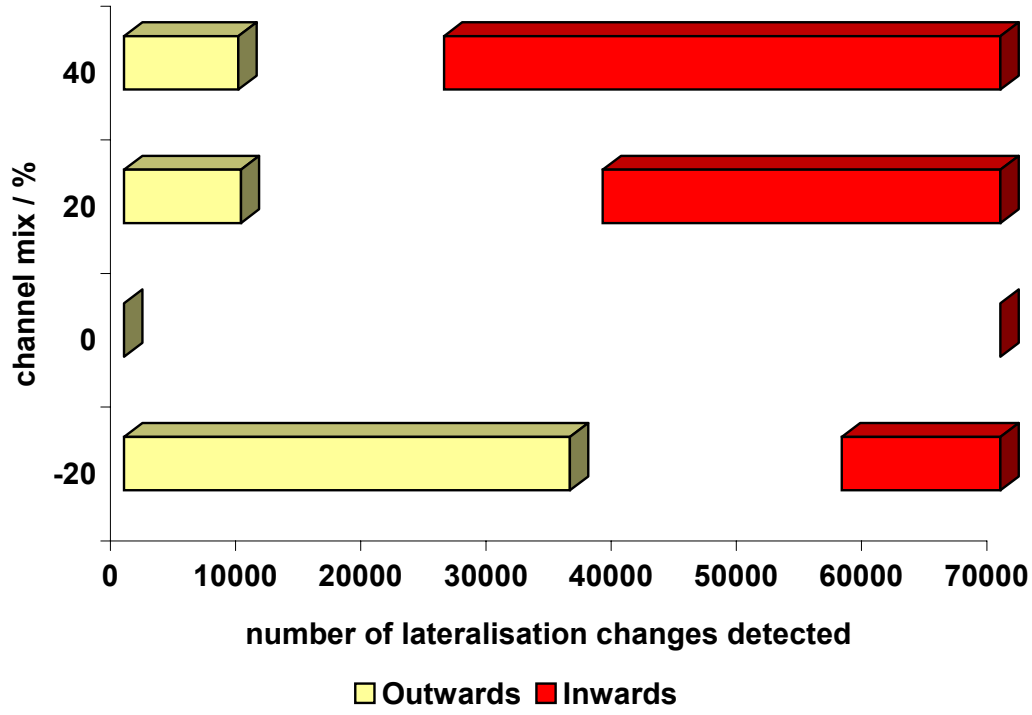


Figure 8.11: Qualitative prediction of image width as a function of channel mixing

To test the ability of the model to judge a change in image width, a stereo signal with extremely wide separation is narrowed by mixing the two channels together slightly. The signal is a twin-track recording from the 1960s which has half the instruments on one channel, the other half on the other, and only ambience in the centre. Several copies of this signal are generated with the stereo channel mixed into each other to varying degrees, and all signals are processed and compared as above. In this test, the qualitative variables "in" and "out" are compared in Figure 8.11.

This is a very difficult test for an auditory model. In order to judge that the image has collapsed inwards, it must first have detected that there are sound sources at two spatially separated positions, and then detected that both sources move inwards. However, the qualitative variables correctly predict the perceived narrowing of the stereo image as the two channels are mixed.

8.7 Conclusion

Three methods for generating suitable "ear signals" for the model have been described, and the simplest of these was used throughout this chapter with some success. A statistical analysis has been introduced that aims to predict the perception of a human listener from the complex time varying outputs of the model. Two codec assessment tests were performed via the model, and assessed using this analysis. The model correctly predicted human performance in one test, but failed in the other. The reasons for this failure were determined, and a possible solution was proposed. Finally, the binaural model was used to assess the change in the stereophonic sound stage of a real music signal. The model was shown to accurately predict human perception of interaural time and level differences, and to correctly detect a change in the width of the stereo image.

9

Conclusion

9.1 Overview

In this chapter, the present work is reviewed. The performance of the model is critically assessed, and areas for further work are outlined. Finally, the original work within this thesis is listed.

In Chapter 2, the principles of psychoacoustic-based audio codecs were introduced, and traditional objective audio quality measurements were demonstrated to be inappropriate for the quality assessment of coded audio. The expensive and time-consuming nature of subjective tests necessitates the development of objective perceptual measurements, such as the one described in this thesis.

In Chapter 3, the processes within the human auditory system were described in detail, and sources of measurable performance limits were identified.

In Chapter 4, existing models of human hearing were critically assessed. The three areas where existing models are inadequate were identified as temporal masking, spatial masking, and binaural hearing.

In Chapter 5, a time-domain monophonic auditory model was developed, based upon the processing present within the human auditory system. This model was calibrated to match human performance in a wide range of psychoacoustic tests.

In Chapter 6, the procedures and results of an investigation into spatial masking were described.

In Chapter 7, a time-domain binaural auditory model was developed. This model used the monophonic model as a pre-processor, and together the two models form a combined predictor of human perception. The model was shown to predict binaural masking and localisation data, which suggested that both phenomena were due to the same internal process. The model was calibrated to match human performance using data from the spatial masking experiment.

In Chapter 8, the model was used for audio quality assessment. The monophonic model was found to correctly predict the human perception of some coded audio extracts, but not all. Solutions to this problem were suggested. The binaural model was shown to correctly predict human perception of spatial artefacts.

9.2 Critical assessment

The monophonic and binaural sections of the model have met with differing levels of success. Each section will be discussed in turn.

9.2.1 Monophonic model

The performance of the monophonic model is disappointing. Possible solutions to the explicit problems demonstrated in Chapter 8 were described therein. It should be noted that the solution to the problem arises from the binaural work, and is supported by recently published psychoacoustic research.

The decision was taken at the start of the project to develop a novel monophonic model based upon the processing within the human auditory system. This approach is often used in models designed to predict human performance in psychoacoustic experiments, but not in the field of audio quality assessment. Had the author chosen to implement an existing monophonic quality assessment model, then its performance would have been guaranteed. However, by employing a new technique in this field, several lessons were learned.

Firstly, the computational burden due to a complex processing model should not be underestimated. Whilst increased complexity may be acceptable if it delivers a corresponding increase in performance, the increases are not proportionate.

Secondly, the accurate simulation of the processes within the human auditory system may impede the accurate prediction of human perception. This statement is superficially contradictory, but it arises because our knowledge of the human auditory system is incomplete. It is possible to generate an accurate representation of the signal at one point within the human auditory system, but to have no knowledge of the subsequent processing. Hence, at the point where this signal must be interpreted, present knowledge of the human auditory system ends. A self-training neural network is one solution to this problem, but not transforming the audio into this representation in the first place is another attractive solution.

Comparing the neural signals to detect the just audible difference between two stimuli is successful. This was demonstrated in Chapter 5 by the performance of the monophonic model in psychoacoustic tests. However, processing the neural signals in order to quantify a large perceived difference is a more challenging task. One possible solution is to use the model presented herein to determine if there is any difference between two audio signals, and to use a conventional model of audio quality assessment to quantify that difference. This may yield increased performance in the assessment of near transparent codecs. In truth, such a method may add little to the prediction of diffgrades, but it may detect the small differences that "audiophiles" claim to hear, which may not be audible within repetitive subjective tests due to listener fatigue.

Thirdly, a system cannot perform more accurately than its weakest link. This was demonstrated during the development of a perceptual loudness meter, based upon the monophonic model. This loudness meter is used in the Replay Gain Proposal, reproduced in Appendix K. By combining the processing of the monophonic model with the model of [Moore, 1997], a very accurate prediction of instantaneous perceived loudness may be computed. However, for a real music signal, the perceived loudness changes dynamically over time. The manner in which the auditory system integrates instantaneous perceived loudness to yield an overall loudness measure is not well understood. [Zwicker, 1984] suggests that a high percentile value of the ordered instantaneous loudness estimates may yield a good prediction of overall perceived loudness, and this approach works well in practice. However, over the entire length of a typical piece of music, this calculation makes the accuracy of the preceding processing redundant. This final stage is so approximate that the overall performance is only slightly reduced if most of the preceding processing is removed. This reduces the computational burden dramatically.

A similar situation arises within the monophonic model. The adaptation circuit calculates the response of the inner hair cells to a high degree of accuracy, and this signal can be used to predict human performance within psychoacoustic experiments. However, the predictions of the human perception of real music are less accurate. Hence, if this is the intended use of the model, it may be appropriate to use a simpler hair cell simulation in order to increase computational efficiency, without compromising performance. If this suggestion is added to the previous one, it is possible that removing the hair cell simulation would actually improve the model's performance in audio quality assessment, although its performance in psychoacoustic tests would be degraded.

Finally, though the monophonic model does not perform as well as intended in audio quality assessment tasks, it does form a vital pre-processor for the binaural model. The accuracy of the monophonic model in simulating the auditory nerve signal is crucial to the excellent performance of the binaural model. The success of the binaural model demonstrates that, in some respects, the monophonic model is performing acceptably.

9.2.2 Binaural model

The performance of the binaural model is very encouraging. The model successfully predicts the perceived lateralisation of a sound source in a time-domain manner. Further, the model correctly detects just noticeable changes in source location, and just detectable binaurally unmasked sounds, using the same detection mechanism.

Previous models either did not operate in the time domain, or were only applicable to a small number of artificial stimuli. The model described in this thesis can accurately judge the perceived spatial qualities of real music stimuli, as demonstrated in Chapter 8.

9.3 Further work

Appropriate solutions to the problems encountered in the monophonic model were discussed in Chapter 8, and these may be implemented to improve the functionality of this part of the model.

The binaural model can detect and judge a change in source location. This is all that is required in order to detect any degradation in the spatial properties of an audio signal due to an audio codec. However, the binaural model could form the basis of a localisation model, which not only detects changes in location, but can also predict the absolute location of a sound source.

The mapping of the lateralisation estimate into a perceived source angle is trivial, but two further modifications are recommended in order to accurately predict the perceived location of a sound source.

Firstly, the role of head movements in localisation is vital. [Theiss and Hawksford, 1999] suggest that localisation estimates should be made for three positions of the head; facing straight forward, and at ±5°. This is appropriate, even for a listener who is apparently motionless, since even small head movements yield important binaural clues. This was demonstrated in Chapter 6. Forming a localisation judgement based upon the three head orientations will remove much of the front/bank confusion that arises from "head clamped" listening.

Secondly, though the model operates well without an interaural intensity difference detector, this quantity is important in determining the perceived location. It is suggested in [Brungart *et al*, 1999] that comparison of the ITD and IID cues may yield a prediction of perceived source distance. Thus, by combining head movement, ITD, and IID in an appropriate manner, the model may predict the true 3-D location of the sound source, not just its angular bearing.

The mapping of ITD to perceived location must be frequency dependent. This will account for the dominant components of the measured HRTF response. If a more accurate method of processing the HRTF cues is required, then it is suggested that a neural network may be employed, as suggested in [Nandy and BenArie, 1996].

Finally, the binaural model's directional accumulators give an indication of the overall change in perceived source location, as demonstrated in Chapter 8. The relationship between source movement and the values of the variables is somewhat signal dependent, but it would be interesting to attempt to calibrate these variables to real world positions. Alternatively, it may be possible to transform the directional accumulator values into a single MOS-type indicator of perceived binaural impairment.

9.4 Original work

To conclude, the original work within this thesis is as follows:

Chapter 5: The implementation of amplitude dependent filter bank.

The use of an accurate inner hair cell simulation within an audio quality assessment model.

Chapter 6: The development of the psychoacoustic experimental tool *audiotest.*

The investigation into spatial masking.

Chapter 7: The entire binaural model, especially:

The use of an Element by Element Vector Multiplication as an approximation to internal binaural processing.

The use of a double-sided exponential binaural temporal window.

The use of a time-reversed double IIR filter as an efficient implementation of double sided exponential window.

The unification of detection and lateraisation in accordance with Webster's hypothesis.

Chapter 8: The ability of the binaural model to correctly perceive changes in the stereo soundstage of real music signals.

Acknowledgements

Many thanks to all of the people who have given advice, guidance, knowledge, support and friendship during this research. Special thanks to Woon Seng Gang, Martin Reed, Malcolm Hawksford, and my loving and supportive wife, Rebecca Robinson.

Thanks also to Steven Huxford, Steven Sumpter, Kelvin Foo, and Andrew Rimell, who subjected themselves to aural torture in the name of research.

The people who have worked in the Audio Research Lab during my time here deserve a special mention. Some helped, some distracted, but all were good company. With apologies to anyone that I have missed out, thanks to Richard Greenfield, Ian Speake, Tan Chew Keong, Goh Yi Wui, Beny, So Ching, Christian Kanakis (aka C. Linton-Frost), Raoul Glatt, Vincent Borel, Antonis Gougoulis, Daisuke Koya, Matthew Williams, Paul French, and Godwin Bainbridge.

Finally, I must give credit to those friends who could only distract me from working. Stuart Wilkinson, Simon James, Katie Shore, Emma Taylor, Becky Skinner, Richard Potter, Stephen Lucas, Karen Miller, Simon Lucas, Caz Jackson, and Fozzie. Life was always work *and* play, and both have been great.

A

PSYCHOACOUSTIC MODELS AND NON-LINEAR HUMAN HEARING

This appendix contains an investigation of in-ear distortion, published by the author during the present research.

In ear distortion is a special instance of in-ear generated sound. The production of extra signal components within the ear may theoretically alter the perceived performance of psychoacoustic-based codecs. In this appendix, the loudest in-ear distortion product, known as the *cubic distortion tone*, is investigated in detail. It is shown that this measurable in-ear distortion has little impact upon the perceived quality of audio codecs.

The original reference for this paper is

Robinson, D. J. M., and Hawksford, M. O. J. (**2000**). "Psychoacoustic models and non-linear human hearing."
preprint 5228, presented at the *109th Convention of the Audio Engineering Society, Los Angeles*, September 2000.

The paper is reproduced here in its entirety.

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: djmrob@essex.ac.uk, 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.

A.1 Introduction

Perceptual audio codecs aim to discard signal components that are inaudible to human listeners. Typical codecs (e.g. [ISO/IEC 11172-3, 1993]) 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.

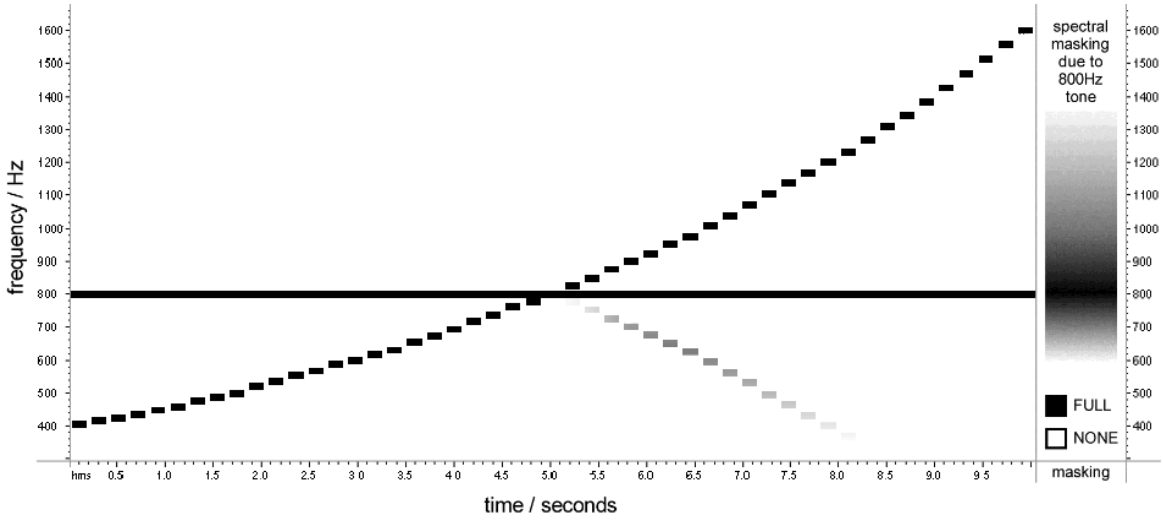


Figure A.1: Masking of stepped tones due to 800Hz tone, and resulting cubic difference tones

This concept first came to the authors' attention through the following example. The signal illustrated in Figure A.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 A.1. They are not present in the actual signal, but are generated by distortion within the human auditory system.

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.

A.2 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

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

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 [Helmholtz, 1870] 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 (A-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 [Probst *et al*, 1991]. 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 [Robinson and Hawksford, 1999] and [Yates, 1995] 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 turn.

A.2.1 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 it was discovered 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 (see [Kemp, 1979]). 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 A.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.

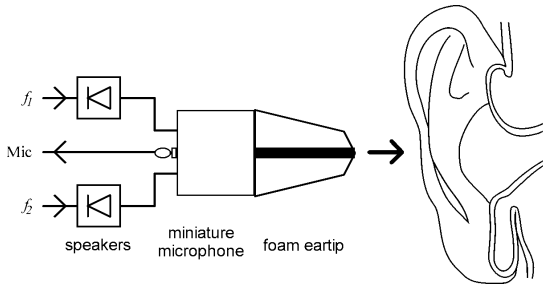


Figure A.2: Apparatus used for the measurement of DPOAEs

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. [Popelka *et al*, 1993]). 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 [Kemp, 1980] 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.

A.2.2 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 A.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 [Smoorenburg, 1972] will be referred to later in this paper.

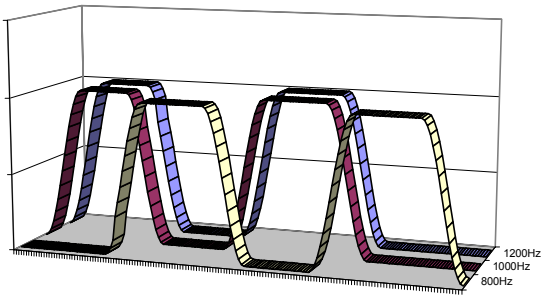


Figure A.3: Measurement of CDT level by pulsing primary tones and probe tone alternately

A.2.3 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. [Goldstein, 1967], [Hall, 1972] and [Smoorenburg , 1972]. Several important 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 [Zwicker and Fastl, 1973].
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 [Zwicker, 1981]. 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 [Giguère *et al*, 1997].

Thus, a wide range of data is 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.

A.3 Modelling the CDT

As mentioned previously, there exists a formula [Zwicker, 1981] 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 [Zwicker, 1981]. The frequency dependent data were taken from [Goldstein, 1967], cancellation-tone results only.

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

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

Where

$$\Delta z = z_2 - z_1 \quad \text{(A-3)}$$

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

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

This equation is taken from [Zwicker and Terhardt, 1980], and was used for the following graphs, for consistency with [Zwicker, 1981]. A more accurate (and more computationally efficient) formula can be found in [Traunmüller, 1990] 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 (A-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 for which human response data was available, however, the equation behaves well outside this range, and gives realistic values (though for levels 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 [Zwicker, 1979], is given by

$$L_{MIN} = \frac{L_1}{2} - 15 \tag{A-5}$$

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 (A-2) may be replaced, thus:

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

The one limit to both equations (A-2) and (A-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 (A-3).

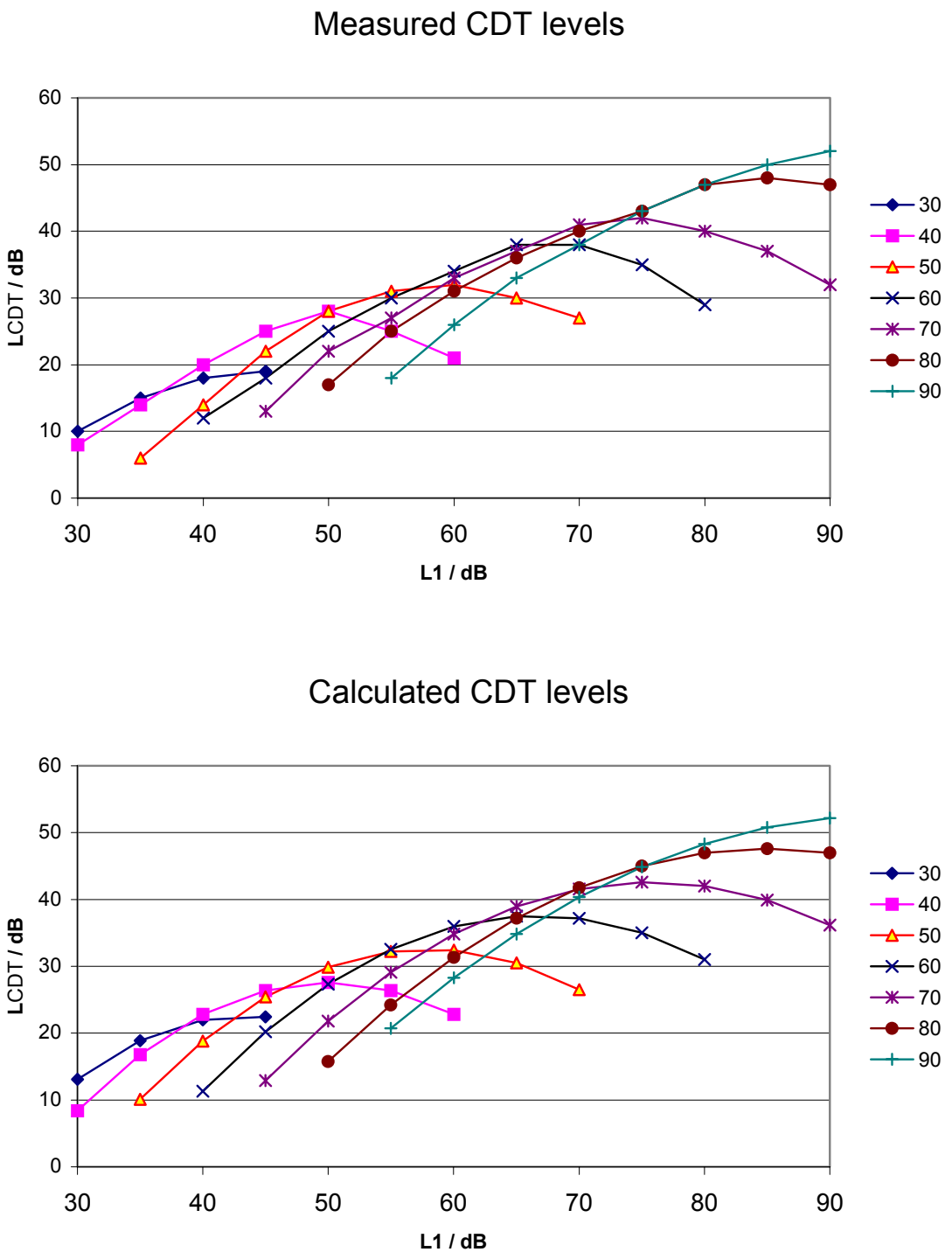


Figure A.4: Variation in L_{CDT} with L_1 and L_2 primary tone levels

upper (a) – levels measured via cancellation method from human subjects [Zwicker, 1979];
lower (b) – levels calculated using equation (A-2).

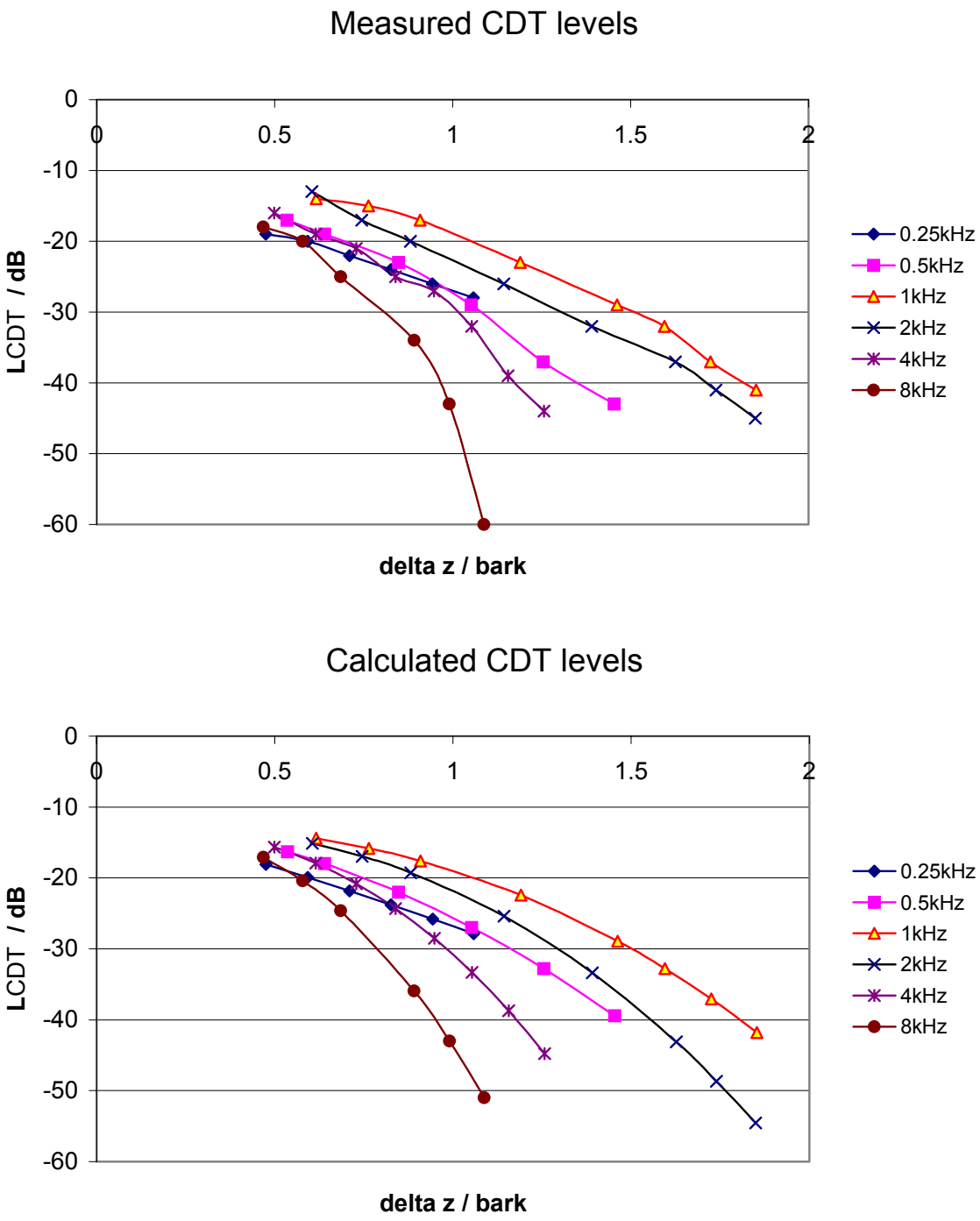


Figure A.5: Variation in L_{CDT} with f_1 and Δz

plot (a) – levels measured via cancellation method from human subjects [Goldstein, 1967];
plot (b) – levels calculated using equations (A-2)-(A-6)

Figure A.4 and Figure A.5 show a comparison of the CDT levels measured from human subjects and the CDT levels predicted by equations (A-2)-(A-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.

A.4 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.

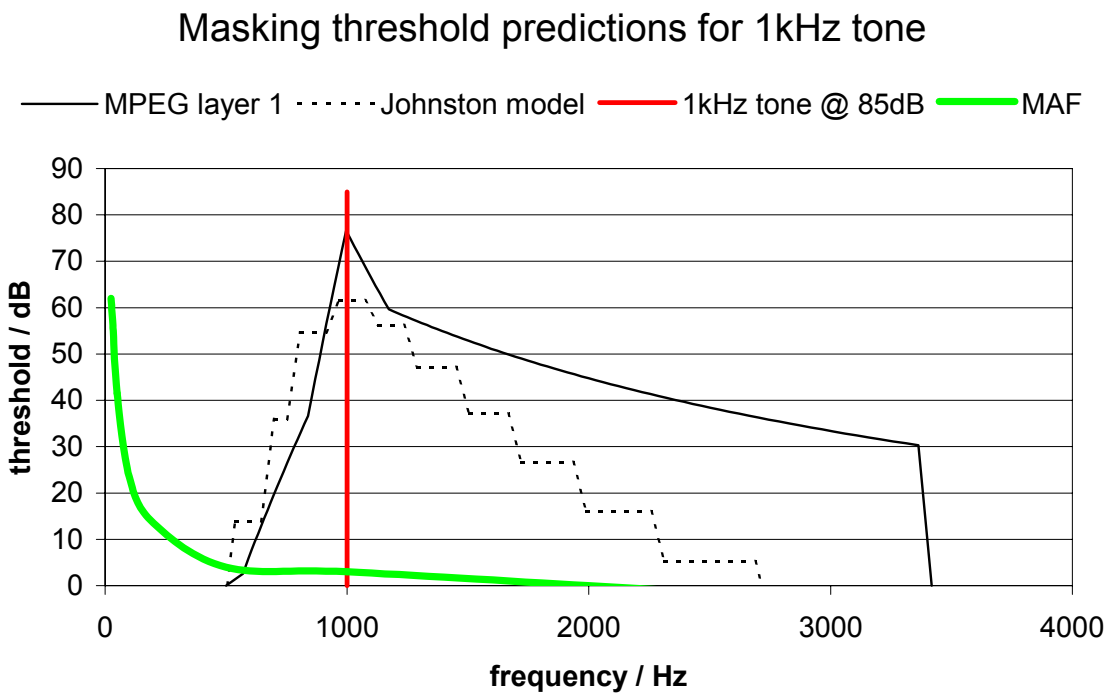


Figure A.6: Masking threshold predictions for 1 kHz tone

A.4.1 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 A.6 shows the masking threshold of the 1 kHz tone as predicted by two psychoacoustic models: The classic Johnston model [Johnston, 1988b] and the MPEG-1 Psychoacoustic model I [ISO/IEC 11172-3, 1993]-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.

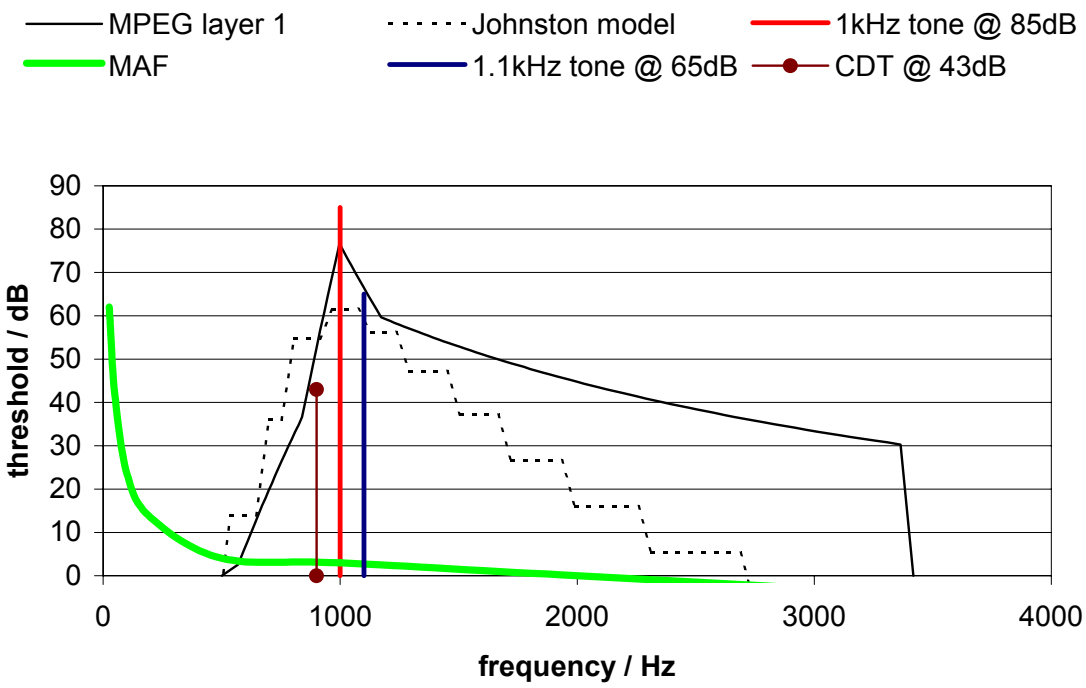


Figure A.7: Cubic Distortion Tone at masked threshold for $L_1 > L_2$

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 A.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 (A-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.

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 psy-

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

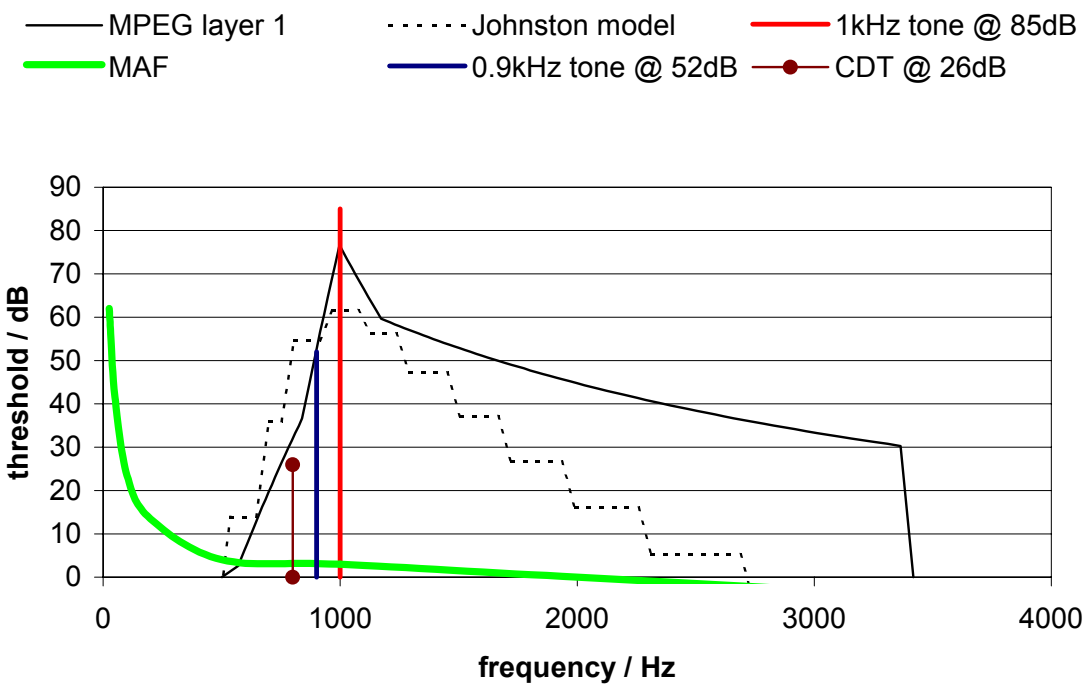
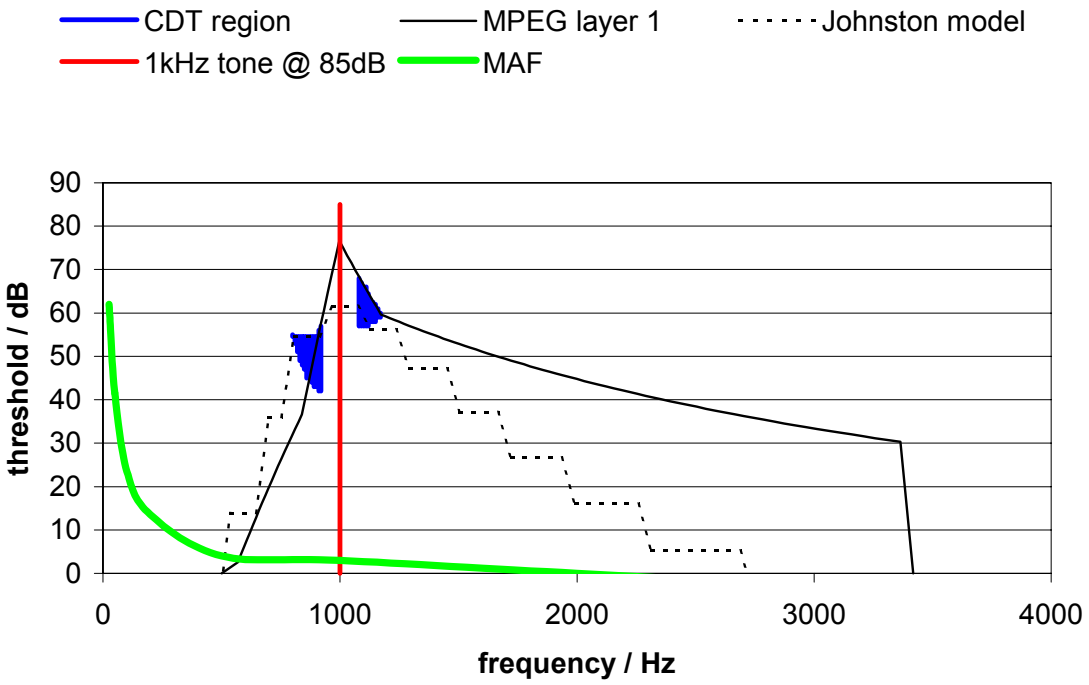


Figure A.8: Cubic Distortion Tone at masked threshold for $L_1<L_2$

choacoustic 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 A.8. However, it is known from the formulae outlined in Section A.3, and from actual experimental data, that this CDT is audible.

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.



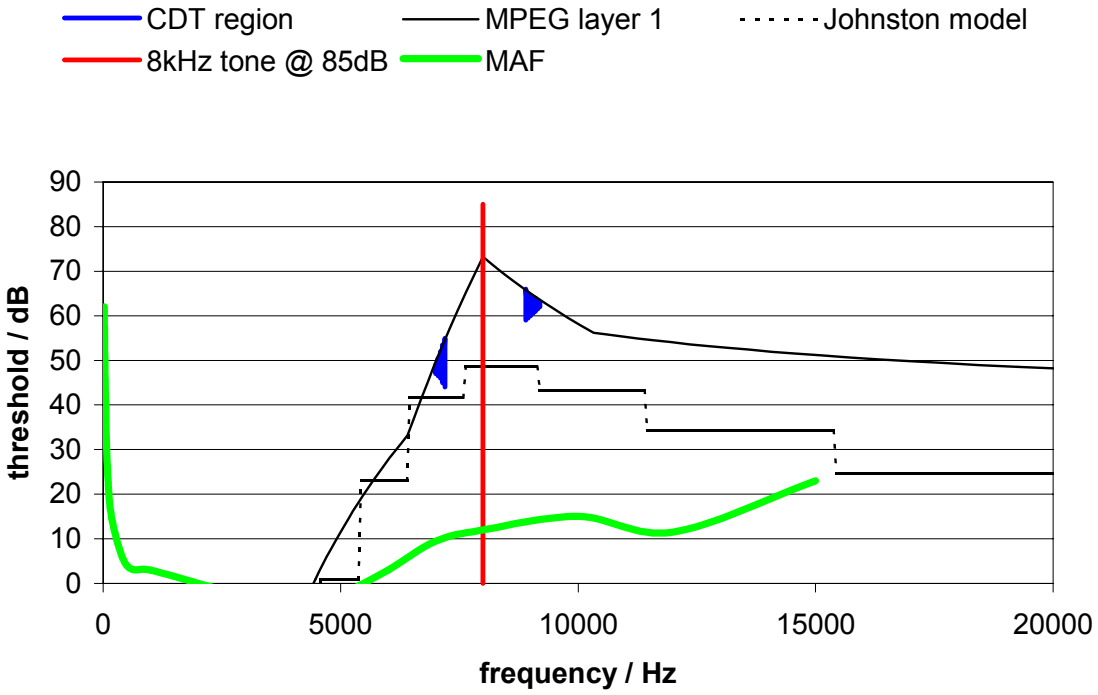


Figure A.9: Region where CDT unmasks f_2

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

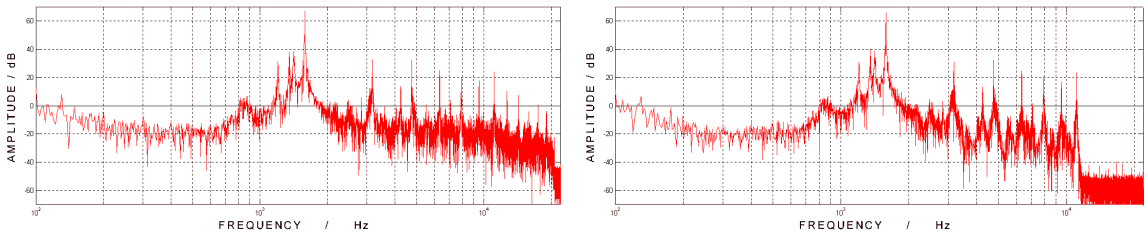


Figure A.10: Spectrum of recorder note: (a) original, (b) coded mp3 @ 96kbps

A.4.1.1 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 [Probst *et al*, 1991]. 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 unmasked f_2 is always within 15 dB of the masking threshold, it is likely that the codec may allocate some 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 A.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 A.4 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 A.11. [Robinson and Hawksford, 1999] 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.

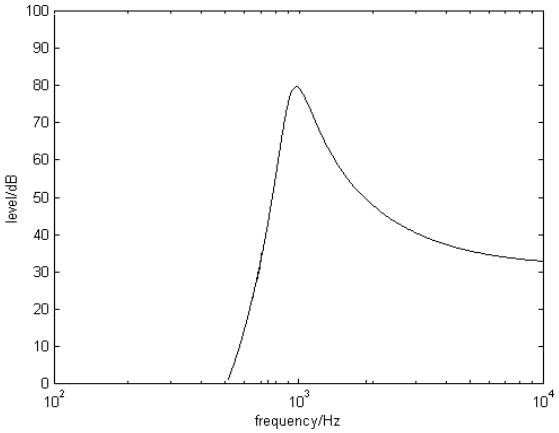


Figure A.11: Excitation within human auditory system due to 1kHz tone

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.

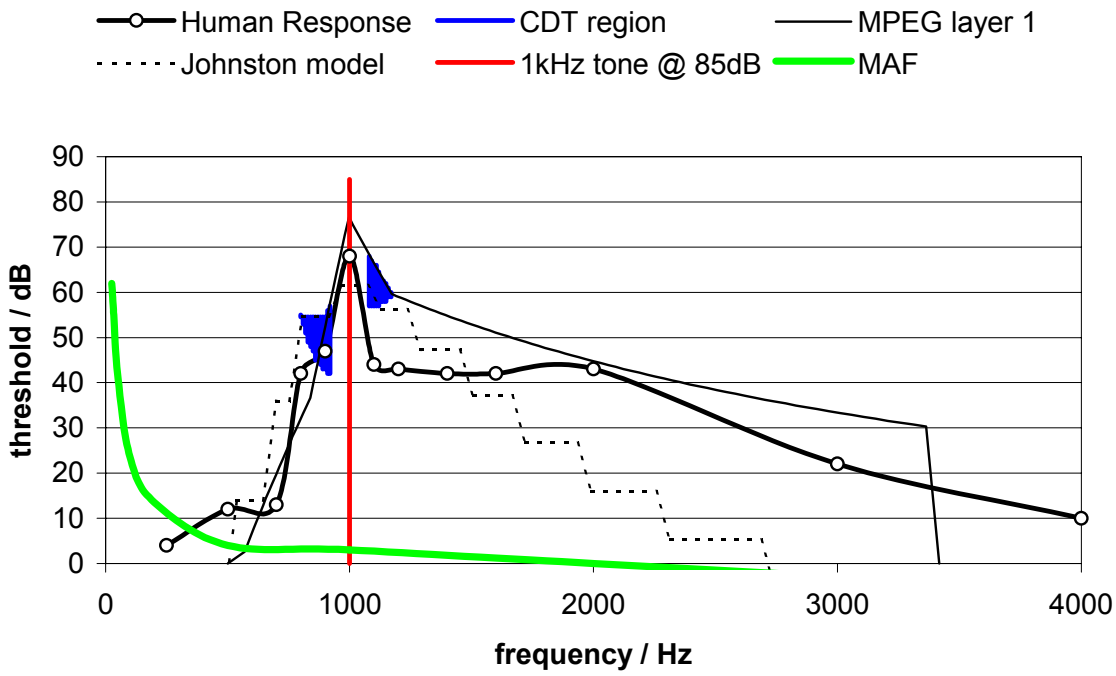


Figure A.12: Masking threshold of human subject

A.4.1.2 Effect of Beats

Figure A.12 shows the human masking thresholds measured using a test of the second type [Moore *et al*, 1998], 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 A.2.3, 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 [Giguère *et al*, 1997]). However, [Moore *et al*, 1998] 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

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

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.

A.4.2 Masking due to CDT

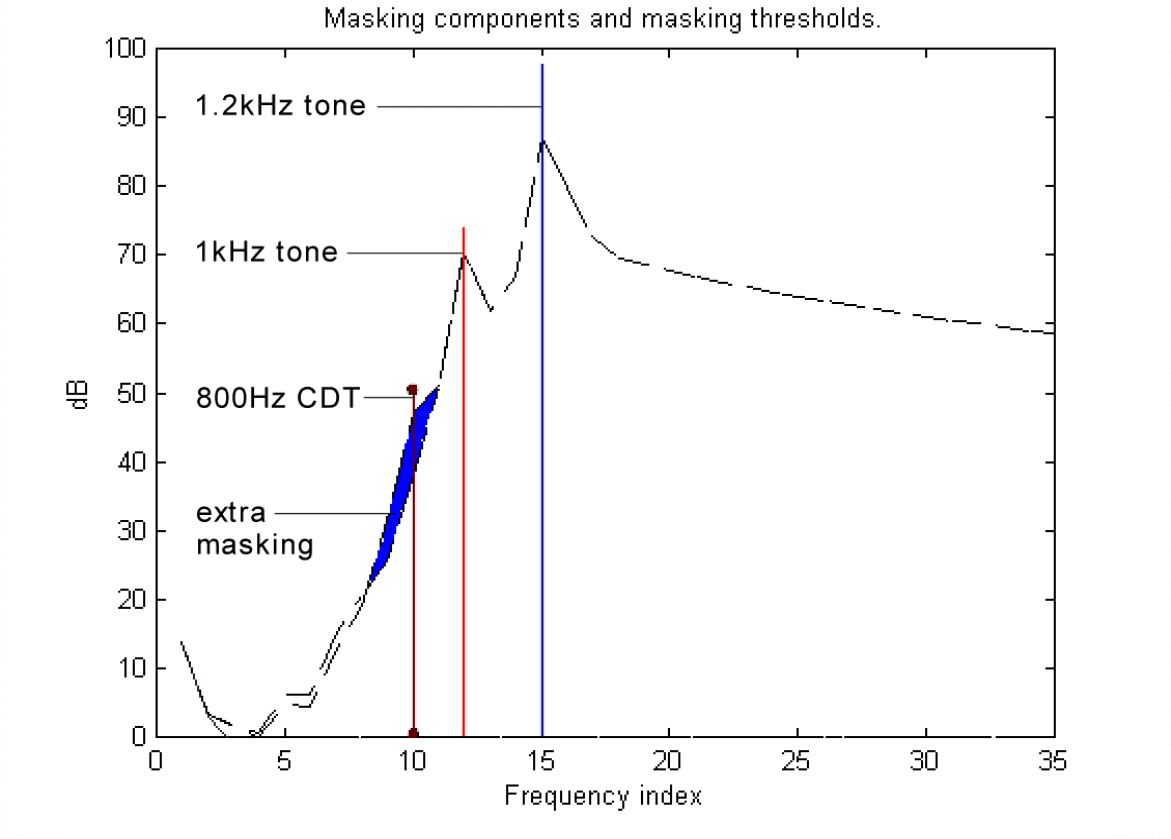


Figure A.13: Extra region of masking (shaded) provided by the CDT

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.

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 **(A-1)**-(A-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 A.13). As the cancellation-tone measured CDT level accounts for the masking due to the CDT to within 2 dB (see Section A.2.3), this is a significant result.

A.4.2.1 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

[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.

1.05 kHz; the auditory filter at 100 Hz will not be excited, hence there will be no resultant masking[3].

A.4.2.2 Real World Applications

In the Section A.4.2, 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.

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 (A-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 A.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.

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

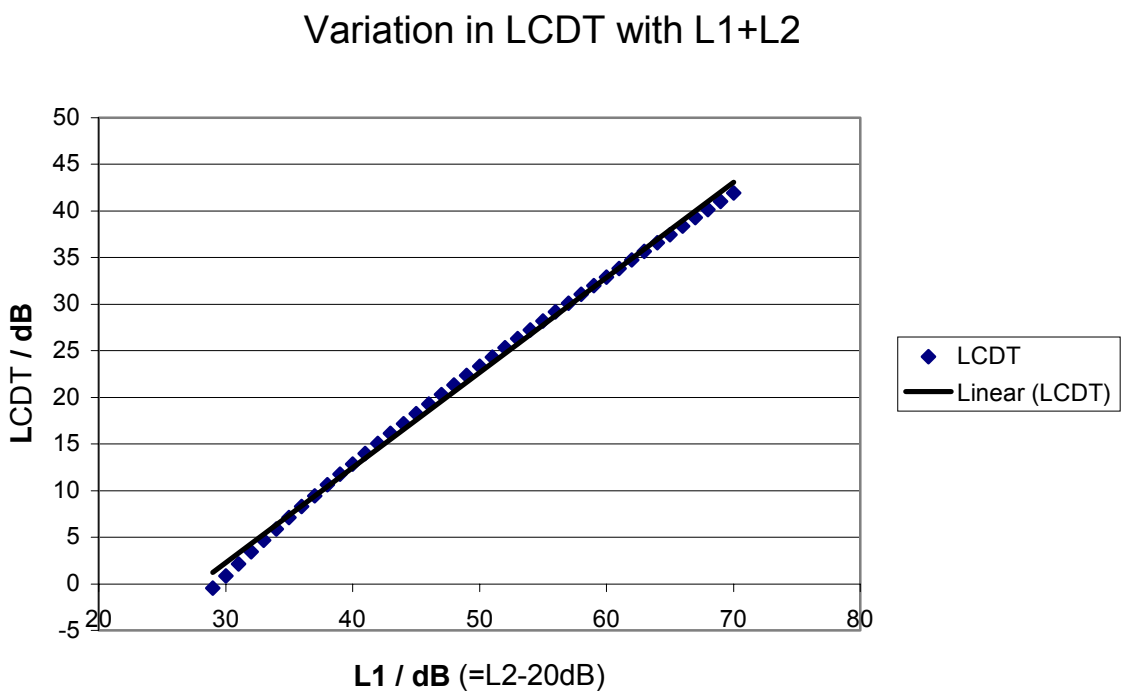


Figure A.14: Showing how L_{CDT} changes linearly as L_1 and L_2 are varied together

A.5 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. However, 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.

B

Masking noise

Throughout this work, it is assumed that the threshold condition within a masking experiment gives rise to some just detectable internal difference.

This implies that if a subject auditions the masker in isolation, and then the masker plus target, the only audible difference between the two signals is, logically, the target. Further, the presence of the target must cause some difference within the internal representation of the signals, and this difference corresponds to a “just noticeable difference”.

It is suggested that the internal difference between the masker only and the masker plus target must be comparable to the internal difference between no sound and a sound at the absolute threshold of hearing. This is discussed extensively in [Zwicker, 1990], where it is demonstrated that this is largely true.

Unfortunately, this ignores one common feature of psychoacoustic experiments. Where a noise masker or target is employed, this noise is usually produced by an analogue noise generator. As such, the masking noise “sample” will be different for each presentation. This means that the difference between the masker, and the masker plus target will not only be due to the target, but also the variability of the random noise. Thus, for the target to be audible, it must not only exceed some internal threshold (assumed to be due to internal noise) but it must also exceed the variation from noise sample to noise sample.

Where the same noise sample is used for each presentation, it is said to be “static” or “frozen”. Such stimuli are rarely used with psychoacoustic tests, though they are becoming more common with the widespread use of PCs for signal generation. Within the present work, all noise stimuli are frozen.

The difference in threshold due frozen noise maskers and random noise maskers is usually no more than 3 dB, due to the power addition of noise. This 3 dB offset is assumed to be unimportant within the present work, since it can easily be corrected for. Further, the variation of threshold between two listeners is usually greater than 3 dB, so it can be said that this factor is insignificant.

In one test, the use of frozen noise was found to have a greater impact on the threshold. The experimental determination of the minimum audible gap in a burst of noise is used to calibrate the monophonic model. The quoted threshold value is a gap of 3 ms [Penner, 1977]. The task is to compare two bursts of noise, and to determine which burst contains the gap. For random noise, the threshold of 3 ms is correct. However, for frozen noise, the threshold is closer to 2 ms, and may be slightly less. Thus, because the noise does not change from presentation to presentation, it is significantly easier to detect the gap.

The reason that frozen maskers are used through the present work is as follows. The original version of an audio extract is always the same. It does not change from one listening session to another. Thus, a listener may learn the features of this extract, even if it is “random” noise, since the noise will sound the same every time the extract is auditioned. If the model were made insensitive to changes in random noise, then the audio codec could generate a new different extract of noise, and the model would not detect this. However, if this noise formed an important part of the audio extract, the listener would detect the change, so it is inappropriate to configure the model to ignore it.

C

Monophonic model calibration data

The table on the following page contains the calibration data for the monophonic model, and the references from which this data is taken.

Reference	Task	Masker	Target	Delay	Target Threshold
[Penner, 1977]	Gap detection	White noise 20 Hz – 20 kHz 87 dB SPL (44 dB SL)	Silence	-	3 ms[1]
[Reed and Bilger, 1973] and [Green *et ak*, 1959]	Noise masking tone	White noise 0 Hz – *fs*/2 78 dB SPL (35 dB SL)	1 kHz tone	-	45 dB
[Gehr and Sommers, 1999]	Temporal pre-masking (backwards mask-ing)	White noise LPF @ 8.5 kHz 80 dB SPL (41 dB SL) 50 ms	500 Hz tone 10 ms duration	1 ms 2 ms 4 ms 6 ms 8 ms 10 ms 20 ms	39 dB 36 dB 33 dB 31 dB 27 dB 27 dB 24 dB
[Zwicker, 1984]	Temporal post-masking (forwards masking)	White Noise 20 Hz – 20 kHz 80 dB SPL 200 ms	2 kHz tone 5 ms duration	0 ms 5 ms 10 ms 20 ms 50 ms 100 ms	75 dB 66 dB 56 dB 43 dB 31 dB 23 dB
[Moore and Alcán-tara, 1998]	Tone masking noise	1 kHz tone 85 dB SPL	Noise band 1 kHz, 80 Hz wide		45 dB SL (64 dB SPL)
	intensity discrimination	1 kHz tone 60 dB SPL	1 kHz tone 61 dB SPL	presented sequentially	1 dB intensity difference

[1] Penner uses gated continuous random noise, such that the presentation without the gap uses a different noise sample to that with the gap. In the current scenario, the two presentations use the same noise sample. This reduces the threshold to 2 ms. For further discussion of this issue, see the previous appendix.

D

Spatial masking reference data

Tonal signals shall be referred to in dB SPL. This is a measure of the energy in the signal. For a pure sine wave, this is equivalent to the peak to peak excursion divided by root two. For a single sine wave is it also equivalent to the spectrum level.

Noise signals shall be referred to in dB SPL / Hz, known as the spectrum level, SL or N_0. This is **not** the total energy of the signal, but a measure of energy per unit bandwidth. If the bandwidth of the noise is known, then energy and spectrum level are related by the following formulae:

$$\text{energy} = 10 \log_{10} \left[\left(10^{\left(\frac{\text{spectrum level}}{10} \right)} \right) \times \text{bandwidth} \right] \qquad \text{(D-1)}$$

$$\text{spectrum level} = 10 \log_{10} \left[\left(10^{\left(\frac{\text{energy}}{10} \right)} \right) \div \text{bandwidth} \right] \qquad \text{(D-2)}$$

If the signal is spectrally white, then it is assumed to have a bandwidth of 20 kHz, unless otherwise stated in the literature.

A time averaged spectral view of a tonal signal, frequency *f* Hz, energy *x* dB SPL would show a single peak at *f* Hz of *x* dB. A time averaged spectral view of a white noise signal band-limited 0-20 kHz, energy *x* dB SPL / Hz would show a flat line from 0-20 kHz at *x* dB. This may seem obvious, but by referring to spectrally flat band limited noise signals in dB SPL / Hz an insight is gained into the relative amplitudes of target and masker which is not apparent when quoting the noise power in dB SPL energy (V_{rms}^2).

To compare data from different experiments, the masking will quantified as target level minus masker level, where the target and masker levels have units of dB SPL for tone or dB SPL / Hz for noise. In the literature, this measure is sometimes referred to as S/N_o for noise masking noise, and E/N_o for tone masking noise. A positive value indicates that the target must be louder than the masker to be audible (typical in noise masking tone tasks); a negative value indicates that the target is audible when quieter than the masker (typical in tone masking noise tasks).

In the following tables, results from the spatial masking experiment in Chapter 6 are quoted first; results from the literature are quoted second. The comparable results are highlighted in **bold type**. Where the results from the literature may not be directly comparable with those from the present experiment, the differences are highlighted, and their importance is discussed.

D.1 Experiment 1

Masker:	35 dB / Hz band-limited (20 Hz – 20 kHz) white noise
Target:	1 kHz tone

Results from the spatial masking experiment in Chapter 6:

thresholds for 1 kHz tone in dB SPL =	WS:52.6, DR:45.7, SS:53.4
diff = SPL (target tone) - Spectrum Level (noise masker) =	WS:17.6, DR:**10.7**, SS:18.4

Results from the literature:

diff (quoted is E/N0 in literature) =	[Reed & Bilger, 1973]:**10.5,** [Green et al, 1959]:**10**

D.2 Experiment 2

Maker:	35 dB / Hz bandlimitted (20 Hz – 20 kHz) white noise
Target:	8 kHz tone

Results from the spatial masking experiment in Chapter 6:

thresholds for 8 kHz tone in dB SPL =	WS:59.9, DR:54.3, KF:58.3
diff = SPL (target tone) - Spectrum Level (noise masker) =	WS:24.9, DR:**19.3**, KF:23.3

Results from the literature:

diff (quoted is E/N0 in literature) =	[Reed & Bilger, 1973]:**18.2**

D.3 Experiment 3

Masker:	80dB SPL 1 kHz tone
Target:	80Hz wide 1 kHz centred noise. reference level: 80 dB SPL (energy) = 61 dB SPL / Hz (SL)

Results from the spatial masking experiment in Chapter 6:

just audible noise in dB SPL (energy) =	WS:61.5, DR:55.8, SS:54.5
just audible noise in dB SPL / Hz (SL) =	WS:42.5, DR:36.8, SS:35.5
diff = SL (noise target) – SPL (tone masker) =	**WS:-37.5, DR:-43.5, SS:-44.5**

Results from the literature [Moore *et al*, 1998]:

masking =	JA:54, BM:60, HF:56
threshold of 1 kHz tone in silence =	JA:11, BM:3, HF[1]:8
just audible noise in dB SPL (energy) = masking + threshold	JA:65, BM:63, HF: 64
just audible noise in dB SPL / Hz (SL)	JA:46, BM:44, HF:45
Masker:	85 dB SPL 1 kHz tone (5 dB louder than above)
diff = SL (noise target) – SPL (tone masker) =	**JA: -39; BM:-41; HF: -40**

This table is discussed on the following page.

[1] There is a typing error in the original document. Throughout the text, and in the caption to Figure 3, one listener is referred to as "HF". However, Table 1 (absolute thresholds) contains no data for listener HF, but does contain data for listener "HV". Since the text directly refers to listener HF as being listed in Table 1, and since the letters F and V are adjacent on a QWERTY keyboard, the author has assumed the data in Table 1 listed under HV correspond to listener HF.

In this comparison, the two experiments are not entirely equivalent. [Moore *et al*, 1998] is different from the present work in the following aspects:

- All levels were measured at the ear drum.
  - This difference will cancel when using narrow band stimuli.
- A 85 dB masker was employed (5 dB louder than in the present experiment)
  - Since the difference between the masker and target is the figure under scrutiny, this will cancel if the amount of masking scales linearly with amplitude. This is not true over a very wide amplitude range. However, over a range of 5 dB, the relationship between masker and target is quasi linear, and the non linearity may be neglected. This demonstrated in [Reed and Bilger, 1973].
- Different noise stimuli were employed in each presentation (in the present experiment, a frozen noise masker was employed)
  - This may be significant, and is probably the cause of the lower thresholds obtained in the present experiment. See Appendix B for further details.

Finally, the use of the threshold tone figure as a basis for calculating the level of the target is suspect, since the target is a noise band , not a pure tone.

D.4 Experiment 4

Masker:	80 dB 1.5kHz tone
Target:	80Hz wide 1 kHz centred noise. Reference level: 80 dB SPL (energy) = 61 dB SPL / Hz (SL)

Results from the spatial masking experiment in Chapter 6:

just audible noise in dB SPL (energy) =	WS:21.7, DR:10.9, SH:14.7
just audible noise in dB SPL / Hz (SL) =	WS:2.7, DR:-8.1, SH:-4.3
diff = SL (noise target) – SPL (tone masker) =	**WS:-77.3, DR:-88.1, SH:-84.3**

Results from the literature [Moore *et al*, 1998]:

masking =	JA:10, BM:10, HF:15
threshold of 1 kHz tone in silence =	JA:11, BM:3, HF:8
just audible noise in dB SPL (energy) =	JA:24, BM:14, HF:22
just audible noise in dB SPL / Hz (SL) =	JA:5, BM:-5, HF:3
masker =	85 dB SPL 1 kHz tone (5 dB louder above)
diff = SL (noise target) – SPL (tone masker) =	**JA:-80, BM:-90, HF:-82**

In this comparison, the two experiments are not entirely equivalent. [Moore *et al*, 1998] is different from the present work, as discussed in the previous section. In addition, [Moore *et al*, 1998] use a 600 Hz noise target with 1 kHz tone masker, as opposed to the 1 kHz noise target with 1.5 kHz tone masker employed in the present work. This is the closest experiment in the literature, but does not match the present work perfectly. This is a problem, because the amount of masking changes with frequency in a non-linear manner. All that can be said is that the two sets of figures are somewhat comparable. The similarity of the results indicates that there was no major error in the present experiment.

D.5 Experiment 5

Masker:	White noise 80 dB SPL (energy) = 37 dB SPL / Hz (SL)
Target:	2 kHz tone burst

Results from the spatial masking experiment in Chapter 6:

just audible tone in dB SPL =	WS:53.8, DR:51.2, AR:57.5
diff = SPL (tone target) – SL (noise masker) =	**WS:16.8, DR:14.2, AR:20.5**

Results from the literature [Moore *et al*, 1998]:

Just audible tone in dB SPL (median, ± one quartile) =	[Zwicker, 1984] 56, 60, 52
diff = SPL (tone target) – SL (noise masker) =	[Zwicker, 1984] **19, 23, 15**

D.6 Conclusion

The threshold values for coincident targets and maskers from the spatial masking experiment match data from the literature.

E

Why lateralisation experiments are not predicted by the model

Possibly the most common binaural hearing experiment involves the comparison of ITD and IID cues, in what is called a “trading experiment”.

When listening over headphones, if an identical stimulus is fed to each ear, then the sound source will be perceived as originating from the centre of the listener’s head. Introducing an interaural time delay will displace the perceived location of the sound towards the ear that first receives the sound. Also, an interaural *level* difference will displace the perceived location of the sound towards the ear which receives the loudest signal.

Psychoacusticians have been fascinated by the relationship between these two phenomena. In particular, they have tried to measure the effective equivalence of these two interaural quantities in the form of trading experiments. In such an experiment, a fixed interaural time delay is added to the stimulus, and the listener must determine the correct, opposing interaural level difference which will bring the perceived location of the sound source back to the centre of their head. The ILD is adjusted by the candidate, and when this has been achieved, the opposing ITD and ILD are said to be equivalent.

It was first hypothesised that the ITD and ILD were traded within the auditory system in the manner suggested by this experiment. That is, the ITD and ILD were combined, and a single localisation position was fed to the brain. Many models include just such a mechanism to com-

bine the two cues into a single lateral position. (See [Colburn and Durlach, 1978] for a review, especially section II: Count comparison models).

There are several strong indicators that the auditory system does not work in such a simplistic manner. Firstly, in trading experiments, many candidates cannot "bring the source back to the centre of their heads" with an opposing ILD – rather, the source splits into two, or become diffuse. If two opposing cues can cause the perception of two separate sources, it is unlikely that the two cues are unified into a single lateral position before they are "perceived". Secondly, even moderate ITDs cancelled by small ILDs do *not* sound the same as the original monophonic signal. Finally, it is quite likely that the comparison of ITD and ILD acts as a distance cue for sound sources near the head. This is possible because a source travelling along a constant bearing gives rise to a constant ITD, whereas the ILD increases dramatically as the source nears the listeners head. It would be sensible for the auditory system to use this property as a distance cue, and there is evidence that this is exactly what occurs [Brungart *et al*, 1999]. If this is the case, then at least some part of the auditory system must have access to the locations implied by *both* the ITD and the ILD, and not just a combined version of them. It seems that the traded ITD and ILD values which give rise to lateralisation are the exception, rather than the norm, and may represent a "collapsed" special case of the distance detection mechanism.

The purpose of this discussion is to justify the lack of an ITD / ILD trade mechanism within the model, and the inability of the model to predict the performance of human subjects in trading experiments. As mentioned in the main text, the model as it stands *does not* predict the perceived location of a sound source, but it can correctly detect changes in the perceived location. The mechanism that detects the just noticeable change in source bearing is the same mechanism that detects a spatially separate target in the presence of a masker. Both these events can be thought of as generating blips on the "binaural radar". In reality, they generate slight interaural phase shifts that are detected by the internal correlation array.

The model predicts that a traded ITD and ILD sounds different from a monophonic signal because these two cues are not equivalent within the auditory system. Though a human listener will perceive both signals as being located at the centre of their head, the two signals will not "sound" the same. The traded sound will have an entirely different character from the monophonic sound – one or the other will sound distinctly more "natural". Whilst it would be possible to extend the model to trade ITD and ILD to make the two sounds equivalent, this would make the model deaf to any similar unnatural localisation characteristics which may be intro-

duced by an audio codec. An accurate perception would be given by the addition of a full localisation device to the model, so that the existing binaural processor would indicate that the sound had changed in some way, while the localisation processor would indicate that the sound was still located at the centre of the listener's head. This is discussed in Chapter 9.

F

Required number of binaural channels

It can be argued that the auditory system uses hundreds of overlapping channels of correlation, rather than the 31 of the Binaural model described in Chapter 7. Consequently, the auditory system contains no interpolating process comparable to the oversampling stage of the binaural model, since sufficient data is already present within the HAS. It is reasonable to question why this approach is not used within the present model, since the design principle of the model is to match known physiology as closely as possible.

The answer is that this approach could be used within the model, but it brings no functional advantage, and one severe disadvantage. The disadvantage is one of computational burden: processing hundreds of channels rather than 31 will increase the computation time by an order of magnitude.

To prove that this approach would bring no advantage, it is necessary to examine sampling theory. It is known that a continuous function may be represented by a series of discrete samples. The continuous function may be *exactly* recreated from these samples (i.e. the samples contain all the information that is present in the original function) if there are no frequency components higher than *fs/2* within the original continuous function, where *fs* is the sampling frequency, and *1/fs* is the spacing between sample points.

A signal with a period of 10 μs would have a frequency of 100000 Hz, or 100 kHz. A sampling frequency in excess of 200 kHz would be required to represent such a signal. However, a 10 μs delay between two 1 kHz signals is preserved within the sampled representation of the two

signals, even if the sampling frequency is only 5 kHz. In the time domain, the 10 μs delay can be verified, even though the sample points are separated by 200 μs.

Likewise, within the binaural model, the ITD is represented by the position of the peak within the time-delay domain. This peak will often fall between sample points, but it is correctly represented by those sample points, so long as no attempt is made to store frequency components in excess of half the sampling rate.

If a rectangular correlation window were used in the model, then frequency components higher than half the sampling rate in the time-delay domain could have been sampled. However, the gaussian windowing of the time-delay domain effectively low passes the data in this domain before it is (re)sampled. From sampling theory, this ensures that the channels contain all the information possible after the low pass filter. The addition of any further channels would not increase the accuracy of the information; instead, the extra channels would provide redundant information, the presence of which would simply slow the subsequent processing.

This raises a further question: is it appropriate to use a gaussian window function to low pass filter the time-delay domain signals? This is discussed in Section 7.4.3. The conclusion is that the rectangular correlation window of Colburn's model is effectively smoothed by the internal noise. The equivalent performance can be achieved in a model without internal noise by gaussian windowing of the EEVM output.

For these reasons, 31 gaussian windowed channels, subsequently oversampled, are equivalent to many hundreds of channels without oversampling.

The economies of storage and computation achieved by the use of oversampling should not be underestimated. The over sampled data is immediately cleared after the peak has been determined, before the next sample is processed. This saves an immense amount of storage space. Gaussian windowing 31 channels, rather than many hundreds, reduces the computational burden considerably.

G

Confidence value

The following figures (from [Zwislocki and Feldman, 1956]) demonstrate the manner in which the just detectable change in inter aural time delay increases as the level of the stimulus decreases.

level / dB	phase difference / °	ITD / μs
10	12.5	34.7
30	7.5	20.8
50	5.5	15.3
70	4.0	11.1
90	4.5	12.5
110	5.5	15.3

Table G.1: level and minimum audible phase difference, converted to ITD.

Stimuli reflecting the above conditions were processed through the model. The confidence values required to match these thresholds were determined. A psychometric function was adjusted to match these values.

A psychometric function is usually used to describe the probability of detecting the target signal in a masking experiment. The probability of detection $p(x)$ is given by

$$p(x) = \frac{1}{1 + e^{-m(x-T)}} \tag{G-1}$$

where x is the target amplitude, T is the amplitude at which the target has a 50% chance of detection, and m is the slope parameter defining the spread of the psychometric function. This equation is calculated by integrating the gaussian distribution function. See [King-Smith and Rose, 1997] for further discussion of the psychometric function equation.

A second psychometric function is used to reduce the confidence value to zero as the threshold falls to zero.

H

High quality listening test

H.1 Overview

This test is described in [Van den Berghe, WEB-1]. The results of this test are included here by permission of Roel Van den Berghe (r3mix). 42 listeners took part, 14 failed to differentiate the codecs, a further 13 scored the hidden reference lower than one of the other codecs. The remaining 15 listeners are categorised as "reliable". Removing the unreliable listeners does not change the ranking of the codecs, but only the range of the results. For this reason, the results of all listeners are included in the data used in Chapter 8, which is calculated by summing all the scores on a codec by codec basis.

H.2 Procedure

Eight different high-quality mp3 codec settings are compared. The purpose of the test is to determine which combination of psychoacoustic model, bit allocation scheme, and bitrate offer the highest quality in an open source mp3 encoder [Taylor, WEB]. The large number of listeners who failed to differentiate between the various settings and the hidden reference is an indication of the high quality of the codec. The lowest quality sample is encoded at 192 kbps constant bitrate. The hidden reference is coded by the MPEG*plus* encoder [Buschmann, WEB], tuned to provide near lossless coding at 740kbps. This near-lossless codec is employed as a hidden reference (instead of the original signal) to prevent listeners from discovering the hidden reference by waveform subtraction. The scores of listeners who mis-scored this codec are *not* discarded, so the hidden reference has indicative status only. Thus, any arguments about the applicability of a (ultra high quality) coded extract being used as the hidden reference are void.

The stimulus consists of a selection of audio extracts. All listeners who submitted reliable results commented that the solo hi-hat extract suffers significantly greater audible degradation than any of the other samples. For this reason, only the hi-hat extract is auditioned by the model. Many listeners report audible pre-echo problems, or temporal smearing. This is a critical test for the model, as a previous version of the monophonic detector (Appendix J) performed poorly at signal onsets, where such distortion occurs.

Each listener graded each extract using the 5-point continuous grading scale described in Chapter 2. Most listeners used a single decimal place, and most of the reliable listeners commented in their feedback that they had used ABX testing to ensure that any differences they perceived were not imagined.

The decoded audio extracts employed in the test are included on the accompanying CD-ROM. Thus, the test is verifiable and repeatable. The reader may wish to audition the audio extracts to gauge the magnitude of the audible errors that are presented within this subjective test.

H.3 Results

clip ref:	0a	24	35	41	61	a1	b1	c3
lame 3.90a7 mode:	cbr192 MS GPSY-CHO	abr224 MJ nspsytune nssafejoint	--r3mix	dm-xtr	MPEG+ insane -nmt99 -tns99 lowpass 19.5	dm-ins	cbr256 nspsytune nssafejoint	dm-std
bitrate/kbps	192	221.3	199.6	231.8	740	272.8	256	221.9
result 01	4.0	4.5	4.1	4.5	5.0	4.0	4.0	4.3
result 02	3.7	3.9	4.0	4.2	3.8	3.5	3.2	3.7
result 03	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 04	4.0	4.0	4.0	4.0	5.0	5.0	5.0	4.0
result 05	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 06	4.5	4.7	*4.7*	*4.7*	5.0	*4.7*	*4.7*	*4.7*
result 07	3.5	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 08	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 09	3.3	4.0	3.3	3.0	5.0	3.2	3.8	3.5
result 10	4.0	4.4	4.5	4.2	4.7	4.3	4.6	4.1
result 11	3.7	4.0	4.5	4.5	4.5	3.7	3.7	5.0
result 12	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
result 13	4.4	5.0	4.5	5.0	4.9	5.0	4.6	5.0
result 14	4.0	3.5	3.0	3.5	4.0	3.5	3.5	3.5
result 15	2.5	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 16	3.0	2.0	2.5	3.0	4.0	4.0	3.5	5.0

result 17	3.5	4.0	4.7	5.0	5.0	5.0	4.9	5.0
result 18	3.5	4.0	3.5	5.0	4.0	3.0	3.0	4.0
result 19	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 20	4.0	4.2	4.3	4.7	4.6	4.1	4.4	4.5
result 21	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 22	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 23	4.2	3.8	3.8	4.0	4.0	4.5	4.3	4.2
result 24	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 25	4.8	4.8	4.8	4.8	4.8	5.0	4.8	4.8
result 26	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 27	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 28	5.0	4.9	5.0	5.0	4.9	5.0	5.0	4.9
result 29	4.9	5.0	4.9	4.9	4.9	5.0	5.0	4.9
result 30	4.0	4.0	4.0	4.0	5.0	5.0	4.0	5.0
result 31	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 32	4.6	4.8	4.6	4.6	4.4	4.6	4.6	4.6
result 33	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 34	4.5	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 35	4.5	4.2	5.0	5.0	5.0	4.4	4.2	5.0
result 36	5.0	4.0	4.0	4.0	5.0	4.0	4.0	4.0
result 37	4.3	5.0	4.8	4.8	5.0	4.7	5.0	4.7
result 38	4.9	4.5	4.6	5.0	4.9	4.9	4.9	4.5
result 39	5.0	4.8	4.6	5.0	5.0	4.5	4.5	4.7
result 40	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
result 41	3.4	4.2	3.5	4.4	3.9	3.7	4.0	4.5
result 42	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
total (/210)	183.7	190.4	189.2	194.6	200.3	192.3	191.2	196.1
lame 3.90a7 mode:	cbr192 MS GPSY-CHO)	abr224 MJ nspsytune nssafejoint	--r3mix	dm-xtr	MPEG+ -insane -nmt 99 -tns99 lowpass 19.5	dm-ins	cbr256 nspsytune nssafejoint	dm-std

Table H-1: Results from the high quality audio test

H.4 Discussion

Post-screening of results is desirable (as discussed in [ITU-R BS.1116, 1997]). In this test the hidden reference reveals that some listeners are simply guessing. Interestingly, removing these listeners from the test does not change the ranking of the codecs, or even the spacing significantly. This suggests that, over the number of listeners who took part (42), the random guesses from listeners who cannot accurately detect any difference between the original and coded samples have cancelled. For this reason, and because the hidden reference is not identical to

the original, no filtering of results is carried out. The totals at the foot of the above table are used in Chapter 8.

Several statistical analyses have been attempted using the above data, to try to define the significance of the difference between the eight extracts. At 95% confidence, cbr192 is worse than the others, and MPC is better, but no other results are significant at the 95% confidence level. Reducing the confidence level to 68%, reveals further trends, but statistically these differences between codecs are less significant.

I

Medium quality listening test

I.1 Overview

This test is described in [Miyaguchi, WEB]. The results of this test are included here by permission of Darryl Miyaguchi (ff123). 16 listeners took part and one failed to differentiate between the codecs. No hidden reference is included in this test, but a low quality anchor is included that is significantly worse than the other codecs. One listener failed to score this codec as the worst on test. Thus, 14 listeners are classified as "reliable". The overall codec ranking is unchanged by excluding the "unrealiable" listeners, so data from all 16 listeners is averaged to produce the Mean Opinion Scores used in Chapter 8.

I.2 Procedure

Six different codecs are compared at a nominal bitrate of 128 kbps. The codecs are as follows:

Format Type	**Encoder Name and Version**	**Settings (128 kbit/s)**	**Decoder Name and Version**
MP3	Lame 3.89beta	--abr 134 -h --nspsytune --athtype 2 --lowpass 16 --ns-bass -8	in_mp3.dll (version 2.75i) default mp3 decoder within Winamp 2.76
MP3	Xing within AudioCatalyst 2.1	128 kbit/s, high frequency mode disabled, simple stereo disabled	in_mp3.dll (version 2.75i) default mp3 decoder within Winamp 2.76
AAC	Liquifier Pro 5.0.0 Beta 2, Build 24	streaming 128, audio bandwidth set at 17995 Hz .	in_lqt.dll (v. 1.055)

MPC	mppenc.exe version 1.7.9c	-radio -ltq_gain 10 -tmn 12 -nmt 4.8	mppdec.exe 1.7.8c
WMA8	Windows Media Player 7.1 (version 7.01.00.3055); wmadmoe.dll version 8.0.0.0371	128 kbit/s	Windows Media Player 7.1 (version 7.01.00.3055); output captured by Total Recorder
Ogg Vorbis	Oggdrop RC2 for Windows 32	128 kbit/s	in_vorbis.dll: Nullsoft Vorbis Decoder v1.13c (RC1)

The Xing MPEG-1 layer III codec is the low quality anchor. This mp3 encoder is optimised for speed, rather than quality, and gives significantly poorer quality than other mp3 encoders at this bitrate.

The stimulus consists of a single audio extract from “Git Along Little Dogies” by Dave Grusin from the album *Discovered Again*. This extract is chosen because it is not especially difficult to code, and may give an indication of the “typical” performance of these audio codecs at 128 kbps.

Each listener graded each extract using the 5-point continuous grading scale described in Chapter 2. Most listeners used a single decimal place, and many commented in their feedback that they had used ABX testing to ensure that any differences they perceived were not imagined.

The decoded audio extracts employed in the test are included on the accompanying CD-ROM. Thus, the test is verifiable and repeatable. The reader may wish to audition the audio extracts to gauge the magnitude of the audible errors that are presented within this subjective test.

I.3 Results

		235 (MPC)	106 (AAC)	854 (OGG)	875 (LAME)	027 (WMA)	740 (XING)	
1	Garf	5.0	2.5	1.5	1.5	1.5	1.0	2.17
2	Filburt	3.5	2.7	2.5	2.2	1.8	1.0	2.28
3	JohnV	3.2	3.0	2.6	2.0	2.7	1.0	2.42
4	r3mix	3.0	4.1	1.5	3.0	0.9	2.0	2.42
5	HansHeijden	4.0	4.5	2.0	3.0	1.0	1.0	2.58
6		2.5	4.0	3.0	2.0	3.0	1.0	2.58
7		4.0	3.0	3.0	2.0	3.0	2.0	2.83
8		4.0	5.0	4.0	3.0	2.0	1.0	3.17
9	ff123	5.0	5.0	3.5	3.0	4.0	1.5	3.67
10	2BDecided	4.7	3.7	3.8	4.0	4.5	3.5	4.03
11		5.0	3.6	5.0	5.0	--	2.8	4.28
12		5.0	4.8	4.2	4.0	5.0	4.0	4.50
13		5.0	5.0	4.5	5.0	4.0	3.8	4.55
14		5.0	5.0	4.5	5.0	5.0	3.0	4.58
15		5.0	5.0	5.0	4.0	5.0	5.0	4.83
16		5.0	5.0	5.0	5.0	5.0	5.0	5.00
		4.31	4.12	3.48	3.36	3.23	2.41	Averages

Table I-1: Results from the medium quality audio test

I.4 Discussion

Due to the lack of a hidden reference in this test, a sophisticated analysis is suggested in [Miyaguchi, WEB], to determine the significance of the results. The analysis takes the form of a Friedman Non-Parametric Analysis [Meilgaard *et al*, 1999]. This analysis takes the implied rankings of each codec by each listener, and determines the statistical probability that each codec is audible different from each other codec. A confidence level of 95% is chosen. The result shows that there are three groups of codecs, which are ranked in the following order:

mpc = 75.0	aac = 71.5	wma8 = 50.5	ogg = 49.5	lame = 44.5	xing = 24.0

Thus, it is *not* possible to state that WMA8 is better than lame at the 95% confidence level.

This analysis is interesting, but does little more than add a statistical basis to the differences that are apparent in the mean opinion scores. The averages in table I-1 are used in Chapter 8.

J

Original difference perception unit

J.1 Overview

The difference perception unit of the monophonic model described in Chapter 5 is the second version of this device. The original version was published in [Robinson and Hawksford, 1999]. This original version was oversensitive to the onset of sounds, hence the new version was developed. However, since the new version is insensitive to differences in real music signals, the old version was re-examined in Chapter 8. The following details of the original difference perception unit are taken directly from [Robinson and Hawksford, 1999]. "The model" became the monophonic model described in Chapter 5. The calibration, validation, and testing of the model with the older detector are also included in this appendix.

J.2 Perceiving a difference

So far, we have simulated the passage of sound through the ear canal and cochlea – see Figure 3.1(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 illus-

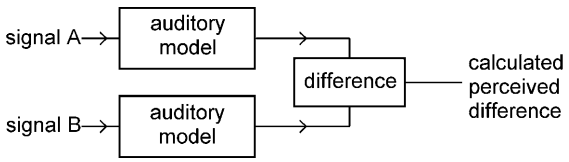


Figure J.1: General method for calculating the perceived difference between two audio signals

trated in Figure J.1. 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. 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.

J.2.1 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.

J.2.1.1 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 J.1, *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 [Moore *et al*, 1998], *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 [Zwicker, 1984], *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.

J.2.2 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 [Theide *et al*, 1999]). However, the discrepancies described in the previous section can be dealt with without a complex cognitive model.

Rather than examining 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 the previous section.

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.

J.2.3 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

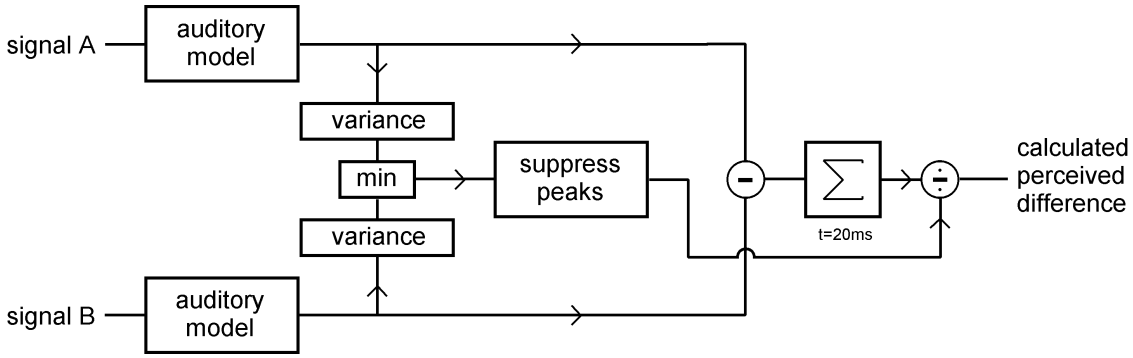


Figure J.2: actual method for calculating the perceived difference between two audio signals

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 J.2, 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.

J.3 Validation of the model

J.3.1 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 J.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 J.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.

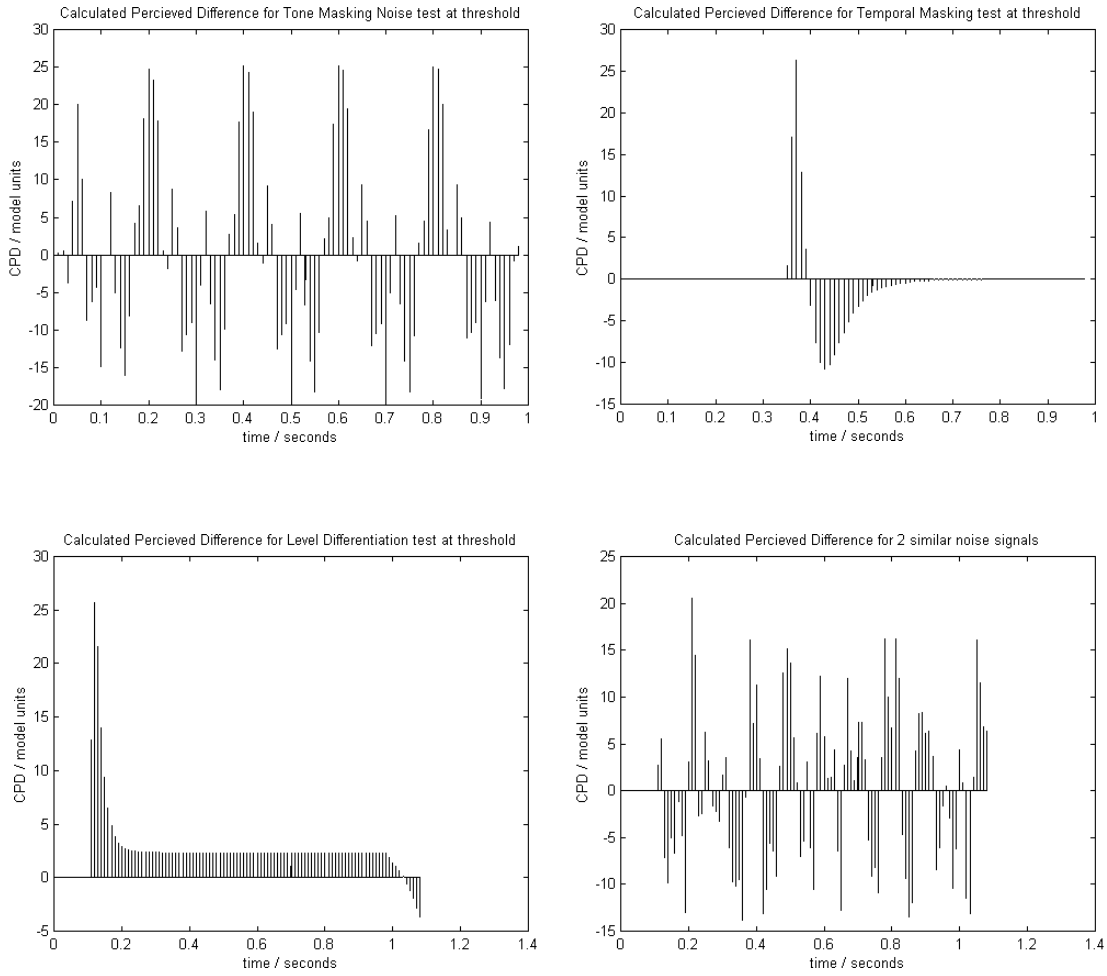


Figure J.3: CPD at threshold

(a) tone masking noise; (b) temporal masking; (c) level differentiation; (d) noise

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

J.3.2 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".

J.3.2.1 MPEG-1 layer II 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 II 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 [Brandenburg and Bosi, 1997]) 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.

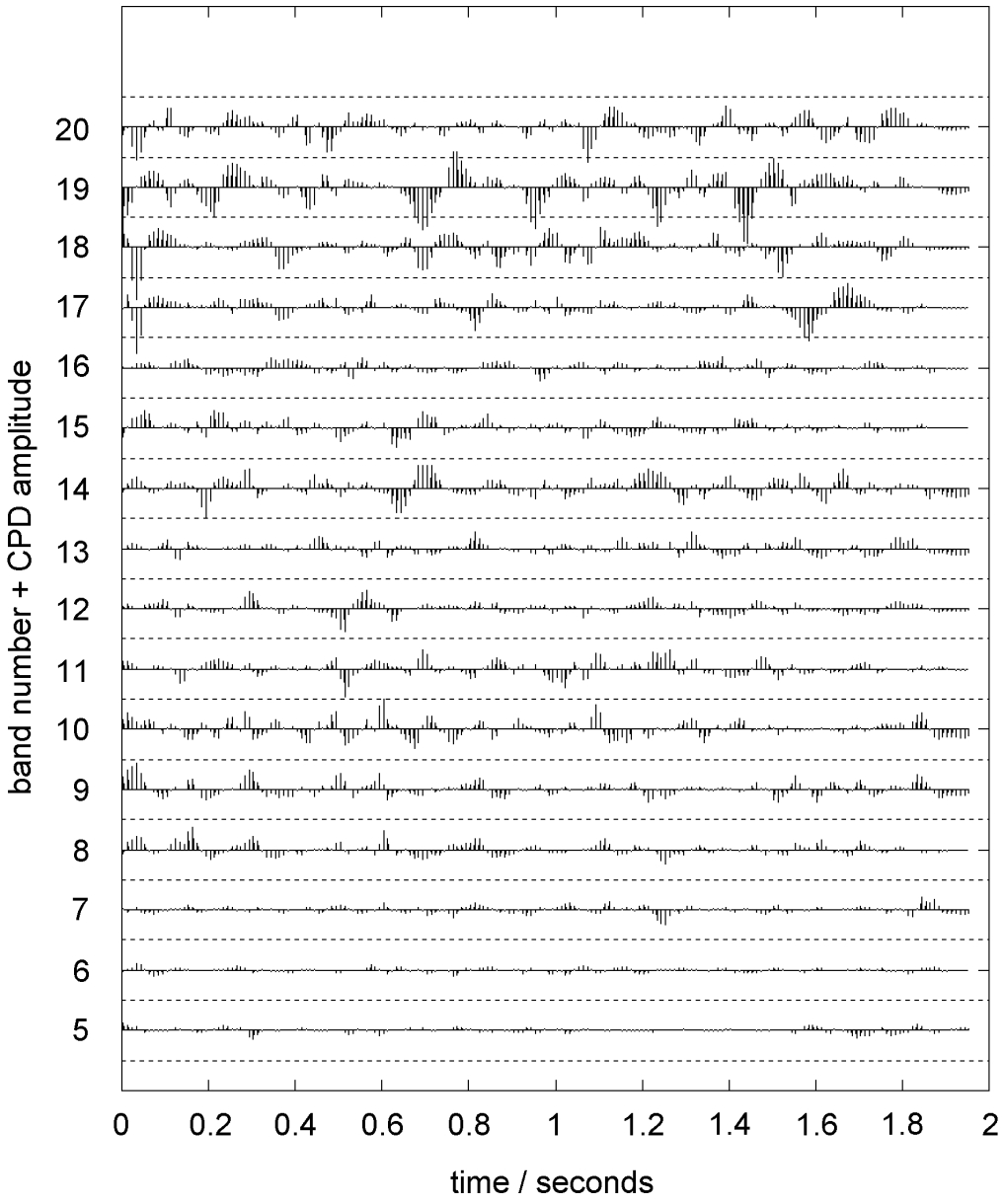


Figure J.4: Calculated perceived difference for an audio signal coded using MPEG-1 layer II

Only 15 of the 25 bands are shown. The dashed lines show the threshold of audibility in each band (where CPD > 25).

Each coded stream was analysed by the model (Figure J.2), and compared with the original. The result is a time varying CPD output from each band (see Figure J.4). Table J.2 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 J.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 J.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.

J.3.2.2 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 III 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 II 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 II 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.

J.4 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.

K

Replay Gain – A Proposed Standard

One possible application of the model is as a perceptual loudness meter. The outer and middle ear processing, and the amplitude dependent gammachirp filter bank are required for this task. The peak amplitude of the signal within each band is used as the excitation in the model of [Moore *et al*, 1997]. This transforms Moore's analytical model into a time-domain model.

The output is a time-domain loudness surface. The instantaneous area under this surface represents the perceived loudness, but for a real music signal, the integration of this area does not correctly predict the overall loudness of the whole track. Instead, the 95th percentile value of the complete set of instantaneous perceived loudness measures gives a good indication of the overall perceived loudness of the whole track.

If this calculation is carried out for a real music signal, the accuracy of the preceding model is redundant, since the statistical processing at the end is significantly more important than the auditory model itself. This raises the possibility of vastly simplifying the whole process, so that it can run at faster than real time on a typical PC, whilst still yielding a reasonably accurate prediction of overall perceived loudness.

This approach forms the basis of the replay gain standard, which is described in this appendix. This is currently published on the world wide web at http://www.replaygain.org/

Replay Gain – A Proposed Standard

K.1 Introduction

Not all CDs sound equally loud. The perceived loudness of mp3s is even more variable. Whilst different musical moods require that some tracks should sound louder than others, the loudness of a given CD has more to do with the year of issue or the whim of the producer than the intended emotional effect. If we add to this chaos the inconsistent quality of mp3 encoding, it is no wonder that a random play through your music collection can have you leaping for the volume control every other track.

There is a remarkably simple solution to this annoyance, and that is to store the required replay gain for each track *within* the track. This concept is called "MetaData" – data about data. It is already possible to store the title, artist, and CD track number within an mp3 file using the ID3 standard. The later ID3v2 standard also incorporates the ability to store a track relative volume adjustment, which can be used to "fix" quiet or loud sounding mp3s.

However, there is no consistent standard method to define the appropriate replay gain which mp3 encoders and players agree on, and no automatic way to set the volume adjustment for each track – until now.

The Replay Gain proposal sets out a simple way of calculating and representing the ideal replay gain for every track and album.

K.1.1 Perceived Loudness

The perceived loudness of an audio track is a combination of several factors. These include the audio data stored on the CD, the position of the listener's volume control, the characteristics of the listener's CD player, amplifier and speakers, and the listener him or herself.

The position of the volume control, and characteristics of the CD player, amplifier and speakers combine to yield an overall system gain. This gain can be calibrated, so that a known digital signal level on a CD will yield a known real world sound pressure level. However, this calibration is rarely carried out in the home listening environment.

Audio engineering experience suggests that the individual listener can be the most variable and unpredictable component a typical audio system. However, "expert" listeners are often in remarkable agreement as to how loud a given audio track should be reproduced. For most acoustic music, this corresponds to the loudness of the original performance. For other types of music, it is difficult to define the origin of this preference, but a consistent preference (to within one or two dB) is often observed across many listeners. For this reason, the listener can be regarded as the most predictable component in this chain.

Surprisingly, the least predictable factor is the audio data on each individual CD. Compare a pop single to a classical album, and the audio data would suggest that the former should be 15 dB louder than the latter. This is clearly nonsense, unless Britney Spears really does sing louder than a full symphony orchestra.

Armed with a volume control, good taste, and common sense, the listener must correct for these unwanted differences in loudness between different discs.

This paper suggests an alternative, which will leave the listener free to enjoy the music. The basic concept is outlined, and an algorithm for calculating the perceived loudness of an audio track is described in detail. The proposed data format is specified, and the requirements for audio players are explained in depth. This entire proposal focuses on the world of PC audio, since everything described in this paper can be accomplished in software. However, the relevance of this proposal to new audio formats should not be underestimated. The author strongly suggests that the meta data defined herein would be a valuable addition to the DVD-audio and SACD specifications.

K.1.2 Definitions

Before continuing, it is necessary to define some terms. These details will be expanded in the following sections.

The Replay Gain is the gain of the entire system, which links the digital signal on a CD to the real world sound pressure level.

All Replay Gain Adjustments are specified relative to SMPTE RP 200. Hence, a Replay Gain Adjustment of -10 dB suggests that the system gain should be reduced by 10 dB relative to SMPTE RP 200. In software, this may achieved by scaling the digital data, at the expense of some SNR.

The ideal loudness is usually equal to the loudness of the original acoustic event. As discussed above, this can often be judged by experienced listeners to within one or two dB.

The Audiophile Replay Gain Adjustment is the gain adjustment, relative to SMPTE RP 200, which will yield the ideal loudness for a given track. If the track is mastered via a system calibrated to SMPTE RP 200, then the Audiophile Replay Gain Adjustment will be 0 dB.

The Radio Replay Gain Adjustment is the gain adjustment, relative to SMPTE RP 200, which will make the overall perceived loudness for a given track equal to 83 dB SPL. Thus, the Radio Replay Gain Adjustment will cause all tracks to sound equally loud.

ReplayGain is the name of the algorithm described within this paper. It calculates the overall perceived loudness of an audio track, and from this calculates a Radio Replay Gain Adjustment. It can also guess a value for the Audiophile Replay Gain Adjustment, but this value is only approximate.

The Replay Gain Standard is defined in this paper, and on the website www.replaygain.org, which together form the Replay Gain Proposal.

Finally, the signal level usually refers to the magnitude of the digital audio data. Without specifying the Replay Gain, this says nothing about the perceived loudness.

K.1.3 Basic Concept

The following basic method is suggested to equalise the loudness of all audio tracks.

1. Calculate the perceived loudness of each track
2. Convert this to a suggested Replay Gain Adjustment
3. Store this value within the header of the audio file
4. Adjust the volume of each track accordingly on replay

The Replay Gain Adjustment is a suggested gain adjustment for each track. Players can scale the audio data by this adjustment in order to achieve a consistent perceived loudness across all tracks.

The reference gain yields a real world loudness of 83dB SPL, as defined in the SMPTE RP 200 standard. If the Replay Gain Adjustment for a given track is -12dB, this means that the track is relatively loud, and the gain should be reduced by 12dB, ideally to 71dB. Players that understand the Replay Gain Standard will do this automatically.

Steps one and two can be carried out automatically, using the *ReplayGain* algorithm described in this paper. Alternatively, a human listener (e.g. the record producer or mastering engineer) may specify the ideal Replay Gain Adjustment relative to SMPTE RP 200, and this value can be stored instead.

K.2 “Radio” and “Audiophile” gain adjustments

Under some listening conditions, it is useful to adjust every track to sound equally loud. However, it is often desirable to leave the *intentional* loudness differences between tracks in place, whilst still correcting for unmusical and annoying changes in loudness between discs.

To account for these two separate listening conditions, the Replay Gain Proposal suggests that two different gain adjustments should be stored in the file header, as follows.

K.2.1 "Radio" Replay Gain Adjustment

This will cause all tracks to sound equally loud (as they typically do on the radio, hence the name). If the Replay Gain is calculated on a track-by-track basis (i.e. an individual *ReplayGain* calculation is carried out for each track), this will be the result. This is something that the *Re-*

playGain algorithm described in this paper does very well. An audio demonstration is included on the accompanying CD-ROM.

K.2.2 "Audiophile" Replay Gain Adjustment

The problem with the "Radio" setting is that tracks which *should* be quiet will be brought up to the level of all the rest. For casual listening, or in a noisy background, this can be a good thing. For serious listening, it would be a nuisance. For example, it would be inappropriate to raise the volume of a flute solo to match that of a symphony orchestra.

To solve this problem, the "Audiophile" setting represents the ideal listening gain for each track. *ReplayGain* can have a good guess at this too, by reading the entire CD, and calculating a single gain adjustment for the whole disc. This works because quiet tracks will remain quieter than the rest of the disc, since the gain won't be changed for each track. It still solves the basic problem of annoying, unwanted level differences between discs because quiet or loud discs are still adjusted overall. This means that a typical pop CD that's 15 dB louder than a classical CD will be brought into line.

ReplayGain will fail to calculate a reasonable “audiophile” setting for an entire CD of quiet music. Instead, it will raise it to an average loudness. This is why the "Audiophile" Replay Gain Adjustment must be user adjustable. The *ReplayGain* whole disc value represents a good guess, and should be stored in the file. Later, the user can tweak this value if required.

If the file has originated from the artist (e.g. download from mp3.com), then the "Audiophile" setting can be specified by the artist. Naturally, the user is free to change this value if they desire.

K.3 Replay Gain Adjustment Calculation

The following *ReplayGain* algorithm is suggested to calculate the perceived loudness, and hence the appropriate Radio Replay Gain Adjustment value for a given audio track. The stages of the calculation are outlined below, and discussed in detail in the following sections.

EQUAL LOUDNESS FILTER

The human ear does not perceive sounds of all frequencies as having equal loudness. For example, a full scale sine wave at 1 kHz sounds much louder than a full scale sine wave at 10 kHz, even though the two contain identical energy. To account for this, the signal is filtered by an inverted approximation to the equal loudness curves (sometimes referred to as Fletcher-Munson curves).

RMS ENERGY CALCULATION

Next, the energy during each moment of the signal is determined by calculating the Root Mean Square of the waveform every 50ms.

STATISTICAL PROCESSING

Where the average energy level of a signal varies with time, the louder moments contribute most to our perception of overall loudness. For example, in human speech, over half the time is silence, but this does not affect the perceived loudness of the talker at all. For this reason, the 95^{th} percentile is chosen to represent the overall perceived loudness of the signal.

CALIBRATION AND REFERENCE LEVEL

A suitable average replay level is 83 dB SPL. A calibration relating the energy of a digital signal to the real world replay level has been defined by the SMPTE. Using this calibration, the current signal level is subtracted from the desired (calibrated) level to give the Replay Gain Adjustment.

Each stage of this calculation will now be described in detail.

K.3.1 Equal Loudness Filter

Figure K.1 shows the Equal Loudness Contours, taken from [Robinson and Dadson, 1956]. The original measurements were carried out by Fletcher and Munson in 1933, and the curve often carries their name.

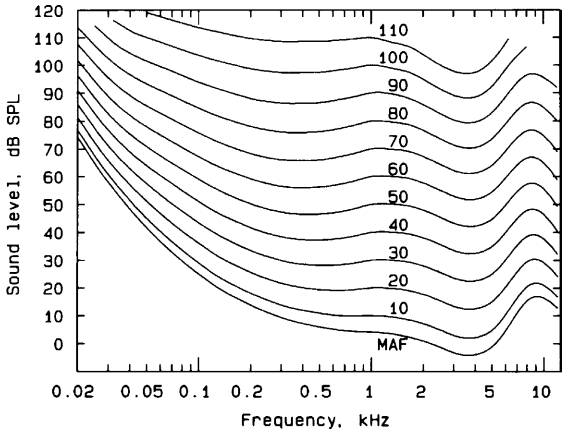


Figure K.1: Fletcher Munsen curves

The lines represent the sound pressure required for a test tone of any frequency to sound as loud as a test tone of 1 kHz. Examine the line marked "60" - at 1 kHz, the line marked "60" is at 60 dB. Following the "60" line down to 0.5 kHz, the value is about 55 dB. This shows that a 500 Hz tone at 55 dB SPL sounds as loud to a human listener as a 1 kHz tone at 60 dB SPL.

If every frequency sounded equally loud, then this graph would consist of a series of horizontal lines. Since it is very far from this ideal, a filter is required to simulate this characteristic of the human auditory system.

Where the lines curve upwards, humans are less sensitive to sounds of that frequency. Hence, the filter must attenuate (reduce) sounds of that frequency in order to simulate the response of the human auditory system. The ideal filter will be the inverse of the equal loudness response. Unfortunately, this response varies with loudness. It would significantly complicate the calculation to use a different filter for different amplitude audio signals, and the performance gains would be questionable. For this reason, a weighted average of the responses is chosen as the target filter response, as shown in Figure K.2.

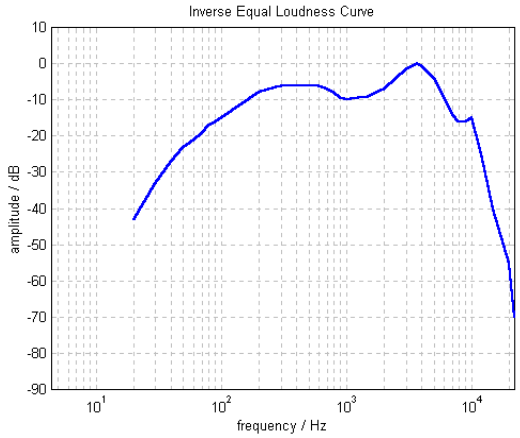


Figure K.2: Target Response – The Average Inverse Equal Loudness Curve

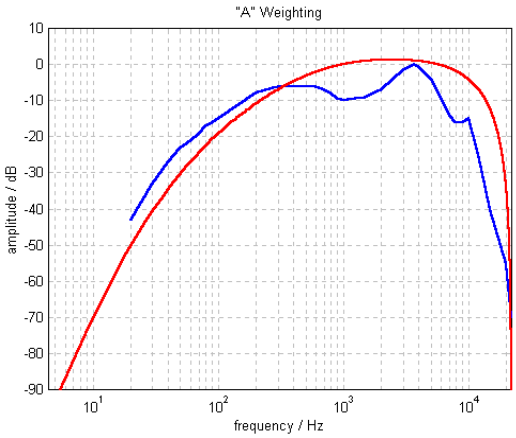


Figure K.3: Possible filter response: A-weighting curve (red)

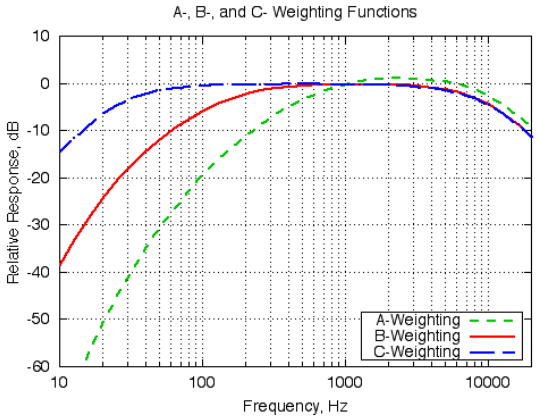


Figure K.4: Other possible filters: A, B, and C-weighting curves

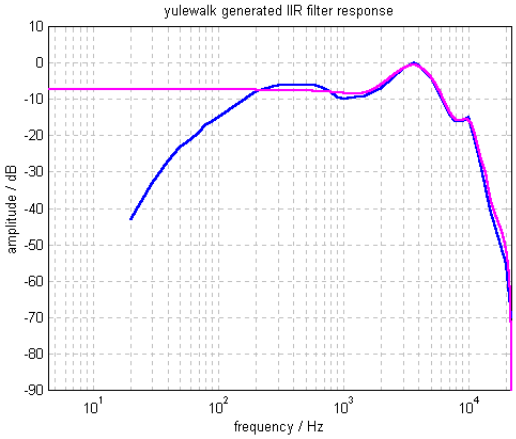


Figure K.5: IIR filter generated by yulewalk.m

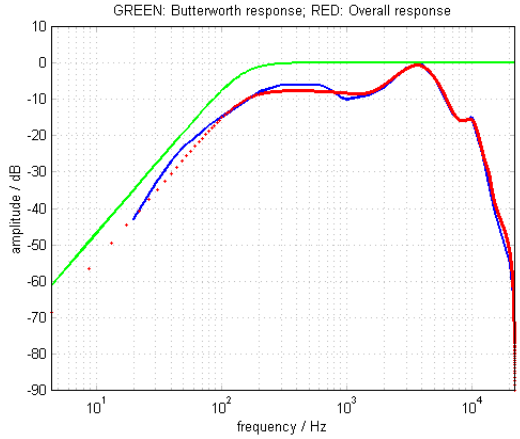


Figure K.6: Composition of final filter. RED: Overall response

A common approximation to the frequency response of the human ear is the "A" Weighting curve [IEC 1672, 1996]. Figure K.3 shows the response of a 2x7 coefficient IIR filter designed to match the A weighting curve. As can be seen from the graph, the A-weighting curve is only an *approximation* to human hearing sensitivity, and is not sufficiently accurate for our purposes. For example, at 15 kHz, the response passes 30 dB *more* energy than the ears of a human listener.

The "B" and "C" weighting curves are even less appropriate, as illustrated in Figure K.4 [Singleton, WEB]. Since none of these standard weighting curves gives a suitable frequency response, a custom filter is designed, as follows.

MATLAB offers several functions to design FIR and IIR filters to match arbitrary amplitude responses. Feeding the target response into `yulewalk.m`, and requesting a 2x10 coefficient IIR filter gives the response shown in Figure K.5

At higher frequencies, this filter is an excellent approximation to the target response. However, at lower frequencies, it is inadequate. Increasing the number of coefficients does not cause the yulewalk function to perform significantly better.

One solution is to cascade the yulewalk filter with a 2nd order Butterworth high pass filter having a high pass frequency of 150 Hz. The resulting combined response is close to the target response, as shown in Figure K.6. This filter is used in *ReplayGain*.

K.3.1.1 Implementation

The MATLAB script `equalloud.m` generated the above graphs. It requires `adsgn.m` [Couvreur, 1997] to generate the "A" weighting filter for comparison. The `yulewalk` and `butter` functions are built in to MATLAB. The filter design function called by the MATLAB implementation of *ReplayGain* is `equalloudfilt.m`. For readers without access to MATLAB, a text document containing all the filter coefficients is provided. All these files are included on the accompanying CD-ROM.

K.3.2 RMS Energy Calculation

It is trivial to calculate the RMS energy over an entire audio file. For example, Cool Edit Pro [Syntrillium, 2000] provides this information via its Analise:statistics box. Unfortunately, this value does not give a good indication of the perceived loudness of a signal. Whilst it is closer than that given by the peak amplitude, it is still less accurate than is required for the current task. For this reason, the RMS energy is calculated on a moment by moment basis (as described in this section), and this information is subsequently processed to yield the overall perceived loudness (as described in Section K.3.3).

The audio data is partitioned into blocks, and the RMS value of the samples within each block is calculated. The block length of 50 ms was chosen after studying the effect of values between 25 ms and 1 second. 25 ms is too short to accurately reflect the perceived loudness of some sounds. Beyond 50 ms there is little change in performance (*after* statistical processing). For

this reason, 50 ms was chosen. Alternatively, the mp3 frame length of 52 ms may be used for convenience without significantly compromising the accuracy of the calculation.

K.3.2.1 Stereo Files

A suitable method to process stereo files was investigated, as the correct calculation was not immediately obvious. It is possible to sum the two channels into a single channel before calculating the RMS energy, but then any out-of-phase components cancel out to zero. This is not the manner in which humans perceive such sounds, so it is not an appropriate solution.

One alternative consists of calculating two RMS values (once for each channel) and then adding them. Unfortunately, a linear addition does not yield the same perceived loudness as listening to the two signals via loudspeakers. To demonstrate this, consider a mono (single channel) audio track. Replay this signal using one loudspeaker, and remember how loud it sounds. Now replay the same track using two loudspeakers, and determine how large a signal is required to *each* speaker such that, overall, the perceived loudness is equal to the previous presentation. A linear addition would suggest that the signal should be half as large, but the correct value is approximately three-quarters, as determined experimentally.

The correct answer can be obtained if the means of the channel-signals are averaged *before* calculating the square root. Possibly, this could be referred to as “Root mean mean square”. In mixing desk terminology, this equates to "equal power" rather than "equal voltage" pan-pots. If it is also assumed that any mono (single channel) signal will always be replayed over two loudspeakers, a mono signal may be treated as a pair of identical stereo signals. Hence a mono signal gives (a+a)/2 (i.e. **a**), while a stereo signal gives (a+b)/2, where a and b are the mean squared values for each channel. The result is square rooted and converted to dB.

K.3.2.2 Implementation

In the MATLAB implementation, the RMS calculation is carried out by the following lines (modified here for clarity) from `ReplayGain.m`:

```
% Mono signal: just process the one channel
if channels==1,
        Vrms_all(this_block)=mean(inaudio(start_block:end_block).^2);
% Stereo signal: calculate average Vrms of both channels
elseif channels==2,
        Vrms_left=mean(inaudio(start_block:end_block,1).^2);
        Vrms_right=mean(inaudio(start_block:end_block,2).^2);
        Vrms_all(this_block)=(Vrms_left+Vrms_right)/2;
```

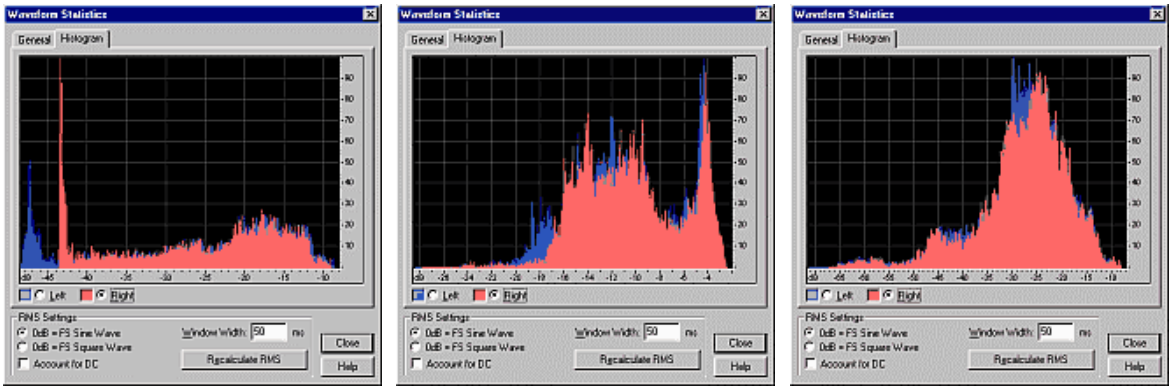


Figure K.7: Histograms of RMS energy values in 3 audio tracks

(a) speech (b) pop music (c) classical music

```
end
% Convert to dB
Vrms_all=10*log10(Vrms_all+10^-10);
```

In the last line, the addition of ten to the power minus ten (10^-10) prevents the calculation of log(0) (which would give an error) during periods of digital silence. A level of approximately -100 dB is calculated instead, which (on this scale) is below the noise floor of a 24-bit recording.

10*log10(signal) is the same as 20*log10(square_root(signal)). Thus, the square root and the conversion to dB are carried out in one simple step, without the use of a square root function.

K.3.3 Statistical Processing

Having calculated RMS signal levels for every 50 ms of the file, a single value must be calculated to represent the perceived loudness of the entire file. The histograms in Figure K.7 show how many times each RMS value occurred in each file, for three different types of music.

The most common RMS value in the speech track was -45 dB (background noise), so the most common RMS value is not a good indicator of perceived loudness. The average RMS value is similarly misleading with the speech sample, and with classical music.

A possible method to determine the overall perceived loudness is to sort the RMS energy values into numerical order, and then pick a value near the top of the list. This suggestion is analogous to one published in [Zwicker and Zwicker, 1991]. After computing the fluctuation of perceived loudness with time, “Comparisons of the loudness perceived by many subjects

have indicated that the average loudness corresponding to N_{50} (the loudness exceeded in 50% of the time) gives an inadequate number, whereas N_5 to N_{10} give adequate readings of what the subjects really perceive."

In the present case, the correct choice of percentile is one of the most critical aspects of *ReplayGain*. Values from 70% to 95% were tested using a wide range of audio material. For highly compressed pop music (e.g. Figure K.7 (b), where there are many values near the top), the choice makes little difference. For speech and classical music, the choice makes a significant difference. The value which most accurately matches human perception of loudness is found to be around 95%, so this value is used by *ReplayGain*.

K.3.3.1 Implementation

Since MATLAB has a dedicated sort function, and simple indexing into arrays, this task is carried out in two lines of code (from `ReplayGain.m`):

```
% Sort the Vrms values into numerical order
Vrms_all=sort(Vrms_all);
% Pick the 95% value
Vrms=Vrms_all(round(length(Vrms_all)*0.95));
```

...where length(Vrms_all)*0.95 is just an index 95% of the way into the sorted array. It must be "round"ed because MATLAB will not accept non-integer array indexing.

K.3.4 Calibration and Reference Level

It is necessary to reference the RMS energy value to a real world sound pressure level. This is essential for the Audiophile Replay Gain Adjustment, but it is also desirable for the Radio Replay Gain Adjustment.

The *audio industry* doesn't have a fixed standard for Replay Gain, but the *movie industry* has worked to an 83 dB standard for years, called SMPTE RP 2000. (see [SMPTE, 1999] for full details). This standard states that a single channel pink noise signal, with an RMS energy level of -20 dB relative to a full scale sinusoid, should be reproduced at 83 dB SPL (measured using a C-weighted, slow averaging SPL meter). The result is that every listener (or cinema) can set their volume control to the same (known, calibrated) gain, and all film soundtracks are mixed to sound best when reproduced at this gain.

As an aside, it is worth noting that the 83 dB SPL reference wasn't picked at random. It was chosen because it represents a comfortable average listening loudness, as determined by professionals from years of listening. The reference level of -20 dB pink noise causes the calibrated average to be 20 dB less than the peak signal level. This yields 20 dB of headroom for louder than average signals. If CDs were mastered to this standard, the average level would be around –20 dB FS, which would leave lots of room for the dramatic peaks which make music exciting. Such a standard has been proposed [Katz, 2000], but this has yet to gain wide acceptance within the audio industry. This standard defines three different calibrated monitor gains, so it does not entirely solve the problem of different CDs requiring different gain settings.

The consequence of adopting a single calibrated monitor gain would be that all CDs should be reproduced using the same gain setting, and they would all “sound right” at that setting. If you (as a listener) didn't want to listen at that particular gain setting, you could always turn it down, but all CDs would *still* sound equalling "turned down" at your preferred setting. You wouldn't have to change the volume setting between discs.

In reality, CDs are not mixed or mastered using a single calibrated monitor gain, hence the huge difference in perceived loudness between discs. However, *ReplayGain* aims to calculate how much louder or quieter a CD is compared to one that has been mastered at the calibrated gain, and adjust appropriately. This process involves a few assumptions, but can give good results. The process is as follows.

For the Radio Replay Gain Adjustment, it is assumed that the perceived loudness should average around 83 dB SPL. SMPTE RP 200 states that a -20 dB pink noise signal will yield 83 dB SPL when replayed via a calibrated system. Hence, the average level of all recordings should be around -20 dB. To achieve this, the reference pink noise signal is processed by *ReplayGain* as outlined thus far, and the result is stored as ref_Vrms. For every audio track, the difference between the calculated *ReplayGain* value for the track and ref_Vrms indicates how much the signal should be scaled in order to match 83 dB perceived loudness. *ReplayGain* calculates overall perceived loudness, so this adjustment will cause all audio tracks to have the same overall perceived loudness as the reference pink noise signal.

There is one difference between the SMPTE calibration, and the *ReplayGain* calibration. The SMPTE system calibration uses a single channel of pink noise (reproduced through a single loudspeaker). However, music is typically played via two loudspeakers. Thus, though one

channel of pink noise is required to calibrate the system gain, the ideal loudness of the music is actually the loudness when two speakers are in use. For this reason, *ReplayGain* is calibrated to two channels of pink noise, since this is the average loudness that the music should match. In the implementation, a monophonic pink noise wavefile is employed, and *ReplayGain* automatically assumes that is will be replayed over two speakers, as it does any monophonic file.

K.3.4.1 Implementation

`ReplayGainScript.m` loads a .wav file containing -20 dB FS pink noise from disk, and processes this via `ReplayGain.m`, storing the result as a reference. The reference wavefile is included on the accompanying CD-ROM. The relevant lines of code are:

```
% Calculate perceived loudness of -20dB FS RMS pink noise
% This is the SMPTE RP 200 reference signal. It calibrates to:
% 0dB on a K-20 studio meter / mixing desk
% 83dB SPL in a listening environment
[ref_Vrms]=replaylevel('ref_pink.wav',a1,b1,a2,b2);
  ...
% Calculate the perceived loudness of the file using "replaygain" function
% Subtract this from reference loudness to give Replay Gain Adjustment relative to 83 dB reference
Vrms=ref_Vrms-replaygain(current_file,a1,b1,a2,b2);
```

This value is stored within the audio file header. It represents an adjustment relative to the SMPTE RP 200 standard.

K.3.5 Overall Implementation

The entire *ReplayGain* algorithm has been implemented in MATLAB. The required files are `ReplayGainScript.m` which batch processes .wav files, and `ReplayGain.m` which carries out most of the calculation. `EqualLoudFilt.m` is required to generate the equal loudness filter coefficients. All files are included on the accompanying CD-ROM.

K.4 Replay Gain Data Format

Three values must be stored in the Replay Gain header.

1. Peak signal amplitude
2. "Radio" = Radio Replay Gain Adjustment required to make all tracks equal loudness
3. "Audiophile" = Audiophile Replay Gain Adjustment required to give ideal listening loudness

If calculated on a track-by-track basis, *ReplayGain* yields (2). If calculated on a disc-by-disc basis, *ReplayGain* will usually yield (3), though this value may be more accurately determined by a human listener if required.

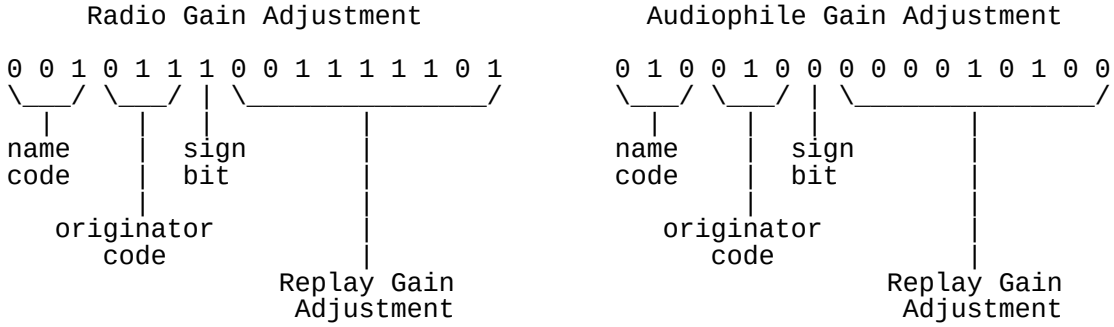
To allow for future expansion, if more than three values are stored, players should ignore those they do not recognise, but process those that they do. If additional Replay Gain Adjustments other than "Radio" and "Audiophile" are stored, they should come *after* "Radio" and "Audiophile". The Peak Amplitude must always occupy the first four bytes of the Replay Gain header frame. The three values listed above (or at least fields to hold the three values, should the values themselves be unknown) are **required** in all Replay Gain headers.

The Replay Gain Adjustment must be between -51.0dB and +51.0dB. Values outside this range must be limited to be within the range, though they are certainly in error, and should probably be re-calculated, or stored as "not set". For example, trying to cause a silent 24-bit file to play at 83 dB perceived loudness will yield a Replay Gain Adjustment of +57dB.

In practice, adjustment values from -23dB to +17dB are the likely extremes, and values from -18dB to +2dB are more usual.

K.4.1 Bit format

Each Replay Gain value should be stored in a Replay Gain Adjustment field consisting of two bytes (16 bits). Here are two example Replay Gain Adjustment fields:



In the above example, the Radio Gain Adjustment is -12.5dB, and was calculated automatically. The Audiophile Gain Adjustment is +2.0dB, and was set by the user. The binary codes are defined in the following sections.

K.4.1.1 Name code

000 =	not set
001 =	Radio Gain Adjustment
010 =	Audiophile Gain Adjustment
other =	reserved for future use

If space has been reserved for the Replay Gain in the file header, but no replay gain calculation has been carried out, then all bits (including the Name code) may be zero.

For each Replay Gain Adjustment field, if the name code = 000 (not set), then players should ignore the rest of that individual field.

For each Replay Gain Adjustment field, if the name code is an unrecognised value (i.e. not 001-Radio or 010-Audiophile), then players should ignore the rest of that individual field.

If no valid Replay Gain Adjustment fields are found (i.e. all name codes are either 000 or unknown), then the player should proceed as if the file contained no Replay Gain Adjustment information (see Section K.7).

K.4.1.2 Originator code

000 =	Replay Gain unspecified
001 =	Replay Gain pre-set by artist/producer/mastering engineer
010 =	Replay Gain set by user
011 =	Replay Gain determined automatically, as described in this paper
other =	reserved for future use

For each Replay Gain Adjustment field, if the name code is valid, but the Originator code is 000 (Replay Gain unspecified), then the player should ignore that Replay Gain Adjustment field.

For each Replay Gain Adjustment field, if the name code is valid, but the Originator code is unknown, then the player should **still** use the information within that Replay Gain Adjustment field. This is because, even if there is uncertainty as to how the adjustment was determined, *any* valid Replay Gain Adjustment is more useful than none at all.

If no valid Replay Gain Adjustment fields are found (i.e. all originator codes are 000), then the player should proceed as if the file contained no Replay Gain Adjustment information (see Section K.7).

K.4.1.3 Sign bit

0 = +

1 = -

K.4.1.4 Replay Gain Adjustment

The value, multiplied by ten, stripped of its sign (since the + or - is stored in the "sign" bit), is represented in 9 bits. e.g. -3.1dB becomes 31 = 000011111.

K.4.2 Default Value

$00 $00 (0000000000000000) should be used where no Replay Gain has been calculated or set. This value will be interpreted by players in the same manner as a file without a Replay Gain field in the header (see Section K.7).

The values of xxxyyy0000000000 (where xxx is any name code, and yyy is any originator code) are all valid, but indicate that the Replay Gain is to be left at 83 dB (0 dB Replay Gain Adjustment). These are **not** default values, and should only be used where appropriate (e.g. where the user, producer, or *ReplayGain* calculation has indicated that the correct Replay Gain **is** 83 dB).

K.4.3 Illegal Values

The values xxxyyy1000000000 are all illegal, since they represent negative zero. These values may be used to convey other information in the future. They must not be used at present. If encountered, players should treat them in the same manner as $00 $00 (the default value).

The value $xx $ff is **not** illegal, but it would give a false synch value within an mp3 file. If this value occurs, it is suggested the Replay Gain should be increased or decreased by 0.1 dB. This change will be inaudible, and it will avoid any possible problems due to false synchronisation.

K.5 Peak Amplitude Data Format

Scanning the file for the peak amplitude can be a time-consuming process. Therefore, it will be useful to store this single value within the file header. This can be used to determine if the required Replay Gain Adjustment will cause the file to clip.

K.5.1 Data Format

The maximum peak amplitude (a single value) should be stored as a 32-bit floating point number, where 1=digital full scale. This typically yields 22-bit accuracy, which is more than sufficient for the purpose.

K.5.2 Uncompressed Files

The maximum absolute sample value held in the file (on any channel) should be stored. The single sample value must be converted to a 32-bit float, such that digital full scale is equivalent to a value of 1.

K.5.3 Compressed files

Compressed audio does not exist as a waveform until it is decoded. Unfortunately, psychoacoustic coding of a heavily limited file can lead to sample values larger than digital full scale upon decoding. However, it is likely that such values will be brought back within range after scaling by the Replay Gain Adjustment. Even so, it is necessary to store the peak value of a compressed file as a 32-bit floating-point representation, where +/-1 represent digital full scale, and values outside this range would usually clip.

K.5.4 Implementation

For uncompressed files, the maximum values must be found and stored. For compressed files, the files must be decoded using a fully compliant decoder that allows peak overflows (i.e. has headroom), and the maximum value stored.

K.6 Replay Gain File Format

Each audio file format represents a unique situation. All audio files would benefit from the inclusion of Replay Gain information. Proposals for additional header information within two popular audio file formats are given below.

K.6.1 "mp3" file format

A suitable location for the Replay Gain header is within an ID3v2 tag, at the start of the mp3 file [Nilsson, WEB].[1] Note that all ID3v2 tag frames are written in Big Endian Byte order.

The ID3v2 standard defines a "tag" which is situated *before* the data in an mp3 file. The original ID3 (v1) tags reside at the end of the file, and contain six fields of information. The ID3v2 tags can contain virtually limitless amounts of information, and new "frames" within the tags may be defined.

K.6.1.1 Replay Gain Adjustment frame

Following the format of the ID3v2 standard document, a new frame is suggested, thus:

```
<Header for 'Replay Gain Adjustment', ID: "RGAD">
Peak Amplitude                          $xx $xx $xx $xx
Radio Replay Gain Adjustment            $xx $xx
Audiophile Replay Gain Adjustment       $xx $xx

Header consists of:
Frame ID                $52 $47 $41 $44 = "RGAD"
Size                    $00 $00 $00 $08
Flags                   $40 $00         (%01000000 %00000000)
```

In the RGAD frame, the flags state that the frame should be preserved if the ID3v2 tag is altered, but discarded if the audio data is altered.

K.6.2 "wav" file format

The RIFF WAV format is defined by [IBM and Microsoft, 1991]. NOTE: .wav files use little endian byte order.

The RIFF format is based around the concept of data chunks. Each chunk has the following basic format: NAME,SIZE,DATA. Wave files always contain a WAVEfmt chunk, and a data chunk. The WAVEfmt chunk contains information such as the number of channels, sampling rate, bits per sample etc etc. The data chunk contains the actual audio data. The entire file is contained within a RIFF chunk.

[1] A new tag format which will include the Replay Gain Adjustment information is proposed in [Van den Berghe, WEB], but is not discussed here.

There are several other optional chunks, which may be placed between the WAVEfmt chunk, and the data chunk. These include Cue and Playlist chunks.

The standard requires that software can successfully read .wav files containing unknown chunks. Should an unknown chunk NAME be encountered, then the accompanying SIZE field should be read, and the DATA field should be skipped. This means that new chunks can be defined without breaking compatibility with legacy software.

K.6.2.1 Replay Gain Adjustment chunk

Following the format of the standard document, the format is extended thus:

```
<WAVE-form> ->
       RIFF( 'WAVE'
            <fmt-ck>                // Format
            [<rgad-ck>]             // Replay Gain Adjustment
            [<fact-ck>]             // Fact chunk
            [<cue-ck>]              // Cue points
            [<playlist-ck>]         // Playlist
            [<assoc-data-list>]     // Associated data list
            <wave-data>   )         // Wave data

 <rgad-ck> ->   rgad( <rPeakAmplitude:REAL>
                      <wAudiophileRgAdjust:WORD>
                      <wRadioRgAdjust:WORD> )
```

Here is the definition in plain English. A new chunk is defined, named "rgad" (Replay Gain Adjustment). The size is eight bytes long. The first value stored is the Peak Amplitude (4-bytes); the second value stored is the Radio Replay Gain Adjustment (2-bytes); the third and final value stored is the Audiophile Replay Gain Adjustment (2-bytes).

EXAMPLE .WAV HEADER

For clarification, there follows an example .wav header, containing the new rgad chunk. This demonstrates the manner in which the "RIFF" chunk wraps around all the other chunks, and also how each chunk starts with its name, followed by its size. The new "rgad chunk" is highlighted in light blue.

Start Byte	Chunk	Chunk	title	contents	contents (HEX)	bytes	format
0	RIFF		name	"RIFF"	52 49 46 46	4	ASCII
4			size	176444	3C B1 02 00	4	uInt32
8		WAVE	name	"WAVE"	57 41 56 45	4	ASCII
12		fmt	name	"fmt "	66 6D 74 20	4	ASCII
16			size	16	10 00 00 00	4	uInt32
20			wFormatTag	1	01 00	2	uInt16
22			nChannels	2	02 00	2	uInt16
24			nSamplesPerSec	44100	44 AC 00 00	4	uInt32
28			nAvgBytesPerSec	176400	10 B1 02 00	4	uInt32
32			nBlockAlign	4	04 00	2	uInt16
34			nBitsPerSample	16	10 00	2	uInt16
36		rgad	name	"rgad"	72 67 61 64	4	ASCII
40			size	8	08 00 00 00	4	uInt32
44			fPeakAmplitude	1	00 00 80 3F	4	float32
48			nRadioRgAdjust	10822	46 2A	2	uInt16
50			nAudiophileRgAdjust	18999	37 4A	2	uInt16
52		data	name	"data"	64 61 74 61	4	ASCII
56			size	176400	10 B1 02 00	4	uInt32
60			waveform data	.....	.....	176400	Int16

In this example, the "rgad" chunk contains the following values:

- The "fPeakamplitude" value of 1 means that the digital data in the .wav file peaks at digital full scale (equivalent to -32768 for this 16-bit file).
- The "nRadioRgAdjust" value of 110822 (0010101001000110) represents a -7.0dB Radio Replay Gain Adjustment, user set.
- The "nAudiophileRgAdjust" value of 18999 (0100101000110111) represents a -5.5dB Audiophile Replay Gain Adjustment, user set.

See Section K.4 for an explanation of the calculation of each binary code.

The rest of the header information shows that this file is a standard 44.1kHz, 16bit, 2 channel wavefile (i.e. CD audio quality). It runs for 1 second, and is 176458 bytes long.

This extension to the .wav format has been tested for backwards compatibility using Winamp [Nullsoft, 2001] and Cool Edit Pro [Syntrillium, 2000]. Both pieces of software correctly play the .wav file, without incorrectly reading the new chunk as part of the audio data. Neither of the programs understand the "rgad" chunk, and both simply ignore it.

K.7 Player Requirements

The player requirements are outlined below, and discussed in detail in the following sections.

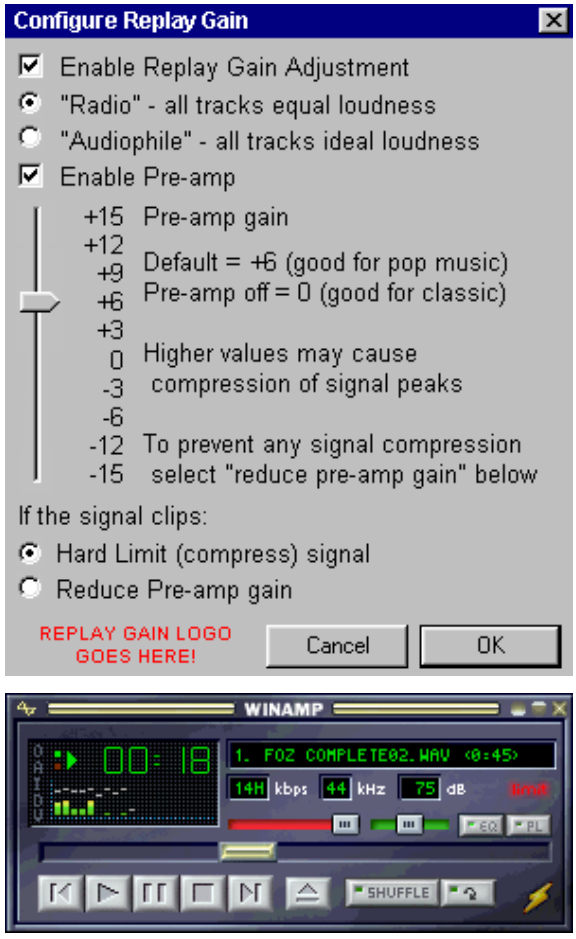


Figure K.8: Artist's Impression of a possible Replay Gain compatible player and control panel (adapted from [Nullsoft, 2001])

SCALE AUDIO BY REPLAY GAIN ADJUSTMENT VALUE

The Player reads the Replay Gain value, and scales the audio data as appropriate.

PRE-AMP

Most users who only play pop music will find that the level has been reduced too far for their liking. An **optional** boost of 6dB-12dB should be included by default, otherwise users will reject the entire concept. Knowledgeable users, or those playing classical music, will disable this. Some may even choose to *decrease* the level. For user friendliness, this part should be referred to as the "pre-amp".

CLIPPING PREVENTION

The player should, by default, apply hard limiting (NOT CLIPPING) to any signal peaks that would go over full scale after the above two operations. This should be user defeatable, so that audiophile users can choose to decrease the overall level to avoid clipping, rather than limiting the signal.

In practice, the first two adjustments can be carried out in a single step, where each sample is multiplied by a fixed amount. The clipping prevention is only required if the peak signal amplitude is above digital full scale *after* the first two steps.

Each stage of this processing will now be described in detail.

K.7.1 Scale audio by Replay Gain Adjustment value

K.7.1.1 Reading the Replay Gain

First, the player should determine whether the user requires "Radio" style level equalisation (all tracks same loudness), or "Audiophile" style level equalisation (all tracks "ideal" loudness). This option should be selectable in the Replay Gain control panel, and should default to "Radio".

Then the player should read the appropriate Replay Gain Adjustment value from the file header, and convert it back to its original dB value (see Section K.4, especially K.4.1.4, noting the factor of ten).

The player should also read (or calculate) the Peak amplitude. This is required for Clipping prevention, as described in Section K.7.3.

K.7.1.2 Scaling by the Replay Gain Adjustment

Changing the level of an audio signal simply means multiplying each sample value by a constant value. This constant is given by:

$$scale = 10^{(replay\_gain\_adjustment/20)}$$

Or, in words: ten to the power of (the Replay Gain Adjustment divided by 20).

After any such operation, it is good practice to dither the result. If this calculation and the preamp are implemented separately, then dither should only be added to the final result, just before the result is truncated back to 16 bits. The bit-depth of the output should be equal to that of the sound card, not the original file. For example, after Replay Gain Adjustment, an 8-bit file should be sent to a 16-bit soundcard at 16-bit resolution.

K.7.1.3 No Replay Gain information

Disabling Replay Gain control for tracks without Replay Gain information would cause these tracks to be louder than the others, thus bringing back the original problem. If neither "Radio" or "Audiophile" Gain Adjustments are set, or if the track does not contain Replay Gain information, then the player should use an average of the previous ten Replay Gain Adjustments. This represents the typical loudness of tracks in the users music collection, and is a **much** better estimate of the likely Replay Gain than 0 dB, or no adjustment at all.

If the file *only* contains one of the Replay Gain Adjustments (e.g. Audiophile) but the user has requested the other (Radio), then the player should use the one that is available (in this case, Audiophile).

K.7.2 Player Pre-amp

The SMPTE calibration level suggests that the average level of an audio track should be 20 dB below full scale. Some pop music is dynamically compressed to peak at 0 dB and average around –3 dB. Hence, when the Replay Gain is correctly set, the level of such tracks will be reduced by 17 dB. If the user is listening to a mixture of highly compressed and not compressed tracks, then Replay Gain Adjustment will make the listening experience more pleasurable, by bringing the level of the compressed tracks down into line with the others. However, if the user is *only* listening to highly compressed music, then they are likely to complain that all their files are now too quiet.

To solve this problem, a Pre-amp should be incorporated into the player. This should take the form of an adjustment to the scale factor calculated in the previous section. The default value should be +6 dB, though some manufacturers may choose +9, +12 or +15 dB. A positive value is chosen so that casual users will find little change to the loudness of their compressed pop music (except that the occasional "problem" quiet track will now be as loud as the rest), while power users and audiophiles can reduce the Pre-amp gain to enjoy all their music without dynamic compression.

If the Pre-amp gain is set to a high value whilst listening to classical music (or nicely produced pop music), then the peaks will be compressed. However, this is the consistent policy of most radio stations, and many listeners like this sound.

K.7.2.1 Implementation

If Replay Gain Adjustment is enabled, the player should read the user selected pre-amp gain, and scale the audio signal by the appropriate amount. For example, a +6 dB gain requires a scale of $10^{(6/20)}$, which is approximately 2. The Replay Gain and Pre-amp scale factors can be multiplied together for simplicity and ease of processing.

K.7.3 Clipping Prevention

The signal may clip for three reasons.

1. In coded audio (e.g. mp3 files) a file that was hard-limited to digital full scale *before* encoding will often be pushed over the limit by the psychoacoustic compression. A decoder with headroom can recover the signal above full scale by reducing the gain (for example [Leslie, WEB]). Typical decoders simply allow the audio to clip.

2. Replay Gain will reduce the level of loud dynamically compressed tracks, and increase the level of quiet dynamically uncompressed tracks. The average levels will then be similar, but the quiet tracks will have higher peaks. If the user pushes the pre-amp gain to maximum (which would take highly compressed pop music back to its original level), then the peaks of the (originally) quieter material will be pushed well over full scale.

3. If a track has a very wide dynamic range, then the Replay Gain Adjustment may instruct the player to increase the level of the track such that it will clip. This will occur for tracks with very low average energy, but very high peak amplitude. The author would be interested to hear of any recordings which cause clipping following Replay Gain Adjustment, with the pre-amp gain set at zero.

There are two solutions to the problem of clipping. In situation 2 above, the user apparently wants all the music to sound very loud. To give them their wish, any signal which would peak above digital full scale should be limited at just below digital full scale. This is also useful at lower pre-amp gains, where it allows the average level of classical music to be raised to that of pop music, without introduction clipping distortion. This could be useful when making tapes for use in an automobile. The exact type of limiting/compression is not defined here, but something similar to the Hard Limiter found in Cool Edit Pro [Syntrillium, 2000] may be appropriate, especially for popular music.

The audiophile user will not want any compression or limiting to be applied to the signal. In this case, the only option is to reduce the pre-amp gain (so that the scaling of the digital signal is lower than that suggested by the Replay Gain Adjustment). In order to maintain the consistency of level between tracks, the pre-amp gain should remain at this reduced level for subsequent tracks.

K.7.3.1 Implementation

If the Peak Level is stored in the header of the file, it is trivial to calculate whether the signal will clip following the Replay Gain Adjustment and Pre-amp gain. If the signal will not clip, then no further action is necessary. If the signal will clip, then either the hard limiter should be enabled, or the pre-amp gain should be reduced accordingly *before* playing the track.

K.7.4 Hardware Solution

The above steps are appropriate for software players operating on the digital signal in order to scale it. However, it is possible to send the digital signal to the DAC *without level correction*, and to place an attenuator in the analogue signal path. The attenuator should be driven by the Replay Gain Adjustment value. This implementation would maintain maximum signal to noise ratio in the digital to analogue conversion stage.

K.8 Typical Results

The result of the *ReplayGain* Radio adjustment is demonstrated by an audio example included on the accompanying CD-ROM. Replay Gain processing was applied to 25 audio tracks, and 5-10 seconds from each track were compiled into two files. The first file contains excerpts from all tracks at their original level, whilst the second file contains the same excerpts after Replay Gain correction, with 6 dB Pre-amp gain and Hard Limiting (i.e. suggested default player settings).

The titles of the tracks, and calculated Radio Replay Gains and Radio Replay Gain Adjustments, are shown in the following table. "Time" refers to the time at which an extract from that track appears in the audio example.

Time:	Artist/CD – Track:	Replay Gain:	Adjustment:
0.00	The Tampera – Feel It (CD single)	67.0849 dB	-15.9151 dB
0.04	Alanis Morissette - Head Over Feet	71.0121 dB	-11.9879 dB
0.10	Celine Dion – Falling into you	70.5880 dB	-12.4120 dB
0.19	Oasis - Wonderwall	68.0897 dB	-14.9103 dB
0.25	Telarc CD80251 - Goldfinger	75.3096 dB	-07.6904 dB
0.37	Telarc CD80183 - Moon River	87.9934 dB	+04.9934 dB
0.53	Telarc CD80221 - Brief Encounter	77.4630 dB	-05.5370 dB
1.15	Elvis - Fever	80.7907 dB	-02.2593 dB
1.23	Oleta Adams - Get Here	76.0916 dB	-06.9084 dB
1.32	HNF+RR test CD 2 - Track 01 (quiet speech)	92.8158 dB	+09.8158 dB
1.38	HFN+RR test CD 2 - Track 03 (pop music)	80.7753 dB	-02.2247 dB
1.46	HFN+RR test CD 2 - Track 19 (classical)	77.6900 dB	-05.3100 dB
1.54	HFN+RR test CD 2 - Track 20 (classical)	80.7789 dB	-02.2211 dB
2.01	HFN+RR test CD 2 - Track 22 (cannon)	101.1475 dB	+18.1475 dB
2.05	Faure - Pavane	77.8484 dB	-05.1516 dB
2.17	1812 Overture - Mercury Recording	70.5373 dB	-12.4627 dB
2.25	Radio show (music - Basement Jaxx)	70.7456 dB	-12.2544 dB
2.33	Radio show (DJ speech over music)	69.4228 dB	-13.5772 dB
2.41	SQAM CD - Track 50 (Male Speech)	75.0974 dB	-07.9026 dB
2.45	SQAM CD - Track 55 (Haydn Trumpet extract)	77.2643 dB	-05.7357 dB
2.53	SQAM CD - Track 56 (Handel Organ extract)	75.0690 dB	-07.9310 dB
3.04	SQAM CD - Track 57 (Bach Organ Extract)	91.5487 dB	+08.5487 dB
3.11	SQAM CD - Track 58 (Sarasate Guitar extract)	81.7892 dB	-01.2108 dB
3.19	SQAM CD - Track 59 (Ravel Violin extract)	80.8753 dB	-02.1247 dB
3.24	SQAM CD - Track 60 (Schubert Piano extract)	80.1139 dB	-02.8861 dB

Table K.1: Replay Gain values for a selection of audio tracks

K.8.1 Discussion

These tracks were processed by *ReplayGain* on an individual track-by-track basis, giving the "Radio" level equalisation. The intention is to make all the tracks sound equally loud, and this is achieved. There are only two problem tracks; the 1812 overture sounds too quiet when brought into line with the other tracks, and the Bach organ extract sounds ridiculous when this small organ is raised to the same loudness as the large pipe organ preceding it. However, these

problems arise because the radio equalisation is operating exactly as intended, rather than because of any fault in the algorithm itself. If the user wishes to hear the tracks at ideal, rather than equalised loudness, they should select the Audiophile Replay Gain Adjustment instead.

The hard limiting on the cannon shot (which was the only track on test which required hard limiting after Replay Gain processing and +6 dB Pre-amp gain) changes the sound significantly. Some listeners would dislike this effect, though others may be impressed by it. If the listener wishes to play tracks with such a large dynamic range without compression, their only option is to disable the hard limiting and pre-amp, as suggested for audiophile listeners.

Overall, the *ReplayGain* adjustment has corrected for the large differences in perceived loudness between the original audio tracks. This demonstrates that the algorithm operates correctly. User feedback from an initial release of this software has indicated that this process is very useful, and urgently sought by many users who are irritated by the problems described in the introduction.

K.9 Further work

The major obstacle facing this proposal is the lack of hardware and software support. At the time of writing, this proposal has been in the public domain for less than three months. However, even at this early stage, two programs support the Replay Gain standard: The latest release of the LAME open source mp3 encoder [Taylor, WEB] calculates the Radio and Audiophile Replay Gain Adjustments, and stores these values within a custom header added to every mp3 files it produces. In addition, a command line tool has been written to calculate and *apply* the Radio Replay Gain Adjustment directly to an mp3 file. This allows mp3 files to be replayed at equalised loudness using existing players without Replay Gain support. It is hoped that player support will follow shortly.

It is sincerely hoped that the acceptance of this proposal within the PC audio world will embarrass the audio industry into embracing a similar concept for consumer audio formats. Whilst the *ReplayGain* algorithm itself may be updated or superseded, the inclusion of meta data specifying the Replay Gain is long overdue, and deserves urgent consideration.

K.10 Conclusion

In this paper, it has been suggested that the ideal replay gain for every audio track should be stored as meta data within the track itself. The SMPTE RP 200 standard is suggested as the calibration reference. Two Replay Gain Adjustments have been defined relative to this reference; "Radio", which causes all tracks to be reproduced at the same loudness, and "Audiophile" which causes all tracks to the reproduced at an ideal loudness. An algorithm has been proposed for calculating the perceived loudness of an audio track, and hence the Radio Replay Gain Adjustment for that track. New header components have been suggested for two file formats in order to store the Replay Gain Adjustment values along side the audio data. The required behaviour of players wishing to implement the Replay Gain standard has been defined. Finally, the *ReplayGain* algorithm has been demonstrated to work as designed, successfully correcting unwanted loudness changes between discs.

K.11 Acknowledgements

Sincere thanks to Bob Katz, for inspiring this proposal with his papers on mastering, and for his suggestions regarding the two Replay Gain Adjustments. Many thanks to Glen Sawyer, who coded the first compiled version of *ReplayGain*. In doing so, he allowed others to use the algorithm, bringing it to a wider audience than the author's MATLAB scripts would have allowed.

References

This list of references contains some URL pointers to information available via the World Wide Web. The author is aware that information on the web can be of a transitory nature. For this reason, web references are backed up with appropriate written references, where such material is available. If a web reference is given first in the text, the web page presents material that is more recent or more detailed than is available in the printed literature at the time of writing. Where a web reference is given second in the text, the web page presents material that is readily available in the printed literature, but in a new or concise manner. For example, [Neely, WEB] includes animations of processes within the cochlea. Whilst these processes are discussed in several printed sources, the animation may aid the readers understanding.

Aladdin Systems (**WEB**).
StuffIt - The complete zip and sit compression solution.
http://www.stuffit.com/

Alcántara, J. I.; Moore, B. C. J.; and Vickers, D. A. (**2000**).
The relative role of beats and combination tones in determining the shapes of masking patterns at 2 kHz: I. Normal-hearing listeners.
Hearing Research, vol. 148, Oct., pp. 63-73.

Allen, J. B.; and Neely, S. T. (**1992**).
Micromechanical models of the cochlea.
Physics Today, vol. 45, July, pp. 40-47.

Anon. (**WEB**).
The Auditory Central Nervous System.
http://serous.med.buffalo.edu/hearing/

Ashland, M. T. (**WEB**).
Monkey's Audio - a fast and powerful lossless audio compressor
http://www.monkeysaudio.com/

Battaeu, D. W. (**1967**).
The role of the pinna in human localization.
Proceedings of the Royal Society of London, Series B, vol. 168, pp. 158-180.

Bauer, B.B. (**1961**).
Phasor analysis of some stereophonic phenomena.
Journal of the Acoustical Society of America, vol. 33, no. 11, Dec., pp. 1536-1539.

Beerends, J. G.; and Stemerdink, J. A. (**1992**).
A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation.
Journal of the Audio Engineering Society, vol. 40, no. 12, Dec., pp. 963-978.

Bernfield, B. (**1973**).
Attempts for better understanding of the directional stereophonic listening mechanism.
preprint C-4, presented at the *44th convention of the Audio Engineering Society in Rotterdam,* Feb.

Bernstein, R. S.; and Raab, D. H. (**1990**).
The effect of bandwidth on the detectability of narrow- and wideband signals.
Journal of the Acoustical Society of America, vol. 88, no. 5, Nov., pp. 2115-2125.

Boerger, G. (**1965**).
Die Lokalisation von Gausstönen
Dissertation, Technische Universität, Berlin.

Bosi, M.; Brandenburg, K.; Quackenbush, S.; Fielder, L.; Akagiri, K.; Fuchs, H.; and Dietz, M. (**1997**).
ISO/IEC MPEG-2 Advanced Audio Coding.
Journal of the Audio Engineering Society, vol. 45, no. 10, Oct., pp. 789·814.

Bower, A. J. (**1998**).
DIGITAL RADIO – The Eureka 147 DAB System.
Electronic Engineering, April, pp. 55-56.

Brandenburg, K. (**1999**).
MP3 and AAC Explained.
paper 17-009, presented at the *17th International Conference of the Audio Engineering Society: High-Quality Audio Coding*; Sept.

Brandenburg, K.; and Bosi, M. (**1997**).
Overview of MPEG Audio: Current and Future Standards for Low Bit-Rate Audio Coding.
Journal of the Audio Engineering Society, vol. 45, no. 1/2, pp. 4-21.

Brungart, D. S.; Durlach, N. I.; and Rabinowitz, W. M. (**1999**).
Auditory localization of nearby sources. II. Localization of a broadband source.
Journal of the Acoustical Society of America, vol. 106, no. 4, pt. 1, Oct., pp. 1956-1968.

Buschmann, A. (**WEB**).
Information on lossy audiocoding and the MPEG*plus*-project.
http://www.stud.uni-hannover.de/~andbusch/audiocoder_eng.html

Carlile, S.; and Wardman, D. (**1996**).
Masking produced by broadband noise presented in virtual auditory space.
Journal of the Acoustical Society of America, vol. 100, no. 6, Dec., pp. 3761-3768.

Carterette, E. C. (**1978**).
Some Historical Notes on Research in Hearing.
in Carterette, E.C.; and Friedman, M. P., Eds., *Handbook of Perception, Volume IV, Hearing* (Academic Press, New York). Chapter 1, pp. 3-34.

Colburn, H. S. (**1973**).
Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination.
Journal of the Acoustical Society of America, vol. 54, no. 6, pp. 1458-1470.

Colburn, H. S. (**1977**).
Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise.
Journal of the Acoustical Society of America, vol. 61, no. 2, Feb., pp. 525-533.

Colburn, H. S.; and Durlach, N. I. (**1965**).
Time-intensity relations in binaural unmasking.
Journal of the Acoustical Society of America, vol. 38, pp. 93-103.

Colburn, H. S.; and Durlach, N. I. (**1978**).
Models of Binaural Interaction.
in Carterette, E. C.; and Friedman, M. P., Eds., *Handbook of Perception, Volume IV, Hearing*. (Academic Press, New York). Chapter 11, pp. 467-518.

Colomes, C.; Lever, M.; Rault, J. B.; and Dehery, Y. F. (**1995**).
A Perceptual Model Applied to Audio Bit-Rate Reduction.
Journal of the Audio Engineering Society, vol. 43, no. 4, April, pp. 233-240.

Couvreur, C. (**1997**).
Environmental Sound Recognition: A Statistical Approach
PhD Thesis, Faculte Polytechnique de Mons, Mons, Belgium, June 1997.

Culling, J. F.; and Summerfield, Q. (**1998**).
Measurements of the Binaural temporal window using a detection task
Journal of the Acoustical Society of America, vol. 103, no. 6, June, pp. 3540-3533.

Dallos, P. (**1978**).
Biophysics of the Cochlea.
in Carterette, E. C.; and Friedman, M.P., Eds, *Handbook of Perception, Volume IV, Hearing* (Academic Press, New York). Chapter 4, pp. 125-162.

Dau, T.; Püschel, D.; and Kohlrausch, A. (**1996a**).
A quantitative model of the "effective" signal processing in the auditory system. I. Model structure.
Journal of the Acoustical Society of America, vol. 99, no. 6, June, pp. 3615-3622.

Dau, T.; Püschel, D.; and Kohlrausch, A. (**1996b**).
A quantitative model of the "effective" signal processing in the auditory system. II. Simulations and measurements.
Journal of the Acoustical Society of America, vol. 99, no. 6, June, pp. 3623-3631.

Dietz, M.; Herre, J.; Teichmann, B.; and Brandenburg, K. (**1997**).
Bridging the Gap: Extending MPEG Audio Down to 8 kbit/s.
preprint 4508, presented at the *102nd convention of the Audio Engineering Society*, March 1997

Doll, T. J.; and Hanna, T. E. (**1995**).
Spatial and Spectral Release from Masking in Three-Dimensional Auditory Displays.
Human Factors, vol. 37, Feb., pp. 341-355.

Durlach, N. I.; and Colburn, H. S. (**1978**).
Binaural Phenomena.
in Carterette, E. C.; and Friedman, M.P., Eds, *Handbook of Perception, Volume IV, Hearing* (Academic Press, New York). Chapter, 10 pp. 365-466.

EBU (**1988**).
SQAM: Sound quality assessment material – Recordings for subjective tests.
Compact Disc. Cat No 422 204-2. Geneva: *European Broadcasting Union.*

EBU (**2000**).
MUSHRA – EBU Method for the Subjective Listening Tests of Intermediate Audio Quality.
Draft EBU Recommendation B/AIM 022 (Rev.8)/BMC 607rev, January. Geneva: *European Broadcasting Union.*

Ehret, G.; and Romand, R. (**1997**).
The Central Auditory System.
(Oxford University Press, New York).

Fantini, D. (**1996**).
Perception of Complex Sounds, Speech and Music.
PS454 course notes, Department of Psychology, University of Essex.

Farrow, J. M. (**WEB**).
The Complete Works of Shakespeare.
http://www.cs.usyd.edu.au/~matty/Shakespeare/

Foo, K. C. K.; Hawksford, M. O. J.; and Hollier, M. P. (**1998**).
Three-Dimensional Sound Localization with Multiple Loudspeakers using a Pair-Wise Association Paradigm and Embedded HRTFs.
preprint 4745, presented at the *104th convention of the Audio Engineering Society in Amsterdam,* May.

Gardner, B.; and Martin, K.D. (**1994**).
HRTF Measurements of a KEMAR Dummy-Head Microphone.
Perceptual Computing Technical Report #280, MIT Media Lab Machine Listening Group, also http://sound.media.mit.edu/KEMAR.html

Gehr, S. E.; and Sommers, M. S. (**1999**).
Age difference in backward masking.
Journal of the Acoustical Society of America, vol. 106, no. 5, Nov., pp. 2793-2799.

Gerzon, M. A. (**1991**).
Problems of Error-Masking in Audio Data Compression Systems.
preprint 3013, presented at the *90th convention of the Audio Engineering Society in Paris*, Feb.

Gerzon, M. A.; Craven, P. G.; Stuart, J. R.; Law, M. J.; and Wilson, R. J. (**1999**).
The MLP Lossless Compression System
paper 17-006I, presented at the *Audio Engineering Society 17th International Conference: High-Quality Audio Coding*, September 1999 .

Giguère, C.; Smoorenburg, G. F.; and Kunov, H. (**1997**).
The generation of psychoacoustic combination tones in relation to two-tone suppression effects in a computational model.
Journal of the Acoustical Society of America, vol. 102, no. 5, pt. 1, Nov., pp. 2821-2830.

Gilkey, R. H.; and Good, M. D. (**1995**).
Effects of Frequency on Free-Field Masking.
Human Factors, vol. 37, April, pp. 835-843.

Goldstein, J. L. (**1967**).
Auditory Nonlinearity.
Journal of the Acoustical Society of America, vol. 41, no. 3, pp. 676-689.

Grantham, D. W. (**1995**).
Spatial Hearing and Related Phenomena.
in Moore, B. C. J., Ed., *Hearing* (Academic Press, London).

Grantham, D. W.; and Wightman, F. L. (**1979**).
Detectability of a pulsed tone in the presence of a masker with time-varying interaural correlation.
Journal of the Acoustical Society of America, vol. 65, pp. 1509-1517.

Green, D. M. (**1976**).
An Introduction to Hearing.
(Lawrence Arlbaum, Hillsdale, NJ)

Green, D. M.; McKey, M. J.; and Licklider, J. C. R. (**1959**).
Detection of a Pulsed Sinusoid in Noise as a Function of Frequency.
Journal of the Acoustic Society of America, vol. 31, pp. 1446-1452.

Green, D. M.; and Swets, J. A. (**1966**).
Signal Detection Theory and Psychophysics.
(Wiley, New York).

Hall, J. L. (**1972**).
Auditory distortion products f2-f1 and 2f1-f2.
Journal of the Acoustical Society of America, vol. 51, pp. 1863-1871.

Hammershøi, D.; Møller, H.; and Sørensen, M. F. (**1992**).
Head-Related Transfer Functions: Measurements on 40 Human Subjects.
preprint 3289 presented at the *92nd Convention of the Audio Engineering Society* in Vienna, March.

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

Hollier, M. P. (**1996**).
Data Reduction – A series of 3 lectures.
Course notes, M.Sc. Audio Systems Engineering, University of Essex.

Hollier, M. P.; Hawksford, M. O. J.; and Guard, D. R. (**1993**).
Characterization of Communications Systems Using a Speechlike Test Stimulus.
Journal of the Audio Engineering Society, vol. 41, no. 12, Dec., pp. 1008-1021.

Hollier, M. P.; Hawksford, M. O. J.; and Guard, D. R. (**1995**).
Algorithms for Assessing the Subjectivity of Perceptually Weighted Audible Errors.
Journal of the Audio Engineering Society, vol. 43, no. 12, Dec., pp. 1041-1045.

Hollier, M. P.; and Cosier, G. (**1996**).
Assessing human perception.
BT Technology Journal, vol. 14, Jan., pp. 206-215.

IBM and Microsoft (**1991**).
Multimedia Programming Interface and Data Specifications 1.0
Issued as a joint design by IBM Corporation and Microsoft Corporation, August 1991.
http://www.seanet.com/Users/matts/riffmci/riffmci.htm

Irino, T.; and Patterson, R. D. (**1997**).
A time-domain, level-dependent auditory filter: The gammachrip.
Journal of the Acoustical Society of America, vol. 101, no. 1, Jan., pp. 412-419.

IEC/CD 1672 (**1996**).
Electroacoustics - Sound Level Meters
International Engineering Consortium, Geneva, Nov. 1996.

IEC 60027-2 (**2000**).
Letter symbols to be used in electrical technology – Part 2: Telecommunications and electronics. Second Edition.
International Engineering Consortium, Geneva, Nov. 2000. see also
http://physics.nist.gov/cuu/Units/binary.html

ISO/IEC 11172-3 (**1993**).
Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio.
Geneva: International Organisation for Standardization.

ISO/IEC 13818-3 (**1998**).
Information technology – Generic coding of moving pictures and associated audio information – Part 3: Audio
Geneva: International Organisation for Standardization.

ISO/IEC 13818-7 (**1997**).
Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)
Geneva: International Organisation for Standardization.

ISO 389-7 (**1996**).
Acoustics - Reference zero for the calibration of audiometric equipment.
in *Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions.* Geneva: International Organisation for Standardization, 1996.

ITU-R BS.468-4 (**1986**).
Measurement of audio-frequency noise voltage level in sound broadcasting.
Geneva: International Telecommunication Union.

ITU-R BS.1116-1 (**1997**).
Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems
Geneva: International Telecommunication Union.

ITU-R BS.1284 (**1997**).
Methods for the subjective assessment of sound quality - General requirements.
Geneva: International Telecommunication Union.

ITU-R BS.[DOC. 6/106] (**2001**).
Method for the subjective assessment of intermediate audio quality.
Draft new recommendation from Radiocommunication Study Group 6. *Geneva: International Telecommunication Union.*

Jeffress, L. A. (**1948**).
A place theory of sound localization.
Journal of Comparative and Physiological Psychology, vol. 41, pp. 35-39.

Johnson, D. H. (**WEB**).
Auditory Neuroscience: How is Sound Processed by the Brain?
http://www-ece.rice.edu/~dhj/neuro.html

Johnston, J. D. (**1988a**).
Estimation of Perceptual Entropy Using Noise Masking Criteria.
ICASSP, A1.9, pp 2524-2527.

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

Katz, B. (**2000**).
Integrated Approach to Metering, Monitoring, and Leveling Practices, Part 1: Two-Channel Metering
Journal of the Audio Engineering Society, vol. 48, no. 9, Sept., pp. 800-809. See also http://www.digido.com/integrated.html

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

Kemp, D. T. (**1980**).
Towards a model for the origin of Cochlea echoes.
Hearing Research, vol. 2, pp. 533-548.

Kidd, G.; Mason, C. R.; Rohtla, T. L.; and Deliwala, P. S. (**1998**).
Release from masking due to spatial separation of sources in the identification of non-speech auditory patterns.
Journal of the Acoustical Society of America, vol. 104, no. 1, July, pp. 422-431.

Klumpp, R. G.; and Eady, H. R. (**1956**).
Some Measurements of Interaural Time difference Thresholds.
Journal of the Acoustical Society of America, vol. 28, no. 5, Sept., pp. 859-860.

Leslie, R. (**WEB**).
MAD: MPEG Audio Decoder
http://www.mars.org/home/rob/proj/mpeg/

Levitt, H. (**1971**).
Transformed Up-Down Methods in Psychoacoustics.
Journal of the Acoustical Society of America, vol. 49, no. 2, pt. 2, pp. 467-477.

Liebchen, T. (**WEB**).
LPAC - Lossless Predictive Audio Compression.
http://www-ft.ee.tu-berlin.de/~liebchen/lpac.html

Macpherson, E. A. (**1989**).
A Computer Model of Binaural Localization for Stereo Imaging Measurement.
Presented at the 87th Convention of the Audio Engineering Society, preprint 2866.

Macpherson, E. A. (**1991**).
A Computer Model of Binaural Localization for Stereo Imaging Measurement.
Journal of the Audio Engineering Society, vol. 39, no. 9, Sept., pp. 604-622.

McFadden, D. (**1968**).
Masking-level differences determined with and without interaural disparities in masker intensity.
Journal of the Acoustical Society of America, vol. 44, pp. 212-223.

Mears, D.; Watanabe, K.; and Scheirer, E. (**1998**).
Report on the MPEG-2 AAC Stereo Verification Tests.
ISO/IEC JTC1/SC29/WG11, N2006, Feb.

Meddis, R. (**1986**).
Simulation of mechanical to neural transduction in the auditory receptor.
Journal of the Acoustical Society of America, vol. 79, no. 3, March, pp. 702-711.

Meddis, R. (**1988**).
Simulation of auditory-neural transduction: Further studies.
Journal of the Acoustical Society of America, vol. 83, no. 3, March, pp. 1056-1063.

Meddis, R.; Hewitt, M. J.; and Shackleton, T. M. (**1990**).
Implementation details of a computation model of the inner hair-cell/auditory-nerve synapse.
Journal of the Acoustical Society of America, vol. 87, no. 4, April, pp. 1813-1816.

Mehrgardt, S.; and Mellert, V. (**1977**).
Transformation characteristics of the external human ear.
Journal of the Acoustical Society of America, vol. 61, no. 6, pp. 1567-1576.

Meilgaard, M. C.; Civille, G. V.; and Carr, T. B. (**1999**).
Sensory Evaluation Techniques, third edition.
(CRC Press, Boca Raton)

Mills, A.W. (**1958**).
On the Minimum Audible Angle.
Journal of the Acoustical Society of America, vol. 30, no. 4, April, pp. 237-246.

Mills, A.W. (**1972**).
Auditory localization.
in Tobias, J. V., Ed., *Foundations of Modern Auditory Theory, Vol. II* (Academic Press, New York).

Miyaguchi, D. (**WEB**).
Group Listening Tests of Various Formats at 128 kbit/s
http://fastforward.iwarp.com/128tests.html

Møller, H.; Hammershøi, D.; Jensen, C. B.; and Sørensen, M. F. (**1995a**).
Transfer Characteristics of Headphones Measured on Human Ears.
Journal of the Acoustical Society of America, vol. 43, no. 4, April, pp. 203-217.

Møller, H.; Sørensen, M. F.; Hammershøi, D.; and Jensen, C. B. (**1995b**).
Head-Related Transfer functions of Human Subjects.
Journal of the Audio Engineering Society, vol. 43, no. 5, May, pp. 300-321.

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

Moore, B. C. J.; Alcántara, J. I.; and Dau, T. (**1998**).
Masking patterns for sinusoidal and narrow-band noise maskers.
Journal of the Acoustical Society of America, vol. 104, no. 2, pt. 1, Aug., pp. 1023-1038.

Moore, B. C. J.; and Glasberg, B. R. (**1987**).
Formulae describing frequency selectivity as a function of frequency and level and their use in calculating excitation patterns.
Hearing Research, vol. 28, pp. 209-225.

Moore, B. C. J.; Glasberg, B. R.; and Baer, T. (**1997**).
A Model for the Prediction of Thresholds, Loudness, and Partial Loudness.
Journal of the Acoustical Society of America, vol. 45, no. 4, April, pp. 224-240.

Moore, G. E. (**1965**).
Cramming More Components onto Integrated Circuits
Electronics, vol. 38, April, pp. 114-117.

Nandy, D.; and BenArie, J. (**1996**).
An auditory localization model based on high-frequency spectral cues.
Annals of Biomedical Engineering, vol. 24, No. 6, pp. 621-638.

Neely, S. T. (**WEB**).
Cochlea Mechanics.
Communication Engineering Laboratory, Boys Town National Research Hospital.
http://www.boystown.org/Btnrh/cel/cochmech.htm

Nilsson, M. (**WEB**).
ID3v2 informal standard – The audience is informed.
http://www.id3v2.org/

Nullsoft (**2001**).
Winamp audio player v2.73
Nullsoft inc, 375 Alabama St. Ste. 350, San Francisco, CA, USA
http://www.winamp.com/

Osman, E. (**1971**).
A correlation model of binaural masking level differences.
Journal of the Acoustical Society of America, vol. 50, pp. 1494-1511.

Paillard, B.; Mabilleau, P.; Morissette, S.; and Soumagne, J. R. (**1992**).
PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals.
*Journal of the Audio Engineering Societ*y, vol. 40, no. 1/2, Jan./Feb., pp. 21-31.

Patterson, R. D.; Allerhand, M. H.; and Giguère, C. (**1995**).
Time-domain modeling of peripheral auditory processing: amodular architecture and a software platform.
Journal of the Acoustical Society of America, vol. 98, no. 4, Oct., pp. 1890-1894.

Patterson, R. D.; Holdsworth, J.; and Allerhand, M. (**1992a**).
Auditory models as preprocessors for speech recognition.
in Shouten, M. E. H., Ed., *The Auditory Processing of Speech: From the Auditory Periphery to words* (Mouton de Gruyter, Berlin), pp. 67-83.

Patterson, R.D.; Robinson, K.; Holdsworth, J.; McKeown, D.; Zhang, C.; and Allerhand, M. (**1992b**).
Complex sounds and auditory images.
in Cazals, Y.; Demany, L.; and Horner, K., Eds., *Auditory Physiology and Perception*, (Pergamon, Oxford), pp. 429-446.

Patterson, R.D. (**WEB**).
Description of Research.
Centre for Neural Basis of Hearing, Physiology Department, University of Cambridge.
<http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/cnbh/>

Patterson, R.D. (**WEB-1**).
The auditory pathway from the cochlea to cortex.
Centre for Neural Basis of Hearing, Physiology Department, University of Cambridge.
<http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/cnbh/audpath.html>

Patterson, R.D. (**WEB-2**).
From the cochlea to the IC: The Cochlea
Centre for Neural Basis of Hearing, Physiology Department, University of Cambridge.
<http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/cnbh/AIC_C_txt.html>

Patterson, R.D. (**WEB-3**).
From the cochlea to the IC: Cochlear Nucleus & Superior Olivary Complex.
Centre for Neural Basis of Hearing, Physiology Department, University of Cambridge.
<http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/cnbh/AIC_CN_SOC_txt.html>

Penner, M. J. (**1977**).
Detection of temporal gaps in noise as a measure of the decay of auditory sensation.
Journal of the Acoustical Society of America, vol. 61, no. 2, Feb., pp. 552-557.

Perrott, D. R.; and Pacheco, S. (**1989**).
Minimum audible angle thresholds for broadband noise as a function of delay between the onset of the lead and lag signals.
Journal of the Acoustical Society of America, vol. 85, no. 6, June, pp. 2669-2672.

PKWARE (**WEB**).
Genuine PKZIP Products.
http://www.pkware.com/

Plack, C.J.; and Carlyon, R. P. (**1995**).
Loudness Perception and Intensity Coding.
in B. C. J. Moore, Ed., *Hearing* (Academic Press, San Diego, California), pp. 123-160.

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

Probst, R.; Lonsbury-Martin, B. L.; and Martin, G. K. (**1991**).
A review of otoacoustic emissions.
Journal of the Acoustical Society of America, vol. 89, no. 5, May, pp. 2027-2067.

Puria, S.; Peake, W.T.; and Rosowski, J.J. (**1997**).
Sound-pressure measurements in the cochlear vestibule of human-cadaver ears.
Journal of the Acoustical Society of America, vol. 101, no. 5, pt. 1, May, pp. 2754-2770.

Püschel, D. (**1988**).
Prinzipien der zeitlichen Analyse beim Hören.
Ph.D. thesis, University of Göttingen.

Putland, G. (**1994**).
Acoustical properties of air versus temperature and pressure.
Journal of the Audio Engineering Society, vol. 42, no. 11, Nov., pp. 927-933.

Reed, C. M.; and Bilger, R. C. (**1973**).
A comparative study of S/N_0 and E/N_0.
Journal of the Acoustical Society of America, vol. 53, no. 4, April, pp. 1039-1044.

Rimell, A.; and Hawksford, M. O. (**1996**).
From the Cone to the Cochlea: Modelling the Complete Acoustical Path.
preprint 4240, presented at the *100th Convention of the Audio Engineering Society* in Copenhagen, May.

Rimell, A. (**1996**).
Psychoacoustic foundations
in *Reduction of loudspeaker polar response aberrations through the application of psychoacoustic error concealment*, PhD thesis, Department of Electronic Systems Engineering, University of Essex.

Robert, A.; and Eriksson, J. L. (**1999**).
A composite model of the auditory periphery for simulating responses to complex sounds.
Journal of the Acoustical Society of America, vol. 106, no. 4, pt. 1, Oct., pp. 1852-1864.

Robinson, D. W.; and Dadson, R. S. (**1956**).
A re-determination of the equal-loudness relations for pure tones.
British Journal of Applied Physics, vol. 7, May, pp. 166-177.

Robinson, D. J. M.; and Greenwood, R. G. (**1998**).
A Binaural simulation which renders out of head localisation with low cost digital signal processing of Head Related Transfer Functions and pseudo reverberation.
preprint 4723, presented at the *104th Convention of the Audio Engineering Society in Amsterdam*, May.

Robinson, D. J. M.; and Hawksford, M. O. J. (**1999**).
Time-domain auditory model for the assessment of high-quality coded audio.
preprint 5017, presented at the *107th Convention of the Audio Engineering Society, New York*, Sept.

Robinson, D. J. M.; and Hawksford, M. O. J. (**2000**).
Psychoacoustic models and non-linear human hearing.
preprint 5228, presented at the *109th Convention of the Audio Engineering Society, Los Angeles*, Sept. Reproduced in Appendix x.

Rosen, S.; and Howell, P. (**1991**).
Signals and Systems for Speech and Hearing.
(Academic Press, London and San Diego).

Saberi, K.; Dostal, L.; Sadralodabai, T.; Bull, V.; and Perrott.; D. R. (**1991**).
Free-field release from masking.
Journal of the Acoustical Society of America, vol. 90, no. 3, Sept., pp 1355-1370.

Santon, F. (**1987**).
"Détection d'un son pur dans un bruit masquant suivant 'angle d'incidence du bruit. Relation avec le seuil de réception de la parole." (Detection of a Pure Sound in the Presence of Masking Noise, and its Dependence on the Angle of Incidence of the Noise).
Acustica, vol. 63, pp. 222-228.

Sayers, B. M.; and Cherry, E. C. (**1957**).
Mechanism of binaural fusion in the hearing of speech.
Journal of the Acoustical Society of America, vol. 29, pp. 973-987.

Schroeder, M. R.; Atal, B. S.; and Hall, J. L. (**1979**).
Optimizing digital speech coders by exploiting masking properties of the human ear.
Journal of the Acoustical Society of America, vol. 66, pp. 1647-1652.

Schubert, E. D. (**1978**). "History of Research on Hearing."
in Carterette, E. C.; and Friedman, M.P., Eds, *Handbook of Perception, Volume IV, Hearing* (Academic Press, New York), pp. 41-80.

Singleton, H. (**WEB**).
A-, B-, and C-Weighting functions
http://www.cross-spectrum.com/audio/newgifs/weighting_1.gif

Smoorenburg, G. F. (**1972**).
Combination tones and their origin.
Journal of the Acoustical Society of America, vol. 52, no. 2, pt. 2, pp. 615-632.

SMPTE RP 200 (**1999**).
Relative and Absolute Sound Pressure Levels for Motion-Picture Multichannel Sound Systems.
Society of Motion Picture and Television Engineers, Recommended Practices document.

Snedecor, G. W.; and Cochran, W. G. (**1989**).
Statistical Methods, Eighth Edition.
(Iowa State University Press).

Soderquist, D. R.; and Lindsey, W. L. (**1972**).
Physiological noise as a masker of low frequencies: The cardiac cycle.
Journal of the Acoustic Society of America, vol. 52, no. 4, pp. 1216-1220.

Soulodre, G. A.; Grusec, T.; Lavoie, M.; and Thibault, L. (**1998**).
Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs.
Journal of the Audio Engineering Society, vol. 46, no. 3, Mar., pp. 164-177.

Stevens, S. S. (**1935**).
The relation of pitch to intensity.
Journal of the Acoustical Society of America, vol. 6, pp. 150-154.

Stoll, G.; and Kozamernik, F. (**2000**).
EBU Listening tests on internet audio codecs.
EBU Technical Review, no. 283, June.

Strube, H. W. (**1985**).
A computationally efficient basilar-membrane model.
Acustica, vol. 58, pp. 207-214.

Stuart, J. R. (**1990**).
High Quality Digital Audio.
*Proceedings of the Institute of Acoustic*s, vol. 12, pp. 1-15.

Syntrillium (**2000**).
Cool Edit Pro version 1.2a
Syntrillium Software Corporation, P.O. Box 62255, Phoenix, AZ 85082-2255, USA
http://www.syntrillium.com/

Taylor, M. (**WEB**).
Lame – Lame Ain't an Mp3 Encoder
http://www.mp3dev.org/mp3/

Terhardt, E. (**1979**).
Calculating Virtual Pitch.
Hearing Research, vol. 1, pp. 318-362.

Thiede, T.; Treurniet, W. C.; Bitto, R.; Schmidmer, C.; Sporer, T.; Beerends, J. G.; Colomes, C.; Keyhl, M.; Stoll, G.; Brandenburg, K.; and Feiten, B. (**2000**).
PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality.
Journal of the Audio Engineering Society, vol. 48, no. 1/2, Jan./Feb., pp. 3-29.

Theiss, B.; and Hawksford, M. O. J. (**1999**).
Auditory Model for Lateralization and the Precedence Effect
preprint 5048, presented at the *107th Convention of the Audio Engineering Society, New York*, Sept.

Traunmüller, H. (**1990**).
Analytical expressions for the tonotopic sensory scale.
Journal of the Acoustical Society of America, vol. 88, no. 4, July, pp. 97-100.

Van den Berghe, R. (**WEB-1**).
Archive Quality Listening test
http://users.belgacom.net/gc247244/extra/AQ.htm (see also appendix x)

Van den Berghe, R. (**WEB-2**).
LAME Tag rev 0 specification - draft 10
http://users.belgacom.net/gc247244/extra/tag.html

von Békésy, G. (**1960**).
Experiments in hearing.
(McGraw-Hill, New York).

Winzip Computing (**WEB**).
Winzip Home Page.
http://www.winzip.com/

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

Young, H. D. (**2000**).
University Physics – tenth edition.
(Addison-Wesley).

Zhou, B. (**1995**).
Auditory Filter Shapes at High-Frequencies.
Journal of the Acoustical Society of America, vol. 98, no. 4, Oct., pp. 1935-1942.

Zwicker, E. (**1961**).
Subdivision of the audible frequency range into critical bands (Frequenzgruppen).
Journal of the Acoustical Society of America, vol. 33, no. 3, p. 248.

Zwicker, E. (**1979**).
Different behaviour of quadratic and cubic difference tones.
Hearing Research, vol. 1, pp. 283-292.

Zwicker, E. (**1981**).
Formulae for calculating the psychoacoustical excitation level of aural difference tones measured by the cancellation method.
Journal of the Acoustical Society of America, vol. 69, no. 5, May, pp. 1410-1413.

Zwicker, E. (**1984**).
Dependence of post-masking on masker duration and its release to temporal effect in loudness.
Journal of the Acoustical Society of America, vol. 75, no. 1, Jan., pp. 219-223.

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

Zwicker, E.; and Fastl, H. (**1990**).
Psychoacoustics – Facts and Models.
(Springer, Berlin).

Zwicker, E.; and Feldtkeller, R. (**1967**).
Das Ohr als Nachrichtenempfänger.
(Hizel Verlag, Stuttgart).

Zwicker, E.; and Terhardt, E. (**1980**).
Analytical expressions for critical-band rate and critical bandwidth as a function of frequency.
Journal of the Acoustical Society of America, vol. 68, no. 5, Nov., pp. 1523-1525.

Zwicker, E.; and Zwicker, U. T. (**1991**).
Audio Engineering and Psychoacoustics: Matching signals to the Final receiver, the Human Auditory System.
Journal of the Audio Engineering Society, vol. 39, no. 3, March, pp. 115-126.

Zwislocki, J.; and Feldman, R. S. (**1956**).
Just Noticeable Differences in Dichotic Phase
Journal of the Acoustical Society of America, vol. 28, no. 5, Sept., pp. 860-864.