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].

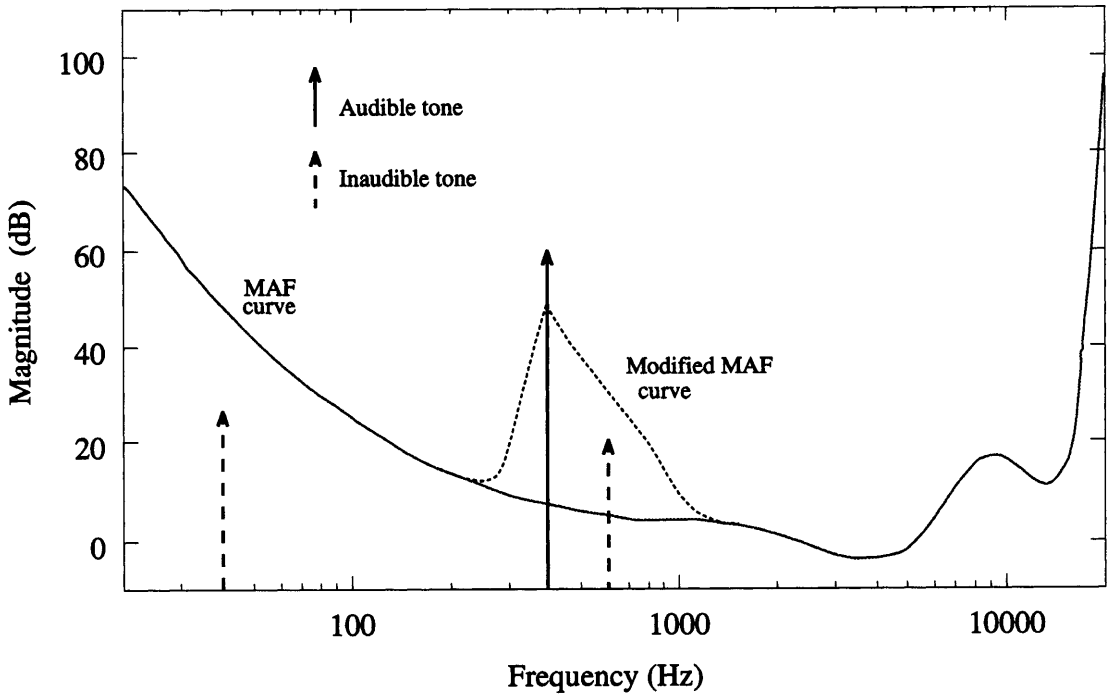

Figure 2.1: Spectral masking

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.

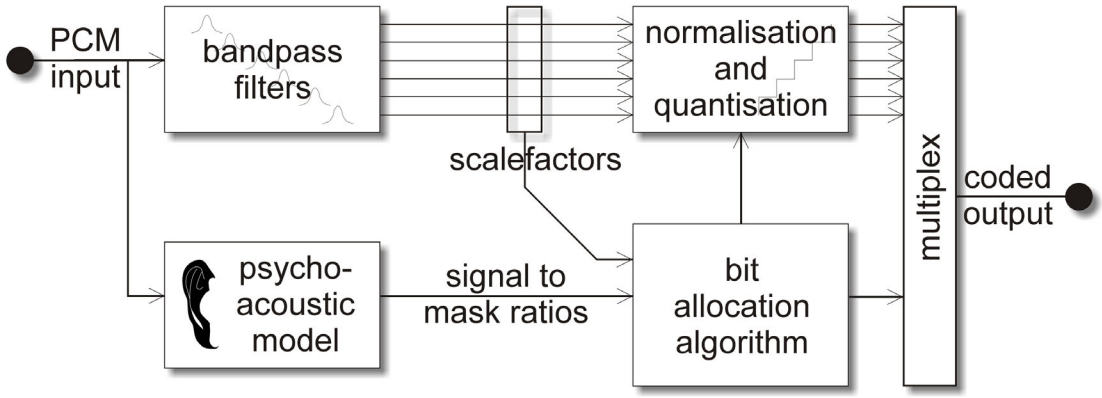

Figure 2.2: General structure of a psychoacoustic codec

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

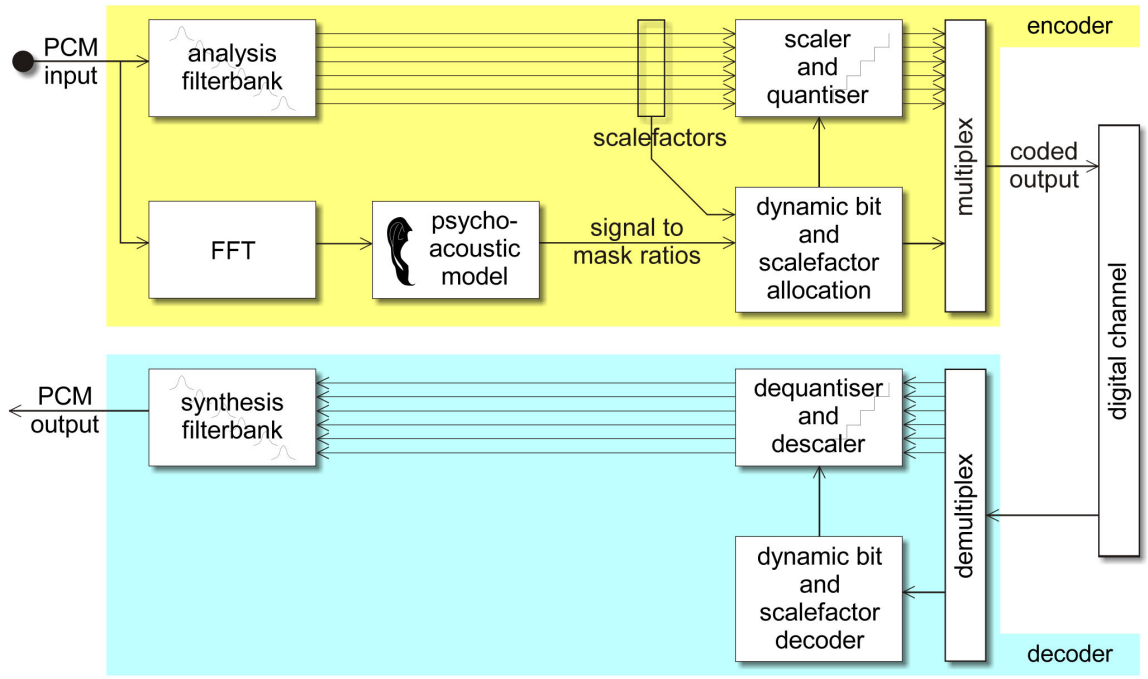

Figure 2.3: Structure of MPEG-1 audio encoder and decoder, Layers I and II

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

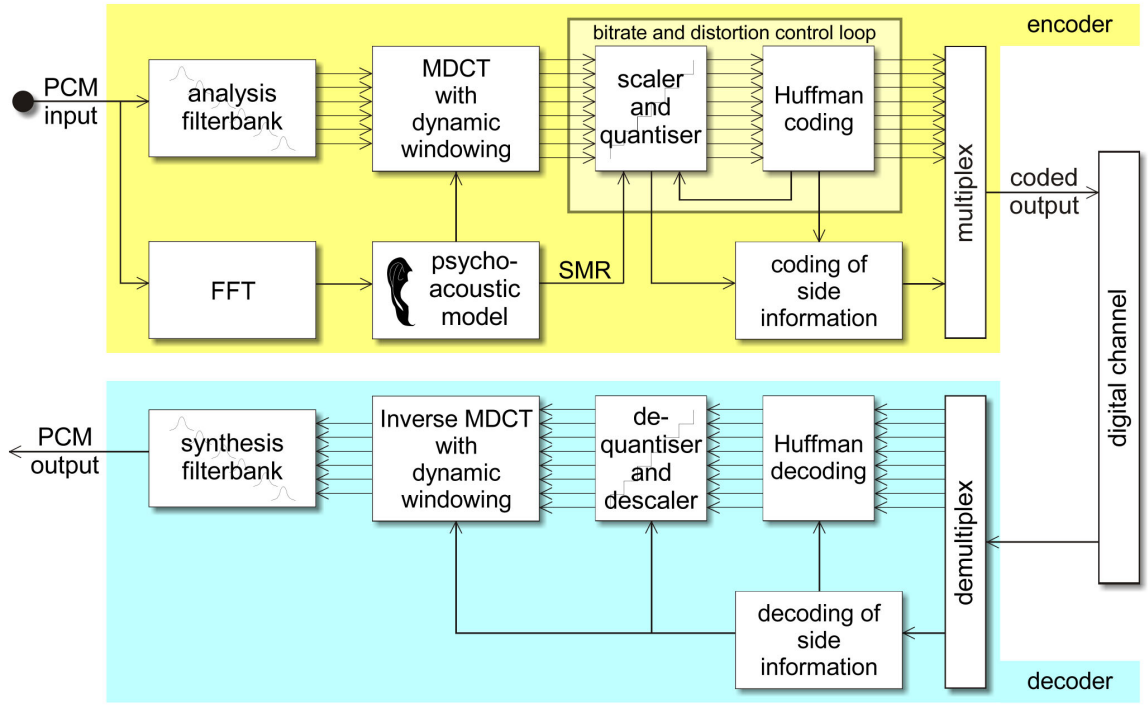

Figure 2.4: Structure of MPEG-1 layer III audio encoder and decoder

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.

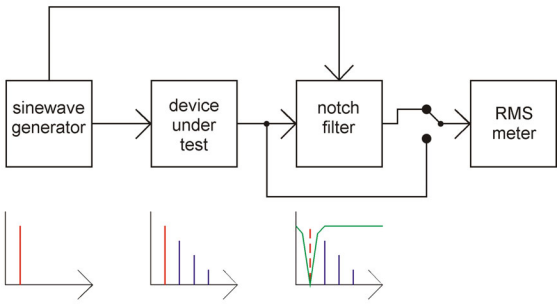

Figure 2.5: THD measurement technique

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

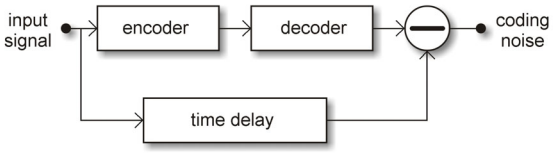

Figure 2.6: Input Output difference analysis

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.

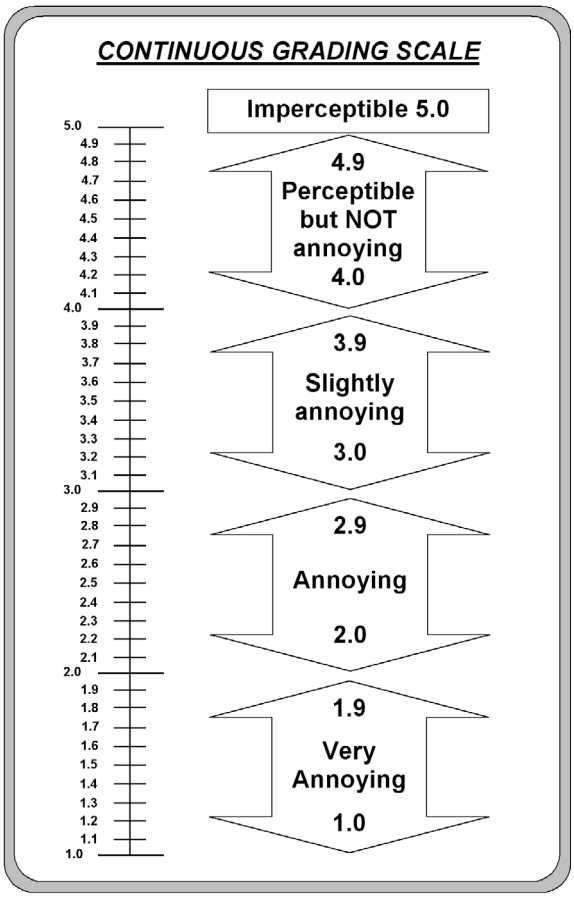

Figure 2.7: Scale used within BS.1116 (from [ITU-R BS.1284, 1997])

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.