AUDIO CODING: BASIC PRINCIPLES AND RECENT DEVELOPMENTS

MARINA BOSI

MPEG LA, LLC, Colorado, USA
mab@mpegla.com

This paper presents an overview of the technologies adopted in perceptual audio coding. Different multichannel configurations, their advantages and disadvantages, and their impact on the system data rates and capacity are examined. The basic principles behind the quantization, time to frequency mapping and psychoacoustic models stages of an audio coding system are analyzed. Emerging audio formats and new trends, such as parametric audio coding, perceptual noise substitution, PNS, spectral band replication, SBR, and sinusoidal modelling are examined.

INTRODUCTION

After the introduction of the CD format, consumer expectations have risen to demand audio quality that corresponds to the equivalent of signal to noise ratios and dynamic ranges above 80 dB and signal spectral content of above 15 kHz. While the CD format specifies only two channels, multichannel audio is providing end users with a more involving experience and is becoming more and more appealing to music producers.

Standardisation bodies including SMPTE, EBU, ITU-R, ISO/IEC MPEG have converged on the so-called 5.1 multichannel format. For example, the DVD-Audio format includes multichannel audio along with increased audio sample resolution (24-bit) and sampling rates (88.2, 96, and 192 kHz). In spite of a steady increase in storage media capacity and transmission bandwidth, high quality multichannel audio creates a challenge for traditional and new delivery media. In this paper, an overview of multichannel audio and the basic principles behind perceptual audio coding are described. Emerging MPEG-4 audio technologies will be discussed. Applications and new directions will be examined.

1. MULTICHANNEL AUDIO: FROM STEREO TO 5.1 AND ABOVE

Since the second half of the 19th century, when Thomas Edison was first able to mechanically record and reproduce the sound of his voice and Alexander Graham Bell succeeded in electrically transmitting the voice over a distance, the art of sound coding, transmission, recording, mixing, and reproduction has been constantly evolving. Starting with monophonic technology, and partially pushed by the progress in the film industry, the art of multichannel sound developed towards stereophonic, quadraphonic, 5.1 channels and more.

If we look back in music history, composers took advantage of the spatial attributes of music since a very early stage. Think, for example, to Vivaldi's concerto "per eco lontano" for violin and orchestra, in which a second violin is placed behind the scenes to simulate an echo effect.

In the eighties, with the introduction of the CD format, stereophonic sound became well established, while few artists were mastering in quadraphonic and a very small number of audiences had access to reproduction systems with more than two channels. Talking about more than two-channel technology was considered an extravaganza fitting only some sort of "modern music".

Today, the general evolution of digital technology and a steady growth in transmission bandwidth and storage capacity have made multichannel audio a more realistic option for artists that want to reach a broad audience.

1.1 The ITU-R 5.1 configuration

The most widely adopted multichannel configuration is the 5.1-channel configuration, often referred to as the 3/2/.1 configuration, with three loudspeakers placed in front of the listener, and two in the side/rear. This arrangement is described in detail in the ITU-R recommendation BS.775 [1]. According to the ITU-R specifications, the five full-bandwidth loudspeakers are placed on a circumference (see Fig 1) centered on the reference listening position. The 5.1 reference channel layout as adopted by recommendation ITU-R BS.1116 [2] is shown in Fig 1. Three front loudspeakers are placed at angles from the listener axis of -30^0 (left channel), $+30^0$ (right channel), and 0^0 (centre channel); the two surround loudspeakers are placed at -110^0 (left surround channel) and $+110^0$ (right surround channel). In addition to the five full-bandwidth channels, a low frequency enhancement channel, covering frequencies below 200 Hz, hence less than 10% of a 20-kHz bandwidth signal, is typically placed in the front,

although its exact location is irrelevant. More recent studies, see for example [3], showed the importance of adopting two LFE channels. In [3] it is shown how the use of two LFE channels enables the presentation of decorrelated low-frequency signals that are particularly effective in producing variation in auditory spatial imagery through the results of three controlled listening experiments.

1.2 A slightly different approach in loudspeakers layout

Independent work done in the seventies studied the minimum number of channels required to reproduce a subjectively diffused sound field [4]. A test was carried out with twenty loudspeakers in a circle separated by 18^0 in the horizontal plane energising 1, 2, 3, etc. loudspeakers with uncorrelated noise in an anechoic space [5]. The minimum number of channels required to reproduce a subjectively diffused sound field was found to be five in agreement with BS.775. The loudspeakers placement, however, differs from the ITU-R BS.775 or BS.1116 specifications. In [5] the loudspeakers were placed at $+-36^0$, $+-108^0$, and 180^0 from the listener axis.

The 5.1 configuration was first introduced by SMPTE in 1987 [4] and then adopted by a number of standardization bodies. It is worth noticing that this configuration was highly influenced by the film industry practice of sound accompanying pictures. Practical implementation reasons, as well as typical sound film material with the dialog in the centre channel and special low frequency effects, had a lot of weight in the choice of this particular multichannel configuration [6]. With a relatively small number of loudspeakers for spatial sound reproduction, the loudspeakers location, as well as their characteristics, levels, etc., play an extremely important role. It would be nice to compare relative results in these two different layouts.

1.3 How many channels is enough?

The question that naturally arises is how well the ITU-R configuration is able to reproduce all-around spatial imaging. The answer to this question depends obviously to some degree on the multichannel program material and the type of the loudspeakers employed. There is an increasing awareness, however, that the choice of five channels for spatial sound reproduction is a coarse way of representing sound fields. Again, we need to keep in mind that this choice was, to some extent, inherited from cinema practises, where the need to store sound on film dictated some heavy restrictions, and faithful reproduction of real-life spatial sound was not the primary goal. One would expect that the adoption of more than five channels would lead to a stronger sense of envelopment and more accurate spatial images. Applications in the film industry and demonstrations at AES Conventions in recent years have shown good performance with eight-channel systems. However no standard with more channels than the ITU-R 5.1 configuration has merged. In the following sections of this paper we will refer generally to the 5.1-channel configuration as the multichannel configuration unless otherwise specified.

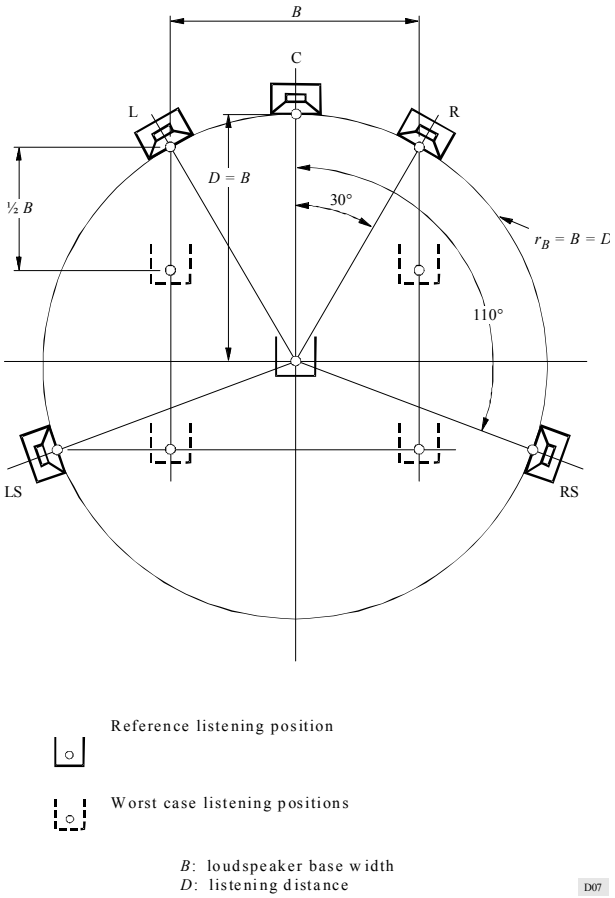

Figure 1: ITU-R 5.1-Channel Reference Configuration from [2].

If more than 5.1 channels are employed, the basic ideas and main principles still hold true, but with even greater compression requirements for the coding technology adopted.

2. MEDIA CAPACITY, MULTICHANNEL AUDIO, AND LINEAR PCM

Going from stereophonic to multichannel sound reproduction adds to the demands on storage and delivery media. If we consider the CD format (linear PCM), the stereophonic audio signals are sampled at a frequency, F_s, of 44.1 kHz and quantized using uniform

quantization with a precision, R, of 16 bits per sample leading to a total data rate, B of:

$$B_{CD} = 2 \cdot R \cdot F_s = 1.411\ Mb/s$$

An hour of music in the CD format needs 635 MB of storage. The total audio capacity of a CD is less than 800MB, allowing for a maximum playback time of about 74 minutes. If we consider a multichannel signal, we have:

$$B_{CD multichannel} = 5.1 \cdot R \cdot F_s = 3.598\ Mb/s$$

An hour of multichannel music in the CD format requires 1.62 GB, way above the CD capacity. In addition to the CD, other current multichannel applications include digital television in the US, which allows a bandwidth of 384 kb/s to multichannel audio, and internet audio, in which the typical user data transfer rate much lower than $B_{CDmultichannel}$. In each case bandwidth/capacity are serious challenges.

The challenge of multichannel audio coding is to minimise the data rate without sacrificing audio quality. While linear PCM is a well-understood and well-established coding method that offers very low-complexity implementations, it requires very high capacity/bandwidth to provide CD-quality audio signals. One should also notice that CD audio signals suffer from some degradation. It was shown in [7] by comparing the hearing threshold with the CD signal resolution levels, that audible quantization noise can be introduced in the mid-range frequencies.

The implication is that, expensive as it may be, one may need to increase the PCM sample resolution, going for example from R=16 to R=24. If, in addition to the augmented sample resolution, we also consider the new trend to adopt higher sampling rates, from F_s = 44.1 kHz or F_s = 48 kHz to F_s = 96 kHz, and F_s = 192 kHz, media restrictions become even more binding, prohibiting multichannel audio even for emerging new high capacity technologies like, for example, DVD-Audio applications. While there is no scientific evidence or published experimental results to the author knowledge, that unequivocally prove the advantages of adopting this increased resolution, many recording engineers and golden ears feel that high resolution, i.e. 96/24, and multichannel audio are essential in providing the end user with high quality audio and a truly enveloping experience

It should be noted that a PCM coder does not take into consideration the characteristics of its input signals, leaving redundancy in the signal representation. Moreover, it is important to emphasise that the last stage in the audio coding chain is the human ear. Some spectral components are more audible than others, therefore a bit distribution modelled upon human hearing characteristics can improve significantly the perceived quality of the signal at a given data rate with respect to a uniform bit distribution as adopted in the PCM technology.

An alternative to linear PCM coding is therefore to eliminate redundancies in the signal and to redistribute the bit pool appropriately in the frequency domain. Additional bits are used in part of the signal spectrum where they are needed to the expense of the number of bits in other parts of the spectrum where a higher number of bits is irrelevant. The added complexity of such a system allows for a more efficient use of the overall number of bits available for the signal representation.

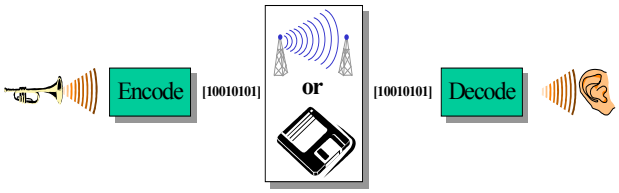

Figure 2: Basic Audio Coding Chain.

3. BUILDING BLOCKS

A variety of approaches for coding audio data are currently available in the marketplace. In order to give the reader an overview of the basic principles, this section describes the most relevant stages of an audio coding system. The basic block diagram of an audio coding scheme is shown in Fig 3. For each audio channel, the PCM audio signal is mapped onto the frequency domain. The time to frequency mapping is implemented via a sub-band filter (MPEG Layers I and II) or via a transform filter bank (MPEG Layer III, AC-3, MPEG AAC). Although mathematically equivalent, these two approaches historically led to different architectures for audio coding.

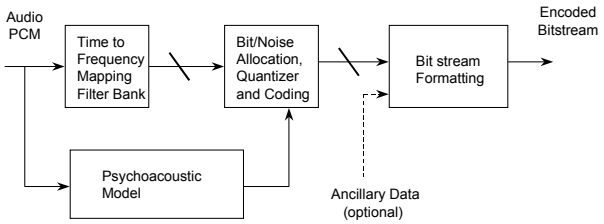

Figure 3: Basic Audio Coding Building Blocks.

Masking thresholds in the bark scale are computed based on a psychoacoustic model for blocks of samples for each channel. The masking thresholds values determine how many bits (MPEG Layers I and II, AC-3) from the common bit pool are allocated to different frequency regions or equivalently how much quantization noise can be injected in a frequency region

without being perceived (MPEG Layer III, MPEG AAC). The quantization stage accordingly quantizes the time-frequency representation of the audio signal and is sometimes followed by a noiseless coding stage (MPEG Layer III, MPEG AAC). Finally, the audio data are interleaved with control parameters and auxiliary data to provide the encoded bistream. The decoding process first de-multiplexes the encoded bitstream, inverse quantizes the audio data based on control parameters transmitted in the bitstream and/or based on the re-computation of some of the psychoacoustic parameters (AC-3) and maps back the signal from the time-frequency to the time domain.

In addition to intra-channel redundancies and irrelevancies removal, multichannel audio coding takes advantages of inter-channel correlations and binaural auditory mechanisms. One of the technology most commonly used to remove inter-channel correlations is the so-called mid/sum, M/S, stereo coding generalised to multichannel [8, 9]. In the M/S approach, the content of two separate channels spectra is summed and subtracted to each other and either the sum/difference signal or the original signal is transmitted depending on the degree of correlation between the two channels.

Based upon the human ear ability to localise sound above 2-3 kHz mostly via interaural intensity differences (IID), in order to reduce the data rate, instead of separate spectral channel information, just a combined spectral signal is transmitted above a certain cut-off frequency [10, 11]. This technique is sometime referred to as intensity stereo coding (MPEG Layers I, II, and III) if applied to two channels or channel coupling (AC-3) if applied to multichannel signals.

3.1 Time to Frequency Mapping

The first stage in perceptual audio coding schemes is usually represented by the time to frequency mapping of audio signals. The basic idea is to filter the signal into its components in various frequency bands. By subdividing the signal into its frequency components and representing the signal by its frequency component parameters, a great reduction in the amount of data needed to reproduce the audio signal can be achieved.

Consider for example a sine wave and its representation in the frequency domain (see Figure 4). While only three parameters, namely frequency, phase, and amplitude, for each block of data fully describe the sine wave in the frequency domain, a large number of PCM data is needed to describe this simple signal in the time domain. This example clearly shows how, by filtering the signal, redundancies can be easily extracted from the audio signals.

In general, although audio signals will not exhibit strict periodicity as in the simple sine wave example, it can be shown that audio signals are quasi-stationary and that they can be modelled by using short-term spectrum analysis.

Once the signal is represented in the time-frequency domain, the number of bits used to encode each frequency component can be adjusted so that greater

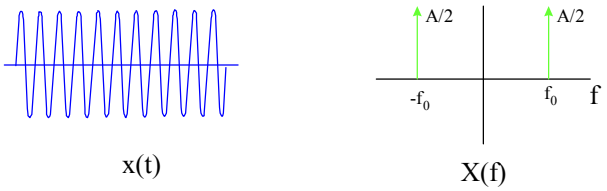

Figure 4: Time vs. frequency representation of a cosine.

encoding accuracy can be placed in frequencies where it is most needed. For example, if we can break the signal into its energy per each critical band, we can apply masking models to separate irrelevant elements of the signal from relevant ones.

A variety of time to frequency mapping algorithms, which differ by the degree to which they allow for source component separation and source redundancy extraction, are available. Just to mentioned only a few, discrete Fourier transform (DFT), discrete cosine transform (DCT), quadrature mirror filters (QMF), pseudo QMF (PQMF), modified DCT (MDCT), hybrid filter banks, wavelet, etc. are time to frequency mapping techniques found in literature for perceptual audio coding [12-20].

Different factors come into play in the design of the filter bank stage in perceptual audio coding. Firstly, we would like to optimally separate the different spectral components so that the perceptual coding gain can be maximised. Since we will be performing short-time analysis/synthesis of the signals, we would like to minimise the audibility of blocking artefacts both in terms of boundary discontinuities and pre-echo effects. The window shape plays an important role in the spectral separation of the signal and blocking artefacts. While no single window provides optimal resolution for all signals, Kaiser Bessel derived [21], KBD, and sine windows are mostly utilized in time to frequency mapping stage of audio coding systems.

Secondly, given that the ultimate goal is to decrease the data rate while maintaining the quality of the audio signal, critically sampled systems are desirable. In these systems, the overall rate at the output of the analysis stage equals the overall rate at the input of the analysis stage.

Thirdly, while this is not a strict requirement, most of the perceptual audio coders currently in use employ perfect reconstruction, PR, or “nearly” PR filter banks, where the output signal to the synthesis filter bank is an identical, delayed replica of the original signal.

A simple example of a critically sampled N-channel filter bank is shown in Figure 5. The input signal is filtered by N band-pass filters, H_k[1]. Each band-pass filter output is then sub-sampled by a factor of N, i.e. it is critically sampled at a rate that is twice the nominal bandwidth of each band-pass filter. In the synthesis filter, the signal is up-sampled and filtered by the set of the N G_k filters. If perfect reconstruction filters can be applied, in absence of quantization, the sum of the output of the G_k filters equals the delayed original signal.

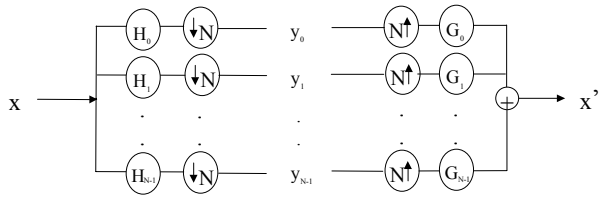

Figure 5: Critically sampled N-channel filter bank.

The down-sampling operation can introduce aliasing in the signal spectrum if there is overlap between adjacent band-pass filters, while the up-sampling operation can introduce imaging. With an appropriate choice of analysis/synthesis filters these distortions in the spectrum cancel each other in the synthesis stage after all components are added together.

Critically sampled filter banks commonly found in audio coding are the pseudo quadrature mirror filters, PQMF [22]. The basic idea is that the filters are designed so that aliasing from adjacent bands is exactly cancelled but aliasing from next-neighbor bands is ignored. Nussbaumer [22] suggested that the band-pass filters, $h_k(n)$, where n is the time index and k is the frequency index, be a modulated version of a single low pass filter with bandwidth fs/N.

$$h_k(n) = h(n)\cos[\pi/N(k+1/2)(n-(L-1)/2)+\varphi_k)]$$

$$k= 0, 1, \ldots N-1 \qquad n= 0, 1, \ldots L-1$$

where N is the number of frequency channels and L is the length of the filters h_k. The reconstruction filters can be derived from the analysis filter as follows:

$$h_k(n) = g_k(L-1-n).$$

The low pass filter prototype should also satisfy (as much as possible) the PR conditions:

$$|H(e^{j\omega})|^2 + |H(e^{j(\pi/K-\omega)})|^2 = 2 \qquad 0<|\omega|<\pi/2N$$

$$|H(e^{j\omega})|^2 = 0 \qquad |\omega|>\pi/N$$

These filters are computationally very efficient since they can be realized via an FFT and are of moderate complexity and low delay. A polyphase QMF with length L = 512, number of channels N =32, and φ_k= -N/2, is used in the MPEG Audio coding schemes [15].

Another example of critically sampled filter bank is the MDCT [14]. This transform is based on time domain aliasing cancellation (TDAC) and was first introduced by Princen and Bradley [23]. The time-invariant TDAC transform provides a critically sampled system with 50% overlap between adjacent windows.

In the analysis stage, N new input time samples are buffered and windowed with a window of length 2N (see Figure 6). The signal is then mapped from time to frequency domain by using the MDCT (oddly stacked TDAC) or alternating an MDCT with a modified discrete sine transform, MDST (evenly stacked TDAC). The inverse-transformed signal contains time aliasing distortion, which, in absence of quantization, is cancelled during the window and overlap-add stage.

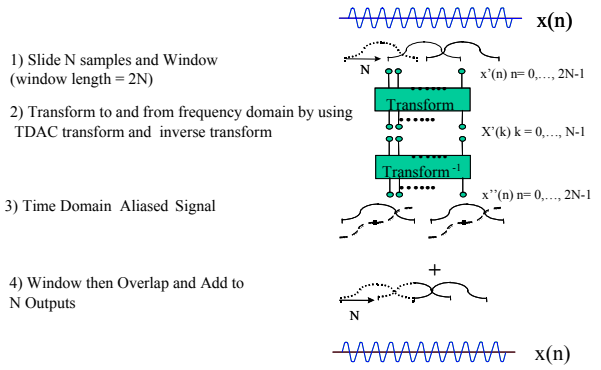

Figure 6: TDAC Transform [14, 23].

The forward TDAC transform can be generalised as follows [21]:

$$X_i(k) = \sum_{n=0}^{2N-1} x_i(n) e^{-j(2\pi/2N)(k+k_0)(n+n_0)} \quad k = 0,1,\ldots,2N-1$$

where

$x_i(n)$ is the windowed input sequence of 2N samples coefficients for the i-th block;

$X_i(k)$ is the sequence of 2N frequency coefficients for the i-th block;

k_0 is a frequency offset in the transform kernel;

$$k_0 = \begin{cases} 1/2 & \text{for the OTDAC} \\ 0 & \text{for the ETDAC} \end{cases}$$

n_0 is a time offset that allows for the cancellation of the time aliasing introduced in the signal; in general n_0 depends on the length of the overlapping region with the

[1] One could apply different structures, e.g. tree-structures cascades of two-band filters, etc.

next block of samples; in the case of the time-invariant TDAC transforms we have:

$$n_0 = \frac{(N+1)}{2}$$

Accordingly, the inverse TDAC transform can be generalised as follows:

$$x_i'(n) = \frac{1}{N}\sum_{n=0}^{2N-1} X_i(k) e^{-j(2\pi/2N)(k+k_0)(n+n_0)} \quad k = 0,1,...2N-1$$

where x'_i equals the delayed, time-aliased input sequence.

The ETDAC and OTDAC MDCT kernel can be obtained from the above equations by taking the real part; the ETDAC MDST kernel can be obtained by taking the imaginary part. If x(n) is real, then the MDCT is odd-symmetric and the MDST is even-symmetric, therefore only N independent frequency coefficients are generated for each transform block. In absence of quantization, after the window and overlap-add stage of the time-invariant TDAC, the output signal becomes an exact delayed replica of the input signal provided that analysis and synthesis windows satisfy the following requirement:

$$W^a(n)W^s(n)+W^a(N+n)W^s(N+n)=1 \quad n=0, 1, \ldots, N-1$$ [2]

The TDAC transforms can be efficiently implemented via an FFT kernel; fast implementations of the MDCT exist in literature see for example [31]. For power of two block lengths[3], the number of complex multiplies /additions is $N/2 + N/2 \log_2(N/2)$, where N is the number of frequency channels. The ETDAC is used in coding schemes like AC-2 [21]; the OTDAC is used in MPEG Audio [15, 24-28] , AC-3 [21], PAC [29], Twin VQ [30], etc..

While historically the PQMF and the MDCT were developed independently, Malvar [31] showed how these approaches can be unified in the frame of the lapped orthogonal transforms, LOT. Given a number of frequency channels N, by appropriately selecting the length, L, and phase, φ_k, of the filters $h_k(n)$ and imposing the following conditions on the prototype filter:

$$h(2N-1-n) = h(n) \quad \text{and}$$

$$h(n)^2 + h(n+N)^2 = 2$$

not only perfect reconstruction is achieved, but the PQMF expression becomes equivalent to the OTDAC MDCT expression where the analysis window is identical to the synthesis window. By setting the filter length L = 2N, and $\varphi_k=(k+1/2)(2+1)\pi/2$, we obtain

$$h_k(n) = h(n)\cos[\pi/N(k+1/2)(n+(N+1)/2)]$$

which is equivalent to real part of the generalized TDAC transform when $k_0 = ½$, i.e. equivalent to the OTDAC MDCT expression.

In general, it is both practical and efficient to represent the signal in the frequency domain. From the human-perception point of view it is particularly meaningful to be able to separately manipulate spectral components of the signal. In summary, filter bank framework provides the best medium for the removal of redundancy, i.e. information that is not necessary to uniquely identify the signal, and irrelevancies, i.e. information that is perceptually not important.

3.2 Psychoacoustic Models

In the psychoacoustic model stage of the coder we dynamically derive the values for the masking thresholds based on the level of the different signal components (maskers) and the hearing threshold. Typically, perceptual audio coders perform a high-resolution DFT (using the FFT algorithm) with blocks of input data solely for use in the psychoacoustic model. The results of this high frequency resolution DFT are then employed to determine the masking curve for each block of coded data. The general idea, as we saw in the previous section, is that the quantization noise can be localized in areas of the signal spectrum where it does not affect (or it least affects) the quality of the audio signal.

3.2.1 The hearing Threshold

The hearing threshold, or threshold in quiet, represents the lowest sound level that can be heard at a given frequency. Even in extremely quiet conditions, the human ear cannot detect sounds at sound pressure levels, SPL, below the threshold in quiet. This curve is extremely important for audio coding since frequency components in a signal that fall below this level are irrelevant to our perception. As long as the quantization noise in frequency components that are transmitted falls below this level, it will not be detectable by the human hearing process. In Figure 8 the threshold in quiet for is shown.

[2] Typically the same window is employed for both the analysis and synthesis stage. Notice that both the sine window and the KDB window satisfy condition (12).

[3] In the case of non-power of two block lengths, usually the FFT kernel is factorised into smaller, power of two length FFT, see for example the MPEG Layer III implementation [4], thus requiring a slightly higher number of multiplies/additions.

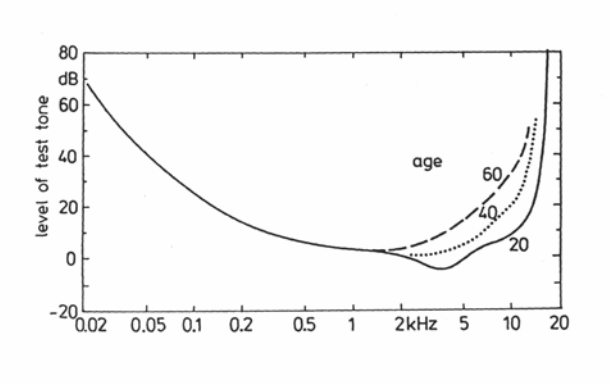

Figure 8: Threshold in quiet from [32].

As shown in [33], one can obtain a good approximation of the threshold in quiet by utilizing the following frequency dependent function:

$$A(f)/dB = 3.64(f/kHz)^{-0.8} - 6.5e^{-0.6(f/kHz-3.3)^2} + 10^{-3}(f/kHz)^4$$

where the threshold in quiet is modelled by taking into consideration the transfer function of the outer and middle ear and the effect of the neural suppression of internal noise in the inner.

3.2.2 Masking

Masking of soft sounds by louder ones is part of our everyday experience. For example, if we are engaged in a conversation while walking on the street, we typically cease conversation while a loud truck passes since we are not be able to hear speech over the truck noise. This can be seen as an example of masking: when the louder masking sound (the truck) occurs at the same time as the maskee sound (the conversation), it is no longer possible to hear the normally audible maskee. This phenomenon is called simultaneous or frequency masking. Another example of frequency masking occurs when in a performance one loud instrument (masker) masks a softer one (maskee) that is producing sounds close in frequency. In general simultaneous masking phenomena can be explained by the fact that a masker creates an excitation in the cochlea's basilar membrane (see also next sections) that prevents the detection of a weaker sound exciting the basilar membrane in the same area.

Masking can also take place when the masker and the maskee sounds are not presented simultaneously. In this case we refer to this phenomenon as temporal masking. For example, in speech a loud vowel preceding a plosive consonant tends to mask the consonant. Temporal masking is the dominant effect for sounds that present transients, while frequency masking is dominant in steady state conditions. For example, in coding sharp instrument attacks like those of castanets, glockenspiel, temporal masking plays a more important role than frequency masking.

3.2.3 Frequency Masking

Figure illustrates frequency masking. In this figure, we see a loud signal masking two other signals at nearby frequencies. In addition to the curve showing the threshold in quiet, the figure shows a curve marked "masking threshold"[4] that represents the audibility threshold for signals in the presence of the masking signal. Other signals or frequency components that are below this curve will not be heard when the masker is present. In the example shown in Figure 9, the two other signals fall below the masking threshold, so they are not heard even though they are both well above the threshold in quiet. Just like with the threshold in quiet, we can exploit the masking thresholds in coding to identify signal components that do not need to be transmitted and to determine how much inaudible quantization noise is allowed for signal components that are transmitted.

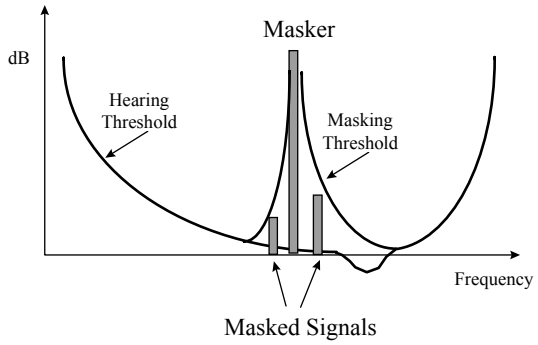

Figure 9: Example of frequency masking.

3.2.4 Temporal Masking

In addition to simultaneous masking, masking phenomena can extend in time outside the period when the masker is present. Masking can occur prior to and after the presence of the masker. Accordingly, two types of temporal masking are generally encountered: pre-masking and post-masking. Pre-masking takes

[4] We shall refer to "masking thresholds" or "masking curves" to indicate the elevation of the hearing threshold due to the presence of one or more masker sounds. We define the "masked threshold" or "masked curve" as the combination of the hearing threshold and the masking threshold.

place before the onset of the masker; post-masking takes place after the masker is removed. Pre-masking is somewhat an unexpected phenomenon since it takes place before the masker is switched on. In general, temporal masking can be explained if we consider the fact that the auditory system requires a certain integration time to build the perception of sound and by the fact that louder sounds require longer integration intervals than softer ones.

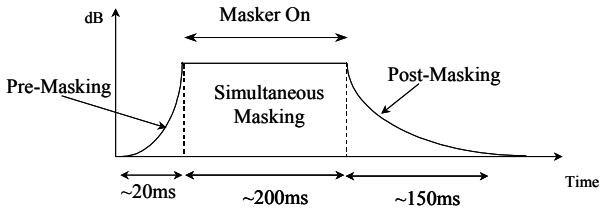

Figure 10: Example of temporal masking.

In Figure 10, an example of temporal masking is shown [32]. A 200 ms masker masks a short tone burst with very small duration relative to the masker. In the figure pre-masking lasts about 20 ms, but it is most effective only in the few milliseconds preceding the onset of the masker. There is no conclusive experimental data that link the duration of pre-masking effects with the duration of the masker. Although pre-masking is a less dramatic effect than post or simultaneous masking, it is nevertheless an important issue in the design of perceptual audio codecs since it is related to the audibility of "pre-noise" or "pre-echo" effects caused by encoding blocks of input samples. Pre-noise or pre-echo distortion occurs when the energy of the coded signal is spread in time prior to the onset of the attack. This effect is taken into consideration in the design of several perceptual audio coding systems both in terms of psychoacoustics models and analysis/synthesis signal adaptive filter design.

3.2.5 Critical Bandwidths

In measuring frequency masking curves, it was discovered that there is a narrow frequency range around the masker frequency where the masking threshold is flat rather than dropping off. A "critical bandwidth" exists around the centre frequency of a masker that exhibits a constant level of masking regardless of the type of masker. The concept of critical bandwidth was first introduced by Harvey Fletcher in 1940 [34]. Fletcher's measurements and assumption led him to model the auditory system as an array of band-pass filters with continuously overlapping pass-bands of bandwidths equal to critical bandwidths. Experiments have shown that the critical bandwidth depends on the frequency of the masker. However, the exact form of the relationship between critical bandwidth and masker frequency is somewhat subject to controversy since differing results have been obtained using different types of measurements. In the pioneering work of Fletcher and later work by Zwicker [35], the critical bandwidth was estimated to be constant at about 100 Hz up to masker frequencies of 500 Hz, and to be roughly equal to 1/5 of the frequency of the masker for higher frequencies. An analytical expression that smoothly describes the variation of critical bandwidth Δf as a function of the masker center frequency f_c is given by [32]:

$$\Delta f / \mathrm{Hz} = 25 + 75\left[1 + 1.4\left(f_c / \mathrm{kHz}\right)^2\right]^{0.69}$$

A number of articles including Greenwood [36], Scharf [37], Patterson [38], Moore and Glasberg [39] disagree in their estimation of the critical bandwidths with that of the standard formula, especially below 500 Hz. In particular, Moore and Glasberg measure a quantity they define called the "equivalent rectangular bandwidth", ERB, which should be equivalent to the critical bandwidth. Their experiments were designed to provide an estimate of the auditory filter shapes by detecting the threshold of a sinusoidal signal masked by notched noise as a function of the width of the notch. The ERB as defined by Moore and Glasberg is about 11% greater than the -3 dB bandwidth of the auditory filter under consideration. The ERB, as a function of the center frequency f_c of the noise masker, is well fit by the function [40]:

$$\mathrm{ERB/Hz} = 24.7\ (4.37\ f_c/\mathrm{kHz} + 1)$$

The ERB function seems to provide values closer to the critical bandwidth measurements of Greenwood [36] than of Flecther or Zwicker at low frequencies. Figure 11 compares the standard critical bandwidth formula with Moore's ERB formula and with other experimental measurements of critical bandwidth. Notice that the critical bandwidths predicted by the ERB formula are much narrower at frequencies below 500 Hz than implied by the standard critical bandwidth formula. Since the critical bandwidth represents the width of high-level masking from a signal, narrower critical bandwidth estimates put stronger requirements on a coder's frequency resolution.

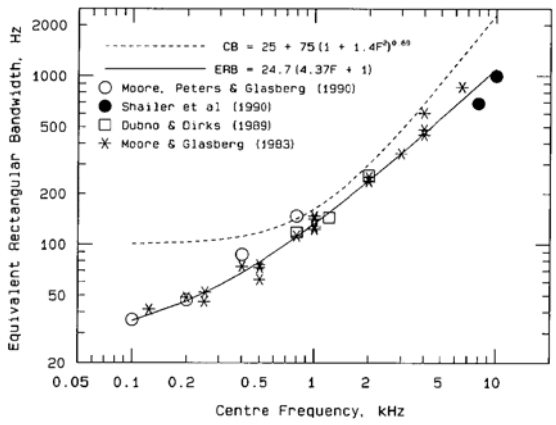

Figure 11: Critical bandwidth function and the ERB function plotted versus different experimental data for critical bandwidth from [40].

We can measure frequency masking curves for various masking and test signals. In all cases, we find that the masking curve levels are highest at frequencies near the masker frequency and drop off rapidly as the test signal frequency moves more than a critical bandwidth away from the masker frequency. The shape of the masking curves depend on the frequency of the masker and its level. The masking curves depend strongly on whether or not the masker is tonal or noise-like, where much greater masking is created by noise-like maskers.

In general, the masking curve relative to the masker M can be dynamically derived from the level of the masker L_M (typically estimated in the psychoacoustic model stage by computing a short-time FFT of the masker M) by:

a) Down-shifting it by a constant that depends on the masker M and
b) Adding a frequency dependent function that describes the spreading of the masker's excitation energy along the basilar membrane.

The down-shift depends both on the characteristics of the masker, namely whether it is noise-like or tone-like, and its frequency. The masker "spreading function" is based on empirical data that describe the contour of masking curves. The masking curve shape is greatly simplified when shown in terms of critical bandwidths rather than on a frequency scale. The critical bandwidth formula introduced above gives us a method for mapping frequency onto a critical bandwidth rate, z(f), or Bark scale.

As a first approximation, a representation of the spreading function is given by a triangular function. We can write this spreading function in terms of the Bark scale difference between the maskee and masker frequency $dz = z(f_{maskee}) - z(f_{masker})$ as follows [41]:

$$10\log_{10}\left(F(dz, L_M)\right) = \left(-27 + 0.37\,\mathrm{MAX}\{L_M - 40, 0\}\,\theta(dz)\right)|dz|$$

where L_M is the masker's level and $\theta(dz)$ is the step function equal to zero for negative values of dz and equal to one for positive values of dz. Notice that dz assumes positive values when the masker is located at a lower frequency than the maskee and negative values when the masker is located at a higher frequency than the maskee. In Figure 12, this spreading function is shown for different levels of the masker L_M.

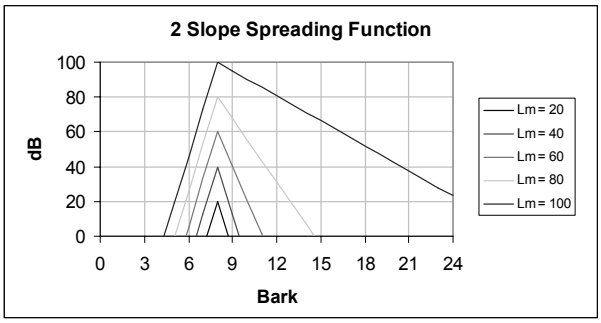

Figure 12: Spreading function described by the two slopes derived from narrow-band noise masking data for different levels of the masker.

Typically, audio sounds may contain several tone-like and noise like components. Once the thresholds are combined to create a global masked threshold, the threshold of hearing is also taken in consideration to derive the masked threshold for the signal during that time interval. Often in perceptual audio coding, the maximum value between the global masked threshold and the threshold of hearing is retained (like, for example, in MPEG Psychoacoustic Model 2 and AC-3) as the masked threshold for the signal at that time interval. Portions of the signal below the masked threshold are considered irrelevant to the signal representation.

In Figure 13, a simple example for different quantization SNR for a signal partially masked by a stronger one to its right is shown. The goal of bit allocation is to make sure that bits are allocated so that the SNR is greater than the SMR across the spectrum. The difference between the SMR and the SNR is referred to as the noise to mask ratio, NMR, and gives an indication of the rate of distortion with respect to the computed masked

threshold (see Figure 13). Given a certain bit budget, when there are extra bits, they are allocated across the spectrum to create a coding margin. When there are not enough bits, bits are allocated to minimize the overall (positive) deviation between SMR and SNR or NMR.

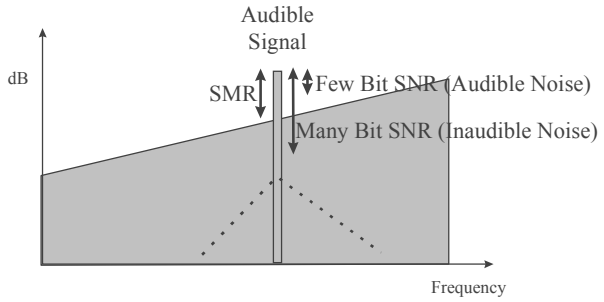

Figure 13: Example of different SNR values allocated to a signal component versus masked threshold.

3.3 Quantization and Bit Allocation

The basic idea in a data reduction scheme is to filter the signal into its components in various frequency bands (see Figure 7), the signal is then quantised in the frequency domain and the total bit pool is allocated dynamically depending on the energy of each spectrum component and its relevancy.

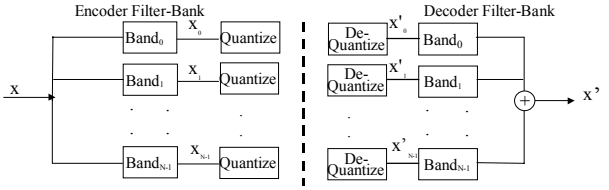

Figure 7: Basic idea in data rate reduction schemes.

Let us assume for a moment that the signal frequency components have equal energy and they populate the full spectrum. Let us also assume that we are not exploiting psychoacoustics models but we are concentrating on redundancies removal only. In this particular case there is really no gain in redistributing the bit pool throughout the spectrum because each component demands the same number of bits.

On the other hand, if we assume that the signal spectrum is coloured, e.g. the spectral components at low frequencies are stronger, then there is an increase in coding gain by redistributing the bit pool throughout the spectrum.

Luckily, the latter case is the most common. In this case the signal contains redundancies. These redundancies can be more or less efficiently removed. The efficiency of the removal depends upon the characteristics of the filter bank.

A measure of the redundancies present in the signal representation is given by the spectral flatness measure (sfm): the flatter is the spectrum, the less redundant is the signal. The sfm is given by the ratio of the geometric mean to the average of the power spectral density of the signal. Low sfm implies potential high coding gains [42].

By comparing various bit allocations at a given level of average block distortion $<q^2>$, where $q_k=x_k-Q^{-1}(Q(x_k))$ is the quantization noise for each spectral component and k is the spectral component index, we can find a method that optimally allocates bits through the spectrum, where our ultimate goal is to localize the quantization noise below the masking thresholds.

We can increase the coding gain with respect to the PCM coding gain if we can find a set of R_k, where R_k represents the number of bits used to code the spectral line of index k, that minimizes the error[5]:

$$< q^2 >= \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{x_k^2}{3\cdot 2^{2R_k}}\right) \qquad (1)$$

such that

$$\frac{1}{N}\sum_{k=0}^{N-1} R_k = R \qquad (2)$$

where N is the number of spectral lines, and R is the average number of bits per spectral line available.

This is a problem of constrained minimisation that can be solved by using a Lagrange multiplier λ, to enforce the average bit rate constraint as specified in (2)[6]. By taking the derivative with respect to each R_k and with respect to λ and by solving the resulting equations for R_k and then enforcing average bit rate constraint we obtain:

$$R_k = R + \tfrac{1}{2}\log_2(x_k^2) - \tfrac{1}{2}\log_2\left(\prod_{j=0}^{N-1} x_j^2\right)^{\frac{1}{N}} \qquad (3)$$

From (3) it is apparent that, for each block of samples, a bit allocation based upon the spectral energy distribution

[5] We are assuming error-free transmission, non-overlapping equal-width sub-bands, and the use of PCM coding of individual sub-bands with a midrise quantizer with maximum non-overload value equal to x_k.

[6] In our derivations we assume that we would always get $R_k>=0$; the above algorithm, however, will sometimes give us negative values of R_k when x_k is much below its geometric mean. (We really should have included Kuhn-Tucker multipliers to keep all of the R_k non-negative.) In practice, one usually rounds those R_ks to zero and takes bits away from other parts of the spectrum. In this case we use an approximate solution allocating bits one by one locally (e.g. water filling algorithms, etc.).

of the signal will introduce an improvement with the respect to uniform allocation, when the geometric mean of the signal power spectral density is much smaller than the average of the signal power spectral density.
The ratio of the geometric mean of the signal power spectral density to the average of the signal power spectral density is the sfm of the signal where:

$$sfm = \frac{\left(\prod_{k=0}^{N-1} x_k^2\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} x_k^2} \quad (4).$$

Notice that the sfm varies between 0 and 1; sfm = 1 implies a signal with a flat spectrum, and no coding gain with respect to uniform distribution of bits throughout the block can be achieved since, by substituting sfm = 1 in (3) we obtain

$$R_k = R .$$

Notice also that the sfm depends not only on the spectral energy distribution of the signal but also on the resolution of the filter bank, i.e. the total number of the frequency lines, N, or block length. If N is >>2, for a given signal, then the sfm decreases by increasing the block size N.

3.3.1 Irrelevancy extraction

In perceptual audio coding, the goal is not just to extract redundancy from the source, but also to isolate the irrelevant parts of the signal spectrum. This translates in not just trying to minimize the average error power $<q^2>$ per block, but trying to get the quantization noise below the masking curves generated by the signal under examination.

For components above the masking curve, i.e. relevant signals, this means that we want to maximize the difference between the signal to noise ratio (SNR), and the signal to mask ratio (SMR), or equivalently, minimize SMR-SNR, where

$$SNR = 10\log\frac{\langle x^2\rangle}{\langle q^2\rangle}$$

and

$$SMR = 10\log\frac{\langle x^2\rangle}{\langle M^2\rangle}$$

with M_k^2 corresponding to the masking threshold (see next section) energy value for the k component of the block spectrum. This differs from the minimization problem described in (1) in that we need to minimize the error weighted by the masking factor, i.e.:

$$<q^2/M^2> = \frac{1}{N}\sum_{k=0}^{N-1}\left(\frac{x_k^2/M_k^2}{3\cdot 2^{2R_k}}\right) \quad (5)$$

with the same constraint as described in (2). The resulting optimal bit allocation leads to:

$$R_k = R + \frac{1}{2}\log_2(x_k^2/M_k^2) - \frac{1}{2}\log_2\left(\prod_{j=0}^{N-1} x_j^2/M_j^2\right)^{\frac{1}{N}} \quad (6).$$

The "perceptual" sfm can then be described as:

$$psfm = \frac{\left(\prod_{k=0}^{N-1} x_k^2/M_k^2\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k=0}^{N-1} x_k^2/M_k^2} \quad (7)$$

Notice that the psfm depends on the spectral energy distribution of the signal weighted by the masking energy distribution. In this case, depending on the characteristics of the input signal, increasing the frequency resolution of the filter bank may or may not imply an increase in the coding gain. While for signals with steady state characteristics increasing the frequency resolution of the filter bank causes an increase in the coding gain, this is not true for transients. Work done by J. Johnston [43], showed that for tonal signals like the harpsichord, increasing the filter-bank block length is reflected in an increase in coding gain, while for transient-like signals, e.g. castanets, the coding gain tends to decrease by increasing the block size.

4. APPLICATIONS AND STANDARDS FORMATS

4.1 ISO/IEC MPEG Audio Layers I, II, and III

Responding to the industry need of interchangeable digital file formats, ISO/IEC MPEG provided the first digital audio compression standard, MPEG-1 Audio Layers I, II and III in 1992 [15]. Two-channel audio at sampling rates of fs_{MPEG-1} = 32, 44.1, 48 kHz, sample resolution equivalent to R = 16 bits per sample, and at data rates between 32 and 384 kb/s per channel was formally specified in this standard. Following the first phase of its standardisation efforts, in 1994 MPEG-2 audio [26] introduced the capability of multichannel (MC) sound and lower sampling frequencies (LSF) than MPEG-1 in order to achieve lower data rates.
In the MPEG-1 Audio coding scheme the time to frequency mapping is implemented via a 512-tap polyphase quadrature mirror filter (PQMF) [44] with 32

frequency channels[7] for Layers I and II. Layer III employs a hybrid filter bank composed by the 512-tap PQMF followed by an 18-point modified cosine transform (MDCT) [14] for a total of 576 frequency channels; during transients, in order to increase the time resolution of the filter bank, the PQMF is followed by 6-point MDCT, for a total of 192 frequency channels. The masking thresholds are computed by applying to the signal a 512-point FFT (Layers I and II) as described in [15] for Psychoacoutsic Model 1 or a 1024-point FFT (Layer III) as described in [45] for Psychoacoutsic Model 2. The output of the FFT analysis is used to detect noise versus tone like components of the signal. The evaluation of the masking thresholds is done via empirical data (Model 1) or by computing a spreading function [45] (Model 2) on a bark scale. The threshold in quiet on a bark scale is also taken into consideration. The quantization is performed by block companding groups of 12 samples (Layers I and II) or by employing non-uniform quantization followed by Huffman coding (Layer III).

The basic audio coding technology employed in MPEG-2 MC and LSF is the same as the technology employed in MPEG-1 (see [25] for more details).

Number of Audio Channels$_{MPEG\text{-}2MC}$ = 1-5.1
Fs$_{MPEG\text{-}2MC}$ = 32, 44.1, 48 kHz
R$_{MPEG\text{-}2MC}$ =16-24 (equivalent bit per sample)
B$_{MPEG\text{-}2MC}$ = 32 – 1,130 kb/s
Frame Size$_{MPEG\text{-}2MC}$ = 384-1152 samples

Number of Audio Channels$_{MPEG\text{-}2LSF}$ = 1-2
Fs$_{MPEG\text{-}2LSF}$ = 16, 22.05, 24 kHz
R$_{MPEG\text{-}2LSF}$ =16 (equivalent bit per sample)
B$_{MPEG\text{-}2LSF}$ = 8 -128 kb/s per channel
Frame Size$_{MPEG\text{-}2LSF}$ = 384-1152 samples.

Backwards compatibility with MPEG-1 was one of the major requirement during the development of the multichannel extension, where the multichannel extension is achieved by using the auxiliary data space for the additional channels and matrixing the multichannel data. This requirement imposed a heavy constraint in the design of the multichannel coder [46], resulting in relatively high data rates to achieve good performance. Published test results [47, 48, 49] showed a good performance of MPEG-2 Layer II at 640 kb/s per 5.1 channels. MPEG-2 MC Layer II is currently used in DVD-Video for the PAL systems.

[7] In this context the term frequency channel identifies the frequency resolution of the filter-bank; the term frequency channel is not used in relationship to the signal audio channels.

4.1.1 The MP3 Format

The decrease in data rates, especially for Layer III, made MPEG-2 LSF useful for low bandwidth Internet applications. This led the audio group at the Fraunhofer Institute to create an even lower sampling rate modification of Layer III that they named “MPEG-2.5”. “MPEG-2.5” reduced the sampling rates by another factor of 2 from MPEG-2 LSF Layer III. In “MPEG-2.5” the allowed sampling rates are 12 kHz, 11.025 kHz, and 8 kHz. The addition of these extensions allow Layer III coders to range from samples rates of 8 kHz (“MPEG-2.5”) up to 32 kHz (MPEG-1).

To allow “MPEG-2.5” decoders to work with the same bitstream format as used in MPEG-1 Audio and MPEG-2 LSF, they removed the final bit from header synchword and merged it with the ID bit into a 2-bit ID code. The result was that “MPEG-2.5” decoders work with an 11 bit synchword (rather than 12 bit) but have 2 bits to identify the bitstream format. The “MPEG-2.5” ID codes are [00] for “MPEG-2.5”, [11] for MPEG-1, [10] for MPEG-2 LSF, and [01] reserved for future extensions. Notice that using a [1] as the first bit of the two-bit ID code bit for the prior formats leads to compatibility with the MPEG-1 and MPEG-2 formats. Hence, the “MPEG-2.5” bitstream is identical to that of MPEG-1 Audio and MPEG-2 LSF when those formats are being encoded.

The high quality of the Layer III encoder coupled with the wide range of sample rates and data rates that can be encoded using the MPEG-2 LSF and “MPEG-2.5” extensions made Layer III a natural choice for Internet applications. The so-called “MP3” file format is typically implemented as MPEG-1 Layer III with both of these extensions supported. Low bandwidth users typically use the 16 kHz sampling rate (for a bandwidth of roughly 8 kHz) and encode stereo sound at 32 kb/s. Compared with the CD format (44.1 kHz stereo at 16 bits/sample), this represents a data rate reduction of over a factor of 40 and allows for an entire CD’s worth of music (about 800 MB) to be stored in under 20 MB. Higher bandwidth users are more likely to operate with the full CD 44.1 kHz sample rate at 128 kb/s for “near CD-quality” sound at a data rate reduced by more than a factor of ten from the CD format.

The use of MP3 files for sharing audio over the Internet has spread so widely that it has become the de facto standard for Internet audio. In addition, MP3 players and portable devices are in widespread use for listening to audio and home audio digital components (e.g. CD players, DVD players) increasingly tout MP3 format playback as one of their features.

4.1.2 AC-3

AC-3 was first introduced in 1991 with the film "Batman Returns". In addition to applications in the film industry, the AC-3 multichannel format is currently adopted for the audio specifications of high definition television in North America [50] and DVD-Video. Following the development of its precursor two-channel scheme, AC-2, AC-3 is also known as Dolby Digital. AC-3's parameters specifications as per [50] are as follows:

Number of Audio Channels$_{AC\text{-}3}$ = 1-5.1
Fs $_{AC\text{-}3}$ = 32, 44.1, 48 kHz
R $_{AC\text{-}3}$ =16-24 (equivalent bit per sample)
B $_{AC\text{-}3}$ = 32 – 640 kb/s
Frame Size $_{AC\text{-}3}$ = 1536 samples.

The time to frequency mapping is implemented via a time varying filter bank. For steady state signal a 256-point MDCT is employed while for transients a 128-point MDCT is employed. The psychoacoustic model involves a two-stage process; the first stage uses a full auditory model, which controls the operation of a simplified, parametric stage present both in the encoder and in the decoder. Quantization is done by differentially coding signal exponents, i.e. the signal envelop, and then uniformly quantizing the signal mantissas according to the output of the psychoacoustic model stage.
Published test results [49] for two-channel AC-3, suggest that that good performance can be achieved at a data rate of about 500 kb/s per 5.1 channels.

4.1.3 ISO/IEC MPEG AAC

MPEG-2 Advanced Audio Coding (AAC) was finalised as an ISO/IEC standard in 1997 [28]. AAC also constitutes the core for the time to frequency (T/F) mapping audio coding algorithms of the newly established MPEG-4 standard. For AAC the following parameters are specified (see also [24]):

Number of Audio Channels$_{MPEGAAC}$ = 1-48
Fs $_{MPEGAAC}$ = 8-96 kHz
R $_{MPEGAAC}$ =16-24 (equivalent bit per sample)
B $_{MPEGAAC}$ = up to 576 kb/s per channel
Frame Size $_{MPEGAAC}$ = 1024 samples.

In the main profile configuration, the time to frequency mapping is implemented via a time varying filter bank. For steady state signal a 1024-point MDCT is employed while for transients a 256-point MDCT is employed. The window function is also adapted to the input signal varying between a Kaiser-Bessel derived window [21] and a sine window. The psychoacoustic model is similar to MPEG-1, 2 Model 2. Non-uniform quantization is adopted in AAC with Huffman coding of audio data and differential scale factors.
Based on a number of test results [48, 49], AAC has shown the capability of providing very good performance at the lowest data rates among the audio coding schemes reviewed. At data rates of 64 kb/s per channel it provides very good audio quality. Applications of AAC include digital audio broadcasting in Japan and IBOC in the US.

4.1.4 New Trends in MPEG-4

The scope of MPEG-4 Audio is broader than the scope of MPEG-1 and 2 Audio. The different types of applications that MPEG-4 is addressing, such as telephony and mobile communication, digital broadcasting, internet networks, interactive multimedia, etc., require a high degree of coding efficiency together with flexible access to coded data, including access to subsets of coded data (i.e. scalability of the coded bitstream), and protection against transmission errors. Reflecting the needs of these requirements, the MPEG-4 Audio goals and functionalities include, in addition to highly efficient audio coding, the provision of speech coding to address telephony applications, universal access through scalability of the coded data to address different transmission channel requirements and robustness in error prone environments. Furthermore content-based interactivity through flexible access and manipulation of the coded data and support to synthetic audio and speech through the structured audio, SA, and TTS interface are addressed by the standard functionalities.
Figure 14 shows the typical data rate requirements for different applications versus the bandwidth of the coded signals and which part of the MPEG-4 Audio standard is applicable. Namely, MPEG-4 addresses two basic types of audio, synthetic (TTS and SA) [51] and natural (parametric, code excited linear predictive or CELP, general audio or G/A, and scalable coders) [52, 53]. The synchronization and mix of natural with synthetic audio is called Synthetic/Natural hybrid coding, SNHC. In addition, the AudioBIFS [54] part of the Systems BIFS framework allows for receiver's mixing and postproduction and 3-D sound presentation.
The TTS interface part of MPEG-4 Audio standardizes a transmission protocol for synthesized speech, where TTS systems translate text information into speech so it can be transferred through speech lines such as telephone lines. In addition, TTS systems can be used for services for the visually impaired, automatic voice response systems, etc. The data rates covered by the TTS systems vary between 200 b/s and 1.2 kb/s.
In the SA part of the audio standard, the delivery of synthetic audio is described. This capability allows for

ultra-low data rates (200 b/s as shown in Figure 14) and interactivity at the receiver end. The SA bitstream format specifies a set of synthesis algorithms that describe how to create the sound, and a set of synthesis control parameters that describe which sounds to create. The set of synthesis algorithms, which can generate "instruments", (such as real-life instruments like the flute, violin, etc., or instruments that reflect the sound of ocean waves, or synthetic-hybrid "instruments", etc.) is specified in the SA orchestra language, SAOL. The control parameters that govern the creation of specific sounds are specified in the SA score language, SASL. A format designed to represent banks of wave-tables, the SA audio sample bank format, SASBF, is included in the standard and was developed in collaboration with the musical instrument digital interface, MIDI, manufactures association. Wave-table synthesis is ideal for applications that don't need interaction and require low complexity structure, such as, for example, karaoke applications. This technology allows for the synthesis of a desired sound from look-up tables where particular waveform types are stored. In this case, extremely low data rates can be achieved.

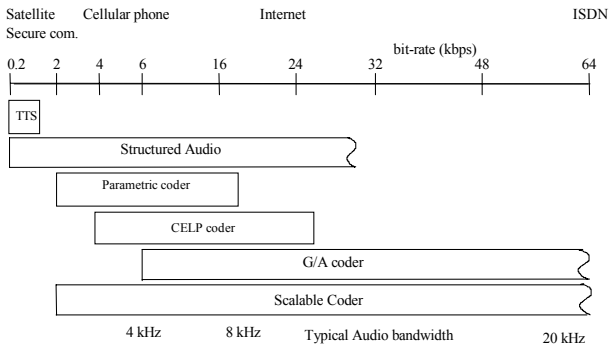

Figure 14:. MPEG-4 Audio data rates and target applications from [55].

General audio covers data rates between 6 kb/s for audio signals with bandwidth of 4 kHz, and 300 kb/s (or above) per channel for signals with bandwidths above 20 kHz for mono to multichannel audio. The work of MPEG-4 Audio in this area represents a continuation of the MPEG-1 and MPEG-2 Audio work with additional tools for addressing natural audio source material. The general audio tools include the basic audio coders such as MPEG-2 AAC and transform-domain weighted interleave vector quantization, TwinVQ. Tools that enhance MPEG-2 AAC efficiency, such as perceptual noise substitution (PNS) and spectral band replication (SBR) are also included.

PNS works in conjunction with MPEG-2 AAC by identifying scale factor bands that consist primarily of noise and transmitting the total noise power rather than the individual spectral coefficients [56]. PNS allows for a parametric description of noise-like signal components. At decoding time the original noise-like spectrum in that scale factor band is replaced ("substituted") by pseudo-random noise with the appropriate signal power. The PNS tool improves the basic quality of the MPEG-2 AAC coder for signal containing noise-like spectral components at data rates below 48 kb/s per stereo channel.

Another tool that allows for a parametric description of the signal is the SBR tool. Based on the fact that there often is a large dependency between the lower and the higher frequency portions of an audio signal spectrum, the high frequency portion is not directly coded, but only control data are transmitted to reconstruct it. SBR is a technology that allows for the bandwidth extension of the frequency components of an audio signal at the receiver side. This method significantly improves the compression efficiency of general audio coders [57].

MPEG-4 provides also the description of parametric coding of general audio signal defined in the harmonic and individual lines plus noise (HILN) tools. A new extension of the capability the HILN parametric coding scheme to higher data rates is in development within the MPEG-4 specifications [58].

Finally, a call for proposals for lossless audio coding was recently issued by the MPEG Committee [59, 60]. The idea behind this call is to extend the general audio coding capability of MPEG-4 Audio to lossless coding.

5. CONCLUSIONS

In this paper we discussed various aspects of multichannel sound and technologies adopted in audio coding. Further discussion on these subjects can be found in [61, 62 and 63]. While a steady increase in storage media capacity and transmission bandwidth and advances in digital audio technology have made high quality multichannel audio a practical alternative to the CD format, we showed that there are still challenges for traditional and new delivery media. Compression algorithms play an important role in the delivery of high quality audio. Increasing the number of channels and added features continue to stress resources for audio storage and transmission. New developments that enhance the traditional performance of waveform and perceptual coding promise a continued improvement in the quality of coding systems at lower data rates.

REFERENCES

[1] ITU-R Recommendation BS.775-1, "Multichannel Stereophonic Sound System with and without Accompanying Picture,"

International Telecommunication Union, Geneva, Switzerland, 1992-1994.

[2] ITU-R Recommendation BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland, 1994.

[3] W. Martens,"The impact of Decorrelated Low-Frequency Reproduction on Auditory Spatial Imagery: Are Two Subwoofers Better than One?," Proceedings of the 16th AES International Conference, pp. 67-77, Rovaniemi 1999.

[4] T. Holman, Private Communications, 1998.

[5] P. Damaske and Y. Ando, "Interaural Crosscorrelation for Multichannel Loudspeaker Reproduction," Acustica, vol. 27, pp. 232-238.

[6] T. Holman, "New Factors in Sound for Cinema and Television," J. Audio Eng. Soc., vol. 39, pp. 529-539 (1991 July/August).

[7] L. Fielder, "Evaluation of the Audible Distortion and Noise Produced by Digital Audio Converters," J. Audio Eng. Soc., vol. 35, pp. 517-535 (1987 July/August).

[8] J. D. Johnston and A. J. Ferreira, "Sum-Difference Stereo Transform Coding," IEEE ICASSP 1992, pp. 569-571.

[9] J. D. Johnston, J. Herre, M. Davis, U. Gbur, "MPEG-2 NBC Audio - Stereo and Multichannel Coding Methods," Presented at the 101st AES Convention, Los Angeles, November 1996, pre-print 4383.

[10] R. G. v.d. Waal and R. N. J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals," IEEE ICASSP 1991, pp. 3601 – 3604.

[11] M. Davis, "The AC-3 Multichannel Coder," presented at the 95th AES Convention, New York, October 1993, pre-print 3774.

[12] N. Jayant, P. Noll "Digital Coding of Waveforms: Principles and Applications to Speech and Video," Prentice-Hall, Englewood Cliffs, 1982.

[13] M. A. Krasner, "Digital Encoding of Speech and Audio Signals Based on the Perceptual Requirements of the Auditory System", Technical Report 535, MIT, Lincoln Laboratory, Lexington, 1979.

[14] J. Princen, A. Johnson, A. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation", Proc. of the ICASSP 1987, pp. 2161-2164.

[15] ISO/IEC 11172-3 "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, Part 3: Audio", 1992.

[16] ATSC, United States Advanced Television Systems Committee Digital Audio Compression (AC-3) Standard, Doc. A/52/10, December 1995.

[17] K. Tsutsui, H. Suzuki, O. Shimoyoshi, M. Sonohara, K. Akagiri, R.M. Heddle, "ATRAC: Adaptive Transform Acoustic Coding for MiniDisc", preprint 3456, 93rd AES-Convention, 1992.

[18] M. Vetterli and J. Kovacevic, "Wavelets and Subband Coding", Prentice Hall, Englewood Cliffs, 1995.

[19] D. Sinha and A. H. Tewfik, "Low Bit-Rate Transparent Audio Compression Using Adapted Wavelets", IEEE Trans. Acoust., Speech, and Signal Processing, 41(12):3463 – 3479, 1993.

[20] J. Princen, J. D. Johnston, "Audio Coding with Signal Adaptive Filterbanks," IEEE Proc. of ICASSP 1995, pp. 3071 - 3074.

[21] L. D. Fielder, M. Bosi, G. A. Davidson, M. Davis, C. Todd, and S. Vernon" AC-2 and AC-3: Low Complexity Transform-Based Audio Coding," in N. Gielchrist and C. Grewin (ed.), Collected Papers on Digital Audio Bit-Rate Reduction, AES 1996, pp. 54-72.

[22] H. J. Nussbaumer, "Pseudo-QMF Filter Bank", IBM Tech. Disclosure Bull., vol. 24, Nov. 1981, pp. 3081-3087.

[23] J. Princen, A. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation", IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, Oct. 1986, pp. 1153-1161.

[24] M. Bosi, K. Brandenburg, S. Quackenbush, K. Akagiri, H. Fuchs, J.Herre, L. Fielder, M.Dietz, Y. Oikawa, G. Davidson, "ISO/IEC MPEG-2 Advanced Audio Coding", JAES, 51, 780 - 792, October 1997.

[25] K. Brandenburg and G. Stoll, "The ISO/MPEG-1 Audio Codec: A Generic Standard for Coding of High Quality Digital Audio", JAES, 42, 780 - 792, October 1994.

[26] ISO/IEC 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 3: Audio", 1994-1997.

[27] ISO/IEC 13818-3 "Information Technology - Generic Coding of Moving Pictures and Associated Audio, Part 7: Advanced Audio Coding, AAC", 1997.

[28] ISO/IEC 14496-3 "Coding of Audio-Visual Objects, Part 3: Audio", 1998.

[29] D. Sinha, J. D. Johnston, "Audio Compression at Low Data Rates Using Signal Adaptive Switched Filter banks", IEEE Proc. of ICASSP 1996, pp. 1053 - 1056.

[30] N. Iwakami, T. Moriya, S. Miki, " High Quality Audio Coding at Less Than 64 kb/s by Using Transform-Domain Interleaved Vector Quantization (Twin-VQ) ", IEEE Proc. of ICASSP 1995, pp. 3095 - 3098.

[31] H. S. Malvar, "Signal Processing with Lapped Transforms," Artech House, Norwood, MA, 1992.

[32] E. Zwicker and H. Fastl, "Psychoacoustics, Facts and Models", Springer 1990.

[33] E. Terhardt, "Calculating Virtual Pitch", Hearing Res., Vol. 1, pp. 155-182, 1979.

[34] H. Fletcher, "Auditory Patterns", Rev. Mod. Phys., Vol. 12, pp.47-55, January 1940.

[35] E. Zwicker, “Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen),” J. Acoust. Soc. of Am., Vol. 33, p. 248, February 1961.

[36] D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane", J. Acoust. Soc. Am., Vol. 33 no. 10, pp. 1344-1356, October 1961.

[37] B. Scharf, "Critical Bands", in *Foundation of Modern Auditory Theory*, New York Academic, 1970.

[38] R. D. Patterson, "Auditory Filter Shapes Derived with Noise Stimuli", J. Acoust. Soc. Am., Vol. 59 no. 3, pp. 640-650, March 1976.

[39] B. C. J. Moore and B. R. Glasberg, "Suggested Formulae for Calculating Auditory-Filter Bandwidths and Excitation Patterns ", J. Acoust. Soc. Am., Vol. 74 no. 3, pp. 750-753, September 1983.

[40] B. C. J. Moore, "Masking in the Human Auditory System", in N. Gilchrist and C. Gerwin (ed.), *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 9-19, AES 1996.

[41] M. Bosi, “Overview of Perceptual Audio Coding”, IEEE Signal Processing Magazine, pp. 43-49, September 1997.

[42] P. Noll and D. Pan, "ISO/MPEG Audio Coding" in N. Jayant (ed.), Signal Compression- Coding of Speech, Audio, Text, Image and Video, World Scientific 1997, pp. 69-118.

[43] J. D. Johnston "Audio Coding with Filter Banks", pages 287-307 in: "Subband and Wavelet Transforms" by A. N. Akansu and M. J. T. Smith (editors), Kluwer Academic Publishers, Norwell 1996.

[44] J. H. Rothweiler, "Polyphase Quadrature Filters - A new Subband Coding Technique", International Conference IEEE ASSP 1983, Boston, pp. 1280-1283.

[45] Schroeder, B. Atal, J. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," JASA, vol. 66(6), pp. 1647-1652, 1979.

[46] M. Bosi, C. Todd, and T. Holman, "Aspects of Current Standardization Activities for High-Quality, Low Rate Multichannel Audio Coding," Proc. of the 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustiscs, 2a.3, October 1993, New Paltz, New York.

[47] ISO/IEC JTC1/SC29/WG11 N1229, "MPEG-2 Backwards Compatible CODECS Layer II and III: RACE dTTb Listening Test Report," Florence, March 1996.

[48] ISO/IEC JTC1/SC29/WG11 N1420, "Overview of the Report on the Formal Subjective Listening Tests of MPEG-2 AAC multichannel audio coding," Maceio', November 1996.

[49] G. Soulodre, T. Grusec, M. Lavoie, and L. Thibault, "Subjective Evaluation of State-of-the-Art Two-Channel Audio Codecs," J. Audio Eng. Soc., vol. 46, pp. 164 - 177 (1998 March).

[50] ATSC, United States Advanced Television Systems Committee Digital Audio Compression (AC-3) Standard, Doc. A/52/10, December 1995.

[51] B. L. Vercoe, W. G. Garnder and E. D. Scheirer, "Structured Audio: The Creation, Transmission, and Rendering of Parametric Sound Representations," Proc. IEEE, Vol. 85 No. 5, pp. 922-940, May 1998.

[52] B. Edler and H. Purnhagen, "Concepts for Hybrid Audio Coding Schemes Based on Parametric Techniques," presented at the 105th AES Convention, San Francisco, Preprint 4808, October 1998.

[53] J. D. Johnston, S. R. Quackenbush, J. Herre and B. Grill, "Review of MPEG-4 General Audio Coding" in *Multimedia Systems, Standards, and Networks,* pp. 131-155, A. Puri and T. Chen (ed.), Marcel Dekker, Inc. 2000.

[54] E. D. Scheirer and R. Väänänen, J. Huopaniemi, "Describing Audio Scenes with the MPEG-4 Multimedia Standard" IEEE Trans. On Multimedia, Vol. 1 no. 3 pp. 237-250, September 1999.

[55] B. Edler, Powerpoint slides shared with the authors, 1997.

[56] J. Herre and D. Schulz, "Extending the MPEG-4 AAC Codec by Perceptual Noise Substitution," presented at the 112th AES Convention, Amsterdam, Preprint 4720, May 1998.

[57] M. Dietz, L. Liljeryd, K. Kjoerling and O. Kunz, "Spectral Band Replication, a novel approach in audio coding," presented at the 112th AES Convention, Munich, Preprint 5553, May 2002.

[58] A. C. den Brinker, E. G. P. Schuijers and A. W. J. Oomen, "Parametric Coding for High-Quality Audio," presented at the 112th AES Convention, Munich, pre-print 5553, May 2002.

[59] ISO/IEC JTC 1/SC 29/WG 11 N5040, "Call for Proposals on MPEG-4 Lossless Audio Coding" Klagenfurt, July 2002.

[60] ISO/IEC JTC 1/SC 29/WG 11 N5040, "Call for Proposals on MPEG-4 1-bit Lossless Audio Coding" Trodheim, July 2003.

[61] "Collected Papers on Digital Audio Bit-Rate Reduction" Neil Gilchrist and Christer Grewin, Editors, Audio Engineering Society 1996.

[62] Proceedings of the AES 17th International Conference on "High-Quality Audio Coding", K. Brandenburg and M. Bosi Co-chairs, Florence September 1999.

[63] M. Bosi and R. E. Goldberg, "Introduction to Digital Audio Coding and Standards", Kluwer Academic Publishers, Dordrecht 2003.