AN OVERVIEW OF MPEG AUDIO

2.1 Introduction

The MPEG Audio compression algorithm is the first International Standard for the digital compression of high-fidelity audio [40,41]. This standard is the result of three years of collaborative work by an international committee of high-fidelity audio compression experts known as the Moving Picture Experts Group (MPEG). The International Organization for Standards and the International Electrotechnical Commission (ISO/IEC) adopted this standard [23] at the end of 1992.

2.2 Features and Applications of MPEG Audio

**Sampling Rate:** The audio sampling rate can be 48 kHz, used in professional sound equipment, 44.1 kHz used in consumer equipment like CD audio or 32 kHz used in some communications equipment.

**Operating Mode:** MPEG 1 audio works for both mono and stereo signals. A technique called joint stereo coding can be used to do more efficient combined editing of the left and right channels of a stereophonic audio signal. The operating modes are

- Single channel
- Dual Channel (two independent channels, for example, containing different language versions of the audio)
- Stereo (no joint stereo coding)
- Joint Stereo

**Predefined Bit-Rates:** The MPEG compressed bit stream can have one of several predefined fixed bit-rates ranging from 32 kb/s per channel to 448 kb/s. A raw PCM
audio stream is about 705 kb/s. Hence 32 kb/s corresponds to a compression ratio of about 22. Normal compression ratio is more like 4:1 (Layer I), 6:1 (Layer II) and 12:1 (Layer III). 96 kb/s is considered transparent for most practical purposes. This means that we will not notice any difference between the original and the compressed signal for rock’n roll or popular music. For more demanding material like piano concerts and such, we will need to go up to 128 kb/s.

**Compression Layers:** The MPEG committee chose to recommend three compression methods and named them Audio Layer I, II, and III. This provides increasing quality/compression ratios with increasing complexity and demands on processing power.

Layer I is the simplest, a polyphase filter bank with a psychoacoustic model. It best suits bit-rates above 128 kb/s per channel. Philips’ Digital Compact Cassette (DCC) uses Layer I at 192 kb/s per channel.

Layer II adds more advanced bit allocation techniques and greater accuracy. It has intermediate complexity and targets bit-rates around 128 kb/s per channel. Possible applications include Digital Audio Broadcasting (DAB).

Layer III adds a hybrid filter bank and non uniform quantization plus advanced features like Huffman coding, higher frequency resolution and bit reservoir technique. It is the most complex but offers the best audio quality, at bit rates around 64 kb/s per channel. This layer suits audio transmission over ISDN.

Thus a wide range of trade-offs between codec complexity and compressed audio quality is offered by the three layers. The reason for recommending three layers was partly that the testers felt that none of these coders was 100% transparent to all material and partly that the best coder (Layer III) was so computation intensive that it would seriously impact the acceptance of the standard.

The specifications say that a valid Layer III decoder shall be able to decode Layer I, II or III MPEG Audio stream. A Layer II decoder shall be able to decode Layer I and Layer II streams. This is the so called “Backward Compatibility” (BC).
The specifications say that a valid Layer III decoder shall be able to decode Layer I, II or III MPEG Audio stream. A Layer II decoder shall be able to decode Layer I and Layer II streams. This is the so called “Backward Compatibility” (BC).

### 2.3 Overview

The basic structures of perceptual audio encoder and decoder are shown in Figs. 2.1 and 2.2. Encoder consists of the following four main parts:

- A time/frequency mapping (filter bank) is used to decompose the input signal into subsampled spectral components. Depending on the filter bank used, these are called subband values (low frequency resolution together with high time resolution) or transform coefficients.
The output of this filter bank or separate calculation of frequency content, is used to calculate an estimate of the actual time dependent masking threshold using rules known from psychoacoustics.

The subband samples or frequency lines are quantized and coded with the aim of keeping the noise, introduced by quantizing, below the masking threshold. Depending on the quantization and coding algorithm, this step is done in very different ways.

In the last step, a frame packing is used to assemble the bit stream, which typically consists of the quantized and coded mapped samples and some side information. Entropy coding is done to remove statistical redundancies.

2.4 Filter Banks

The following list provides a short overview of the most common filter banks [58,59] used for coding of high quality audio signals:

- **Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT)**
  These were the first transforms used in transform coding of audio signals. They implement equally spaced filter banks with at least 128 to 512 bands at a low computational complexity. They do not provide critical sampling, i.e., the number of time/frequency components is greater than the number of time samples represented by one block length. Another disadvantage of these transforms are possible blocking artifacts.

- **Polyphase Filter Banks**
  These are equally spaced filter banks which combine the filter design flexibility of generalized QMF banks with low computational complexity [58]. A polyphase filter bank using 32 bands is used for Layer I and Layer II of the MPEG coder. The main disadvantage of the polyphase filterbank is that, it is not of "perfect reconstruction" type.

- **Modified Discrete Cosine Transform (MDCT)**
  MDCT using time domain aliasing cancellation is proposed in [56,60]. This transform combines with a good frequency resolution provided by a sine window and the
computational efficiency of a fast FFT like algorithm [59]. Typically 128 to 512 equally spaced bands are used.

- Hybrid structures (eg. Polyphase +MDCT)
  Using hybrid structures as proposed in [61], it is possible to combine different frequency resolution at different frequencies with moderate implementation complexity. A hybrid scheme consisting of a polyphase filter bank and a MDCT is used in Layer III. However it does not exploit the human auditory system’s frequency dependent behaviour.

- Quadrature Mirror Filters, QMF tree filter banks.
  Different frequency resolutions at different frequencies is possible. Typical QMF tree filter banks uses up to 32 bands. The computational complexity is also low. The advantage of QMF filterbanks is that near perfect stop band rejection is possible. Theoretically MDCT and polyphase filter banks belong to the same class of time to frequency domain mappings, called Lapped Orthogonal Transform [58].

### 2.4.1 Polyphase Filterbank

The polyphase filter bank is common to all three layers of MPEG Audio. This filter bank divides the audio signal into 32 equal width frequency subbands. It should be noted that the polyphase filterbank and its inverse are not lossless transformations. Even without quantization, the inverse transformation cannot perfectly recover the original signal.

The ISO / MPEG Audio standard [23] describes steps for computing the polyphase filterbank, analysis and synthesis algorithms. The analysis algorithm is given by the following equation:

\[
S_i[t] = \sum_{k=0}^{63} \sum_{j=0}^{7} M[i][k] \times (C[k + 64 j] \times [k + 64 j])
\]  

(2.1)

where \(i\) is the subband index and ranges from 0 to 31; \(S_i[t]\) is the filter output sample for subband \(i\) at time \(t\), where \(t\) is an integer multiple of 32 audio sample intervals;
C[n] is one of 512 coefficients of the analysis window defined in the ISO/MPEG audio standard: X[n] is an audio input sample read from a 512 sample buffer; and

\[ M[i][k] = \cos \left( \frac{(2i + 1) \times (k - 16) \times T}{64} \right) \]  

(2.2)

are the analysis matrix coefficients.

The function within the parenthesis in Eq. (2.1) is independent of the value of i, and \( M[i][k] \) is independent of j, so the 32 filter outputs need \( 512 \times (32 \times 64) = 2,560 \) multiplications and \( (64 \times 7) + (32 \times 63) = 2,464 \) additions or roughly 80 multiplications and additions per output.

However, the polyphase filterbank is one of the most computational intensive operations in MPEG coding. For example, MPEG audio decoding showed that the polyphase synthesis operation represented 40% of the overall decoding time. Hence, fast algorithms are of prime importance here, especially for applications such as real time audio encoding and decoding. Substantially further reductions in multiplications and additions are possible with a fast Discrete Cosine Transform or a Fast Fourier Transform implementation. For example, the original 2048 multiply - accumulate operations in the matrixing operation, can be reduced to 80 multiplications and 209 additions by using 32 point Lee's fast DCT algorithm [62]. Overall this reduces the original \( 512 \times (32 \times 64) = 2,560 \) multiplications down to \( 512 + 80 = 592 \) and the additions from \( (64 \times 7) + (32 \times 63) = 2,464 \) down to \( (64 \times 7) + 209 = 657 \) or roughly 20 multiplications and additions per output. Note also that this polyphase filterbank is critically sampled. For every 32 input samples, the filterbank produces 32 output samples.

Equation 2.1 can be rewritten as

\[ S_i[t] = \sum_{n=0}^{511} X[t-n] \times H_i[n] \]  

(2.3)

where \( X[t] \) is an audio sample at time t, and

\[ H_i[n] = h[n] \times \cos \left( \frac{(2i + 1) \times (k - 16) \times \pi}{64} \right) \]  

(2.4)

with \( h[n] = -C[n] \) if the integer part of \((n/64)\) is odd and \( h[n] = C[n] \) otherwise, for \( n = 0 \) to 511. In this notation, each subband of the filterbank has its own band pass
filter response, $H_i[n]$. The coefficients, $h[n]$, correspond to the prototype low-pass filter response for the polyphase filterbank. Eq.(2.4) clearly shows that each is a modulation of the prototype response with a cosine term to shift the low pass response to the appropriate frequency band. Hence these are called polyphase filters. The polyphase analysis and synthesis algorithms in pseudo code are following:

<table>
<thead>
<tr>
<th>ANALYSIS ALGORITHM</th>
<th>SYNTHESIS ALGORITHM</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Begin</td>
<td>• Begin</td>
</tr>
<tr>
<td>• for $i=511$ down to 32 do</td>
<td>• Input 32 new subband samples $S_i$</td>
</tr>
<tr>
<td>$X[i]=X[i-32]$</td>
<td>$i=0 \ldots \ldots 31$</td>
</tr>
<tr>
<td>• for $i=31$ down to 0 do</td>
<td>• Shifting for $i=1023$ down to 64 do</td>
</tr>
<tr>
<td>$X[i]=\text{next_input_audio_sample}$</td>
<td>$V[i]=V[i-64]$</td>
</tr>
<tr>
<td>• Window by 512 coefficients, produce vector $Z$ for $i=0$ to 511 do $Z_i=C_i \times X_i$</td>
<td>• Matrixing for $i=0$ to 63 do</td>
</tr>
<tr>
<td>• Partial calculation for $k=0$ to 63 do</td>
<td>$V_i=\sum_{k=0}^{31} N_{ik} \times S_k$ where</td>
</tr>
<tr>
<td>$Y_k=\sum_{j=0}^{7} Z_{k+64} \times j$</td>
<td>$N_{ik}=\cos\left[\frac{(16+i) \times (2k+1) \times \pi}{64}\right]$</td>
</tr>
<tr>
<td>• Calculate 32 samples by matrixing for $i=0$ to 31 do</td>
<td>• Build a 512 values vector $U$</td>
</tr>
<tr>
<td>$S_i=\sum_{k=0}^{63} M_{ik} \times Y_k$ where</td>
<td>for $i=0$ to 7 do; for $j=0$ to 31 do</td>
</tr>
<tr>
<td>$M_{ik}=\cos\left[\frac{(2i+1) \times (k-16) \times \pi}{64}\right]$</td>
<td>$U[i<em>64+j]=V[i</em>128+j]$</td>
</tr>
<tr>
<td>• Output 32 subband samples</td>
<td>$U[i<em>64+32+j]=V[i</em>128+96+j]$</td>
</tr>
<tr>
<td>• Output 32 reconstructed PCM samples</td>
<td>• Window by 512 coefficients, produce vector $W$.</td>
</tr>
<tr>
<td></td>
<td>for $i=0$ to 511 do</td>
</tr>
<tr>
<td></td>
<td>$W_i=U_i \times D_i$</td>
</tr>
<tr>
<td></td>
<td>• Calculate 32 samples for $j=0$ to 31 do</td>
</tr>
<tr>
<td></td>
<td>$S_j=\sum_{i=0}^{15} W_{j+32} \times i$</td>
</tr>
<tr>
<td></td>
<td>• Output 32 reconstructed PCM samples</td>
</tr>
</tbody>
</table>
2.5 Psychoacoustic Principles

The number of bits needed to represent an audio signal can be reduced without affecting the perceptual quality by examining the perception of sound by a human listener, identifying the components that will not be audible and throwing these components.

2.5.1 A walk through the Human auditory system

The main components of the human auditory system are shown in Fig. 2.3 [19].

Sound waves incident from different angular positions are spectrally shaped by the pinna in a direction dependent manner. The ear canal further filters the waveform, before it passes through two small bones, and on to the cochlea. The ear canal is the resonant cavity between the outer and middle ear, which has a resonance at around 3-5 kHz. Hence it attenuates higher and lower frequencies. Cochlea is a fluid filled coil within the ear, and is partially partitioned by the Basilar Membrane (BM) (see Fig. 2.4). Sensory cells (outer hair cells and inner hair cells) are distributed along the basilar membrane (see Fig. 2.5).
basilar membrane

sound as pressure wave in fluid

sound as standing wave on basilar membrane
cross section > next figure

lower frequency & higher frequency resonances

basilar membrane

cochlea (straightened)

fluid flow

Fig. 2.4: Cochlea

outer hair cells

inner hair cells

sound as neural impulses in auditory nerve

stereocilia

inner hair cell

auditory nerve

outer hair cells

basilar membrane (cross section)

Fig. 2.5: Cross section of basilar membrane
The different points of the basilar membrane resonate at different frequencies. Thus the BM acts as a spectrum analyser. The spacing of frequency resonances along the BM is not linear with frequency. The resonant frequencies of various points along the BM are shown in Fig 2.6. The scale that relates the resonant frequency to position on BM is called the Bark scale or Critical Band scale [20,63,64]. It approximates to a \(-\log\) scale.

\[\text{Fig.2.6: Resonant frequencies of various points along the basilar membrane}\]

Sound waves enter the cochlea and set the fluid within it in motion. The movement of the fluid stimulates the hair cells of BM. Auditory nerve endings carry these stimuli to the auditory centre of the brain. Interpretations of these impulses by the brain results in hearing.

2.5.2 Absolute Threshold of Hearing

The absolute threshold of hearing [20,63,64] characterizes the amount of energy needed in a pure tone such that a listener in a noiseless environment can detect it. The absolute threshold is typically expressed in terms of dB SPL (Sound Pressure Level). The SPL gives the level (intensity) of sound pressure in decibels (dB) relative
Fig 2.7: Hearing threshold in quiet

The quiet threshold is well approximated by the non-linear function

\[ T_q(f) = 3.64(\frac{f}{1000})^{-0.8} - 6.5 \exp(-0.6(\frac{f}{1000} - 3.3)^2) + 10^{-3}(\frac{f}{1000})^4 \text{ (dB SPL)} \]  

(2.5)

which is representative of a young listener with acute hearing. When applied to signal compression, \( T_q(f) \) could be interpreted as a maximum allowable energy level for coding distortions introduced in the frequency domain. Variation of threshold in quiet with frequency is given in Fig.2.7. It is the outer ear canal that is responsible for the high sensitivity of hearing at frequencies near 4 kHz, indicated by the dip of threshold in quiet around 4 kHz.

2.5.3 Critical Bands

Using the absolute threshold to shape the coding distortion spectrum represents the first step toward perceptual coding. The detection threshold for spectrally complex quantization noise is a modified version of the absolute threshold, with its shape determined by the stimuli present at any given time. Since stimuli are in general time
varying, the detection threshold is also a time varying function of the input signal. Ear performs spectral analysis as follows. A frequency-to-place transformation takes place in the cochlea (inner ear), along the basilar membrane. A sound wave generated by an acoustic stimulus moves the eardrum and the attached ossicular bones, which in turn transfer the mechanical vibrations to the cochlea. Once excited by mechanical vibrations at its input, the cochlear structure induces travelling waves along the length of the basilar membrane. Neural receptors are connected along the length of the basilar membrane. The travelling wave generate peak responses at frequency-specific membrane positions, and therefore different neural receptors are effectively tuned to different frequency bands according to their locations. For sinusoidal stimuli, the travelling wave on the basilar membrane propagates from the oval window, until it nears the region with a resonant frequency. The wave then slows, and the magnitude increases to a peak. The location of the peak is referred to as the best place or characteristic place for the stimulus frequency, and the frequency that best excites a particular place is called the 'best frequency' or 'characteristic frequency'. Thus a frequency-to-place transformation occurs. An example is given in Fig.2.8.

![The frequency-to-place transformation](image_url)

**Fig.2.8:** The frequency-to-place transformation

The above figure gives a schematic representation of the travelling wave envelopes that occur in response to an acoustic tone complex containing sinusoids of 400, 1600, and 6400 Hz. Peak responses for each sinusoid are localized along the
membrane surface, with each peak occurring at a particular distance from the oval window (cochlear window). As a result of the frequency-to-place transformation, the cochlea can be viewed from a signal-processing perspective as a bank of highly overlapping band pass filters. The cochlear filter pass bands are of non-uniform bandwidth, and the bandwidth increases with increasing frequency.

Table 2.1. Idealized critical band filter bank

<table>
<thead>
<tr>
<th>Band No.</th>
<th>Lower Edge (Hz)</th>
<th>Upper Edge (Hz)</th>
<th>BW (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>300</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>510</td>
<td>110</td>
</tr>
<tr>
<td>6</td>
<td>510</td>
<td>630</td>
<td>120</td>
</tr>
<tr>
<td>7</td>
<td>630</td>
<td>770</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>770</td>
<td>920</td>
<td>150</td>
</tr>
<tr>
<td>9</td>
<td>920</td>
<td>1080</td>
<td>160</td>
</tr>
<tr>
<td>10</td>
<td>1080</td>
<td>1270</td>
<td>190</td>
</tr>
<tr>
<td>11</td>
<td>1270</td>
<td>1480</td>
<td>210</td>
</tr>
<tr>
<td>12</td>
<td>1480</td>
<td>1720</td>
<td>240</td>
</tr>
<tr>
<td>13</td>
<td>1720</td>
<td>2000</td>
<td>280</td>
</tr>
<tr>
<td>14</td>
<td>2000</td>
<td>2320</td>
<td>320</td>
</tr>
<tr>
<td>15</td>
<td>2320</td>
<td>2700</td>
<td>380</td>
</tr>
<tr>
<td>16</td>
<td>2700</td>
<td>3150</td>
<td>450</td>
</tr>
<tr>
<td>17</td>
<td>3150</td>
<td>3700</td>
<td>550</td>
</tr>
<tr>
<td>18</td>
<td>3700</td>
<td>4400</td>
<td>700</td>
</tr>
<tr>
<td>19</td>
<td>4400</td>
<td>5300</td>
<td>900</td>
</tr>
<tr>
<td>20</td>
<td>5300</td>
<td>6400</td>
<td>1100</td>
</tr>
<tr>
<td>21</td>
<td>6400</td>
<td>7700</td>
<td>1300</td>
</tr>
<tr>
<td>22</td>
<td>7700</td>
<td>9500</td>
<td>1800</td>
</tr>
<tr>
<td>23</td>
<td>9500</td>
<td>12000</td>
<td>2500</td>
</tr>
<tr>
<td>24</td>
<td>12000</td>
<td>15500</td>
<td>3500</td>
</tr>
<tr>
<td>25</td>
<td>15500</td>
<td>22000</td>
<td>6500</td>
</tr>
</tbody>
</table>
The critical bandwidth is a function of frequency that quantifies the cochlear filter pass bands. Approximate critical bands of auditory system are shown in Table 2.1. The critical band can be loosely defined as the bandwidth at which subjective responses change abruptly. For example, the perceived loudness of a narrowband noise source at constant sound pressure level remains constant even as the bandwidth is increased up to the critical bandwidth. The loudness then begins to increase. For an average listener, the critical bandwidth is approximated by

\[ BW_c(f) = 25 + 75 \left[ 1 + 1.4(f/1000)^2 \right]^{0.69} \text{Hz} \]  

(2.6)

A distance of one critical band is referred as one bark in literature. The function

\[ z(f) = 13 \arctan(0.00076f) + 35 \arctan\left(\frac{f}{7500}\right)^2 \]

(2.7)

is used to convert from frequency scale to bark scale.

2.5.4 Masking

Masking \cite{1,20,21,63,64} refers to a process where one weak sound is rendered inaudible because of the presence of another strong sound. Simultaneous masking is a frequency domain phenomenon within critical bands when two or more stimuli are simultaneously present.

The mechanism underlying simultaneous masking phenomena is that the presence of a masker creates an excitation of sufficient strength on the basilar membrane at the critical band location to effectively block the transmission of a weaker signal. There are two types of masking, namely Noise Masking Tone (NMT) and Tone Masking Noise (TMN).

a. **Noise - Masking - Tone (NMT)**

In the NMT scenario, a narrow band noise masks a tone within the same critical band, provided that the intensity of the masked tone is below a predictable threshold directly related to the intensity of the masking noise. At the threshold of detection for the masked tone, the minimum signal-to-mask ratio (SMR), i.e. the smallest difference between the intensity (SPL) of the masking noise and the intensity of the masked tone occurs when the frequency of the masked tone is close to the masker's centre.
frequency. In most studies, the minimum SMR tends to lie between -5 and +5 dB. Fig. 2.9 (a) shows the NMT scenario. In this figure, a critical band noise masker centered at 410 Hz with an intensity of 80-dBSPL masks a 410 Hz tone, and the resulting SMR at the threshold of detection is 4 dB. Masking power decreases for probe tones above and below the frequency of the SMR tone, in accordance with a level-and frequency-dependent spreading function (discussed in Section 2.5.6).

b. Tone- Masking- Noise (TMN)

In the case of TMN, a pure tone occurring at the center of a critical band masks noise of any sub critical bandwidth or shape, provided the noise spectrum is below a predictable threshold related to the strength of the masking tone. At the threshold of detection for a noise band masked by a pure tone, it was found that the minimum SMR, i.e. the smallest difference between the intensity of the masking tone and the intensity of the masked noise, occurs when the masker frequency is close to the center frequency of the probe noise. Minimum SMR for TMN tends to lie between 21-28 dB. This is shown in Fig.2.9 (b). In the figure, a narrow band noise of one Bark band width centered at 1 kHz is masked by a 1 kHz tone of intensity 80 dB SPL.

2.5.5 Asymmetry of Masking

The NMT and TMN examples in Fig. 2.9 clearly show an asymmetry in masking power between the noise masker and the tone masker. In spite of the fact that both maskers are presented at a level of 80 dB SPL, the associated threshold SMR's differ by about 20 dB. For each temporal analysis interval, a codec's perceptual model should identify across the frequency spectrum noise- like and tone-like components. The model should apply the appropriate masking relationships in a frequency specific manner. In conjunction with the spread of masking, NMT and TMN properties can be used to construct a global masking threshold.

2.5.6 The Spread of Masking

Simultaneous masking effects are not band limited to within the boundaries of a single critical band. Interband masking also occurs, i.e., a masker centred within one
critical band has some predictable effect on detection thresholds in other critical bands. This effect, also known as spread of masking, is often modeled in coding applications by an approximately triangular spreading function.

Fig. 2.9 (a): Noise masking tone.

Fig. 2.9 (b): Tone masking noise
An analytical expression for the spreading function can be given as:

$$SF dB (x) = 15.81 + 7.5 (x + 0.474) - 17.5(1+(x+0.471)^2)^{1/2} \text{ dB.}$$

(2.8)

where 'x' has unit of Barks.

After critical band analysis is done and spread of masking has been accounted for, masking thresholds in perceptual coders are established by the relations.

$$T_{HN} = E_T - 14.5 - B$$

(2.9)

$$T_{HT} = E_N - K$$

(2.10)

$T_{HN}$ and $T_{HT}$ are noise and tone masking thresholds, respectively, due to TMN and NMT. $E_N$ and $E_T$ are critical band noise and tone masker energy levels respectively. $B$- critical band number; $K$ is typically set to 5 dB.

The above equations capture only the contributions of individual tone like or noise-like maskers. In the actual coding scenario each frame typically contains a collection of both masker types. After they have been identified, these individual masking thresholds are combined to form a global masking threshold. The global masking threshold comprises an estimate of the level at which quantization noise becomes just noticeable.

![Schematic representation of simultaneous masking](image)

Fig.2.10: Schematic representation of simultaneous masking
Notions of critical bandwidth and simultaneous masking in the audio coding context give rise to some convenient terminology illustrated in Fig. 2.10. Consider the case of a single masking tone occurring at the center of a critical band. This generates an excitation along the basilar membrane that is modeled by a spreading function and a corresponding masking threshold. For the band under consideration, the minimum masking threshold denotes the spreading function in-band minimum. Assuming the masker is quantized using an m-bit uniform scalar quantizer, noise might be introduced at level m. SMR and noise-to-mask ratio (NMR) denote the log distances from the minimum masking threshold to the masker and noise levels respectively.

2.6 Psychoacoustic Model Implementation (Layer I)

In this section emphasis is given on the implementation details of the Psychoacoustic model I, as has been used in common by both MPEG and the wavelet based codecs proposed in this thesis.

Step I Spectral analysis and SPL Normalisation

The goal of this step is to obtain a high resolution spectral estimate of the input signal. A 512 point FFT (Fast Fourier Transform) is used for the purpose.

First the input samples, $s(n)$, are normalised according to the FFT length $N$ and the number of bits per sample $b$.

$$X(n) = \frac{s(n)}{N(2^b-1)}$$  \hspace{1cm} (2.11)

Next, a power spectral density (PSD) estimate is obtained using 512 point FFT.

$$P(k) = P_N + 10 \log_{10} \left| \sum_{n=0}^{N-1} w(n)x(n)e^{-j2\pi kn/N} \right|^2 \quad 0 < k < \frac{N}{2}$$  \hspace{1cm} (2.12)

where the power normalization term $P_N$ is fixed at 90 dB and the Hann window, $w(n)$ is defined as

$$w(n) = \frac{1}{2} \left[ 1 - \cos \left( \frac{2\pi n}{N} \right) \right]$$  \hspace{1cm} (2.13)

Since playback levels are unknown during psychoacoustic analysis, the normalisation term $P_N$ is used to estimate SPL (Sound Pressure Level) conservatively.
from input power. For example, a full scale sinusoid which is precisely resolved by the 512-point FFT in bin \( k_0 \) will yield a spectral line, \( P(k_0) \), having 84 dB SPL. With 16-bit sample resolution, SPL estimates for very low amplitude input signals will be at or below the absolute threshold.

**Step 2: Tonal and Noise Masker Identification**

Local Maxima in the PSD which exceed neighbouring components within a certain bark distance by at least 7 dB are taken as tonal components. The tonal set \( S_T \) is defined as

\[
S_T = \left\{ P(k) \left\{ \begin{array}{l}
    P(k) > P(k + 1) \\
    P(k) > P(k + \Delta_k) + 7 \text{ dB}
\end{array} \right. \right\}
\]

(2.14)

where

\[
\Delta_k \in \begin{cases}
    2 & 2 < k < 63 \\
    [2,3] & 63 \leq k \leq 127 \\
    [2,6] & 127 \leq k \leq 256
\end{cases}
\]

0.17 - 5.5 kHz

5.5 - 11 kHz

11 - 20 kHz

The tonal maskers \( P_{TM}(k) \) are computed for the peaks obtained from the above step and listed in \( S_T \) as

\[
P_{TM}(k) = 10 \log_{10} \sum_{j=-1}^{1} 10^{0.1P(k+j)}
\]

(2.15)

The remaining spectral components in each critical band not within a certain bark distance (as explained earlier) of tonal components are added up into a single noise masker,

\[
P_{NM}(k) = 10 \log_{10} \sum_{j} 10^{0.1P(j)} \text{ dB}
\]

(2.16)

\( \forall P(j) \not\in \{ P_{TM}(k, k \pm 1, k \pm \Delta_k) \} \)

where \( \bar{k} \) is defined as the geometric mean spectral line of the critical band, i.e.,

\[
\bar{k} = \left( \prod_{j=1}^{u} j \right)^{1/(u-1)}
\]

where \( l \) and \( u \) are the lower and upper spectral line boundaries of the critical band, respectively.
**Step 3 Decimation of Maskers**

In this step the number of maskers is reduced using two criteria. First, any tonal or noise masker below absolute threshold are discarded. That is, only maskers that satisfy the inequality given in Eq. 2.17 are retained.

\[ P_{TM, NM}(k) \geq T_q(k) \]  

(2.17)

Next, a sliding 0.5- Bark-wide window is used to replace any pair of maskers occurring within a distance of 0.5 Bark by the stronger of the two.

**Step 4 Calculation of Individual Masking Threshold**

Each individual masking threshold represents the masking contribution at a particular frequency bin, say \( i \), (due to a tone or noise masker located at frequency bin, say \( j \)). Total masking thresholds are given by,

\[ T_{TM}(i, j) = P_{TM}(j) - 0.275z(j) + SF(i, j) - 6.025 \text{ dB SPL} \]  

(2.18)

where \( P_{TM}(j) \) is the SPL of the tonal masker in frequency bin \( j \), \( z(j) \) is the bark frequency of bin \( j \), and \( SF(i, j) \) is the spread of masking from masker bin \( j \) to maskee bin \( i \), and is given by the expression,

\[ SF(i, j) = \begin{cases} 
17\Delta_z - 0.4P_{TM}(j) + 11 & -3 \leq \Delta_z < -1 \\
(0.4P_{TM}(j) + 6)\Delta_z & -1 \leq \Delta_z < 0 \\
-17\Delta_z & 0 \leq \Delta_z < 1 \\
(0.15P_{TM}(j) - 17)\Delta_z - 0.15P_{TM}(j) & 1 \leq \Delta_z < 8 
\end{cases} \]

Individual noise masker thresholds are given by,

\[ T_{NM}(i, j) = P_{NM}(j) - 0.175z(j) + SF(i, j) - 2.025 \text{ dB SPL} \]  

(2.19)

**Step 5 Calculation of Global and Minimum Masking Thresholds**

The individual masking thresholds are combined to estimate a global masking threshold for each frequency. Global masking threshold is given by the sum,

\[ T_e(i) = 10 \log_{10} \left( 10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)} \right) \text{ dB SPL} \]  

(2.20)
where,

\[ T_q(i) \] : absolute threshold for frequency bin \( i \);

\[ T_{TM}(i,l) \text{ and } T_{NM}(i,m) \] : individual masking thresholds from step 4;

\( L \) and \( M \) : numbers of tonal and noise maskers, respectively, identified during step 3.

The minimum value of global masking threshold in each critical band is taken as the minimum masking threshold of that particular critical band. From this the SMR (Signal-to-Mask Ratio) in each critical band is calculated. The bit allocation is then done on the basis of SMR s calculated in various subbands.

2.7 Summary

Basic concepts of ISO/MPEG audio coding standard are presented in this chapter. Human auditory system and psychoacoustic properties like absolute threshold of hearing, critical bands, masking etc. are discussed briefly. Implementation details of MPEG Layer I psychoacoustic model is also described here. Draw backs of MPEG standard by using uniform filterbank for time/frequency analysis are:

- Analysis does not match with the properties of speech and audio signals.

  Speech and audio signals are non-stationary and hence fixed time-frequency resolution windows are not suitable for their analysis. These signals call for narrow windows in the analysis of high frequency components and wide windows in the analysis of low frequency components.

- Analysis does not match with the properties of human auditory system.

  Human ear analyses various frequency components with different resolutions. Critical bands are non-uniform and the bandwidth of the critical bands increases as the frequency increases. That is, our ear analyses high frequency components with good time resolution and low frequency components with good frequency resolution.
Filterbank and its inverse do not yield perfect reconstruction. This introduces errors even in the absence of quantization error.

Hence, for the efficient exploitation of perceptual irrelevancies in audio coding, analysis filterbank should match to the properties of human auditory system. MPEG standard faces with a serious artifact known as pre-echo distortion, because of employing uniform filterbank for analysis purpose. Pre-echo is noise, spread out over some time, even before the music event, causing the noise. To avoid pre-echo distortion, high frequency components should be analysed with narrow windows (good time resolution) and low frequency components should be analysed with wider windows (good frequency resolution).

Since, critical bands are of almost constant Q type, in order to fully exploit the masking thresholds in various frequency bands and to place the quantization noise in the least sensitive regions of the spectrum, the analysis filterbank should be of either constant Q type or whose subbands mimic the various critical bands of the human auditory system.

Major attraction of wavelet analysis is that it uses basis functions which are well localized in time and frequency. Hence, wavelet transform will concentrate the energy of a signal in very few transform coefficients. Non-uniform filterbank is used for the implementation of wavelet transform. A filterbank emulating the human hearing process can be constructed. Unlike Fourier analysis in which basis functions are only sines and cosines, a number of wavelet basis functions are available in the literature. Hence, wavelet analysis is more flexible in the sense that each audio frame can be represented with the most matching wavelet basis. DWT provides a good approximation to the Karhunen-Loeve transformation (KLT) of a wide class of stationary and non-stationary process. In this transform, high frequency components of the signal are analysed using narrow windows and low frequency components are analysed using wide windows. Hence, wavelet analysis is readily applicable to the task of perceptual audio coding. Brief theory of wavelets and its implementation details are discussed in the next chapter. Wavelet based perceptual audio coding schemes with various features are proposed in Chapters 4-6.