國立交通大學

資訊工程學系

碩士論文

MPEG-4 AAC 與 MP3 的聽覺感官模型設計

Design of Psychoacoustic Model for MPEG-4 Advanced Audio Coding and MPEG Layer III

研究生：邱　挺

指導教授：劉啟民　教授

李文傑　教授

中華民國 九十三 年 六 月

MPEG-4 AAC 與 MP3 的聽覺感官模型設計

Design of Psychoacoustic Model for MPEG-4 Advanced Audio Coding and MPEG Layer III

研 究 生：邱　挺　　Student：Ting Chiou

指導教授：劉啟民　　Advisor：Dr. Chi-Min Liu

李文傑　　Dr. Wen-Chieh Lee

A Thesis
Submitted to Institute of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National ChiaoTung University
in partial Fulfillment of the Requirements
for the Degree of Master in
Computer Science and Information Engineering
June 2004
HsinChu, Taiwan, Republic of China

中華民國 九十三 年 六 月

MPEG-4 AAC 與 MP3 的聽覺感官模型設計

學生：邱挺　　指導教授：劉啓民 博士
李文傑 博士

國立交通大學資訊工程所碩士班

中文論文摘要

本論文提出一個有效率的聽覺感官模型，以 filterbank 取代 FFT 的計算，使其可以大量減少計算時間，並且在本文內提出了頻率上的 attack 偵測。本文更利用 energy floor 來估測量化誤差以及提出一個有效率的 energy floor 估算方式，而能有效的減少 fishy noise 達到良好的壓縮品質。最後將此聽覺感官模型實作在 NCTU-AAC 及 NCTU-MP3 上，效率上均獲卓越的提升比傳統的聽覺感官模型達到 60%以上的改進。並且使用 ODG 評斷品質，而本論文的方法均能獲得 0.2 的進步在 MPEG12 bitstream。

Design of Psychoacoustic Model for MPEG-4 Advanced Audio Coding and MPEG Layer III

Student: Ting Chiou Advisor: Dr. Chi-Min Liu

Dr. Wen-Chieh Lee

Institute of Computer Science and Information Engineering
National ChiaoTung University

ABSTRACT

This thesis presents an efficient psychoacoustic model providing better quality than the psychoacoustic model II. This thesis considers the design of the psychoacoustic models from two aspects. First, we improve the psychoacoustic model from the aspect of varying tonal and noise masking offset with bands and energy normalization to suppress the distortion, which is called the fishy noise or the birdie noise, caused by the overestimated masking in the harmonic-rich signals. Second, we consider the design issue in implementing the psychoacoustic model in the filterbank used in MP3 and AAC instead of the independent FFT to reduce the computing complexity and storage. The efficient psychoacoustic model provides 60 percentage performance gain compared to the psychoacoustic model II in MPEG-2/4 AAC and MP3. For the quality comparison based on Objective Difference Grade (ODG) and the subjective test, the efficient psychoacoustic model provides quality gain of 0.26 at 128k bit rates and 0.3 at 112k bit rate for MPEG testing bitstream in NCTU-AAC.

致謝

感謝劉啓民老師兩年來的栽培及李文傑博士給予的指導，實驗室的楊宗瀚學長、同學許瀚文、蕭又華、彭康硯和張子文，以及學弟陳立偉和蘇明堂的協助，在研究上提供我寶貴的意見，讓我在專業知識及研究方法獲得非常多的啓發。最後，感謝我的父母與家人及系上同學，在我研究所兩年的生活中，給予我無論在精神上以及物質上的種種協助，使我能全心全意地在這個專業的領域中研究探索在此一併表達個人的感謝。

Contents

Figure List

Table List

Chapter 1 Introduction

During the last decade, analog audio as FM-quality audio is gradually fading out and high-fidelity digital audio like CD-quality audio is going to dominate audio. Moreover, digital audio can be used in conjunction with network, wireless, and multimedia. Nevertheless, digital audio relatively needs some demands which are reducing the channel bandwidth, limiting the storage capacity, and low cost. For the purposes of resolving the above demands, there are numerous researchers having devoted to the development of algorithms for perceptually transparent coding of digital audio. And in consequence, the considerable audio coding algorithms are presented for the transparent CD-quality digital audio.

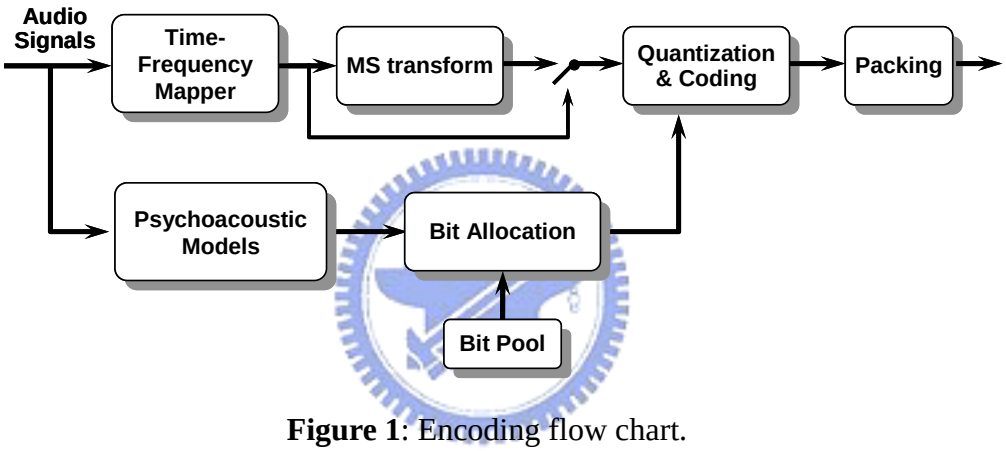

Figure 1: Encoding flow chart.

As illustrated in Figure 1, audio signals are segmented into overlapped blocks and transformed into frequency domain through the time-frequency mapper. The L/R signals are transformed to M/S signals if signals achieve a threshold. Those signals are then quantized and coded with the parameters decided by the bit allocation. The psychoacoustic model analyzes the signal contents and calculates the associated perceptual information on the human auditory system. According to the perceptual information and the available bits, the bit allocation decides the suitable quantization manner to fit the bit rate. The packing module packs all the coded information by the standard format.

Therefore, the psychoacoustic model is considerable critical to an encoder. First, the psychoacoustic model according to the human auditory system calculates SMR for the bit allocation such that it minimizes quantization error and costs fewer bits. Second, psychoacoustic model decides the suitable block type in order to obtain better resolution

of frequency or time for any input signal. Last, the psychoacoustic model calculates PE for M/S which can effectively reduce redundant information between channels.
Conventional psychoacoustic model uses the Fast Fourier Transform (FFT) and prediction magnitude and phase as tonality index. Using the foregoing tonality forms the Signal-to-Masking Ratio (SMR) consisting of noise masking and tone masking.
Therefore, there are three issues in conventional psychoacoustic model. First, a non-consistent spectrum is between analysis and coding. Second, the noise masking effect is stronger than tone masking effect but the energy is dominated by the tone that will cause the overestimation of masking threshold. Third, the conventional psychoacoustic can only detect attack in the time domain without the attack in the frequency domain.

Therefore, the thesis is based on this concept which replaces FFT with MDCT [1][2] in the filterbank. Moreover, only the noise masking effect is considered to calculate the masking threshold. Detection of tonal attack band and tone-rich signal is proposed in the thesis. The proposed psychoacoustic model can speed up 70% in AAC and 65% in MP3. The quality has also improved 0.2 in AAC and 0.1 in MP3 than conventional psychoacoustic model.

The thesis is organized as follows. Chapter 2 introduces the concept of the psychoacoustic principle and detailed psychoacoustic model II. Chapter 3 reviews the filterbank and implementation in AAC and MP3. In chapter 4, the efficient psychoacoustic model and experiments are described. Finally, chapter 5 is the conclusion and following reference.

Chapter 2 Psychoacoustic Model

2.1 Psychoacoustic Principle

The purpose of the psychoacoustic model is to characterize the human auditory system. Although, nowadays the precise psychoacoustic model for the high quality audio coding is not existence, audio coding algorithms can optimize the coding efficiency and quality depending upon the psychoacoustic model. However from the viewpoint of audio coding, the final receiver is human ears. Therefore, hearing quality is significantly affected by the properties of human auditory system, especially for masking effect. Audio coding coders usually employed the irrelevant signal information which is not detectable by even a sensitive listener to reduce the compression rate. Thus, using the signal analysis incorporating into the several psychoacoustic principles including absolute hearing thresholds, critical band analysis, simultaneous masking, the spread of masking along the basilar membrane, and temporal masking is to identify irrelevant information. Combining these psychoacoustic principles with basic properties of signal quantization has also led to the theory of perceptual entropy, a quantitative estimate of the fundamental limit of transparent audio signal compression. Last, the psychoacoustic model has the several subjects including absolute hearing threshold, critical band analysis, masking effect, and perceptual entropy which will be introduced in the following sections.

2.1.1 Absolute Hearing Threshold

A minimum threshold of the pure tone which is can be detected by a listener in a noiseless environment is the absolute hearing threshold (ATH) also called threshold in quite. And it is always expressed in terms of dB SPL (sound press level, a standard metric that quantifies the intensity of an acoustical stimulus). Fletcher [3] addressed that the frequency dependence of this threshold. Furthermore, Terhardt [4] proposed a well approximated nonlinear function:

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5e^{-0.6\left(\frac{f}{1000}-3.3\right)^2} + 10^{-3}\left(\frac{f}{1000}\right)^4 \text{ (dB SPL)}, \quad (1)$$

The curve in the Figure 2 represents the threshold which is representative of a training listener with acute hearing. However, $T_q(f)$ could be deemed the maximum allowable energy level for coding distortion while applying to the audio coding. Nevertheless, it has two issues in implementation to shape the coding distortion. First, the quantization noise in the audio coding is caused by the complicated spectrum containing not only pure tone

stimuli but also other stimuli such that the noise can not be masked by the ATH. Second, the relation between the ATH function and energy is not definitely clear for audio coder. Therefore, implementation the threshold in the audio coding is necessarily conservative for estimation of the masking capability. Thus, most audio coders usually degrade the threshold in applying to audio coders. For example in the standard psychoacoustic model II makes the dB values of ATH are relative to the level that a sine wave of + or – 1/2 least significant bit has in the FFT used for threshold calculation.

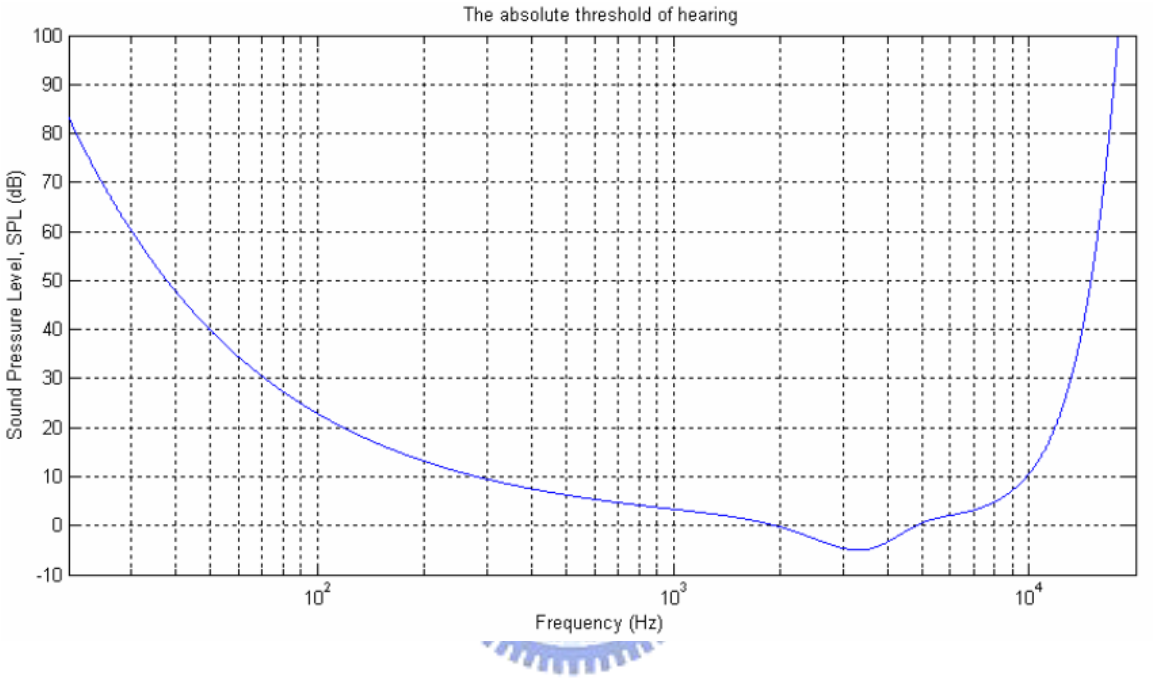

Figure 2: The curve of absolute hearing threshold (by Terhardt [4]).

2.1.2 Critical Band Analysis

The critical band is used to define that the unit of human auditory system. Nevertheless, it has been not clear and definite understanding for audio coding algorithm designers. Realizing the human ear structure as shown in Figure 3 is necessary for researching the human auditory system. Thus, emulating the human auditory system first is to know how to perform the spectral analysis in the cochlea. However, the cochlea is a highly overlapping bandpass filter in which the frequency-to-place transformation can take place along the basilar membrane. For example, when the oval window receives the excited sound like mechanical vibration, the sound follows the cochlea structure traveling along the length of the basilar membrane. The peak responses are produced at frequency-specific membrane position by the excited sound. Therefore, different frequency ranges are effectively fit for the different neural receptors according to their basilar membrane position.

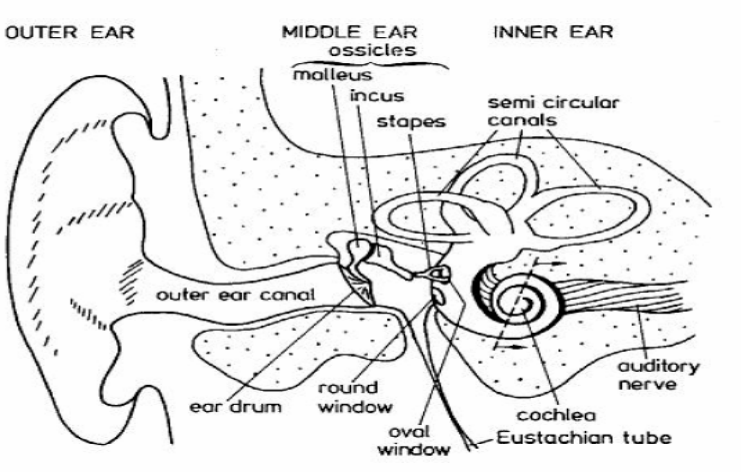

Figure 3: The structure of the human ear (by Zwicker [5]).

Consequently, using the critical band analysis models the behavior of cochlea. As illustrated in Figure 4, a unit of the critical band is one bark as defined by the formula [5]:

$$z(f) = 13\arctan(0.00076 f) + 3.5\arctan[(\frac{f}{7500})^2] (\text{Bark}) \quad (2)$$

It is used to convert from frequency in Hertz to the Bark scale. Moreover, the critical bandwidth is narrower in low frequency and wider in the high frequency because the sensibility of human auditory relies upon the frequency. Another formula is represented for the critical bandwidth [5], which is designed as:

$$BW_c(f) = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69} (Hz). \quad (3)$$

Figure 5 illustrates the curve of the critical bandwidth and describes the relation between frequency and critical band.

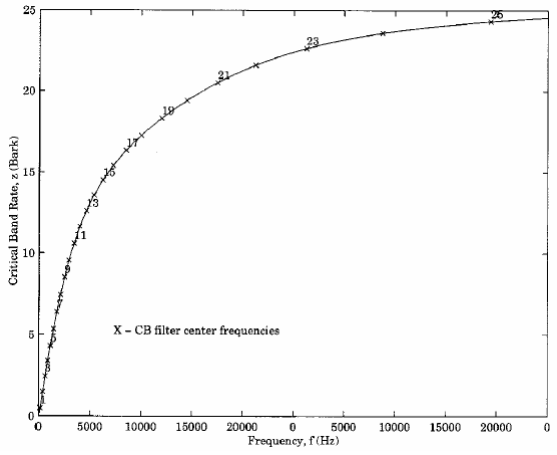

Figure 4: The critical band rate (by

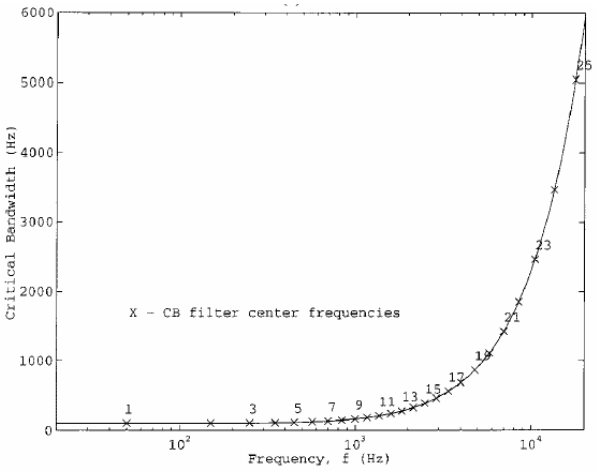

Figure 5: Critical bandwidth (by Zwicker

Zwicker [5]). [5]).

Table 1: Critical Bandwidth (by Zwicker [5]).

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

2.1.3 Masking Effect

Masking effect is that the rendered sound is inaudible because another sound is raised at the same time. Masking is an important characteristic of the human auditory system that can help audio coder designers optimize the bit allocation for an input signal in the perceptual audio coding system. Because a sound is likely to be masked by another sounds, the audio coder can allocate the prime bits to the most audible sound which may be the strong masker and allocate rest bits to the insensitive one which is possibly almost be masked. However, the sound is generally considerable complicated since the masker is possibly masked by other maskee and masker is also masked by another masker. Therefore, it actually has difficulties in exactly analyzing the relation between the masker and the maskee. Moreover, the masking effect can part into two categories from the temporal perspective: simultaneous masking which also called spectral masking and nonsimultaneous masking also called temporal masking.

1. Simultaneous masking

From the viewpoint of the time-domain, simultaneous masking is a phenomenon that simultaneous presences of the stimuli cause some of them to be not sensitive to human hearing as shown in Figure 6. For instance, only little power just can be heard by a solo piano in a quite environment, but when another instrument like bass drum presents at the same time the piano sound may be no loner heard. As far as the human auditory system is concerned, the strong masker makes a sufficient excitation on the basilar membrane at the critical band location block effectively detection of a weaker signal. In other words, the weak signal can save the bits for quantization due to this masking effect. However, it is very difficult in how to find the masker in order to calculate the masking threshold for quantization. Thus, for the proposes of simplifying the estimation of coding distortion it is usually used only two types of simultaneous masking, namely, tone-masking-noise(TMN) [6], noise-masking-tone(NMT) [7] to compose the signal masking ratio (SMR) in Figure 7 for the quantization in the perceptual coding. For

example, in the MPEG-AAC[8][9] , it defines that：NMT (b) = 6 dB for all b. NMT (b) is the value for noise masking tone (in dB) for each partition band; TMN (b) = 18 dB for all b. TMN (b) is the value for tone masking noise (in dB) for each partition band. And, in the MPEG-1 Layer III [10] it is defined as：NMT is set to 6.0 dB for all partition bands, and TMN is set to 29.0 dB. Therefore, the simultaneous masking is a strong signal either tone or noise which can mask other weaker concurrence signal.

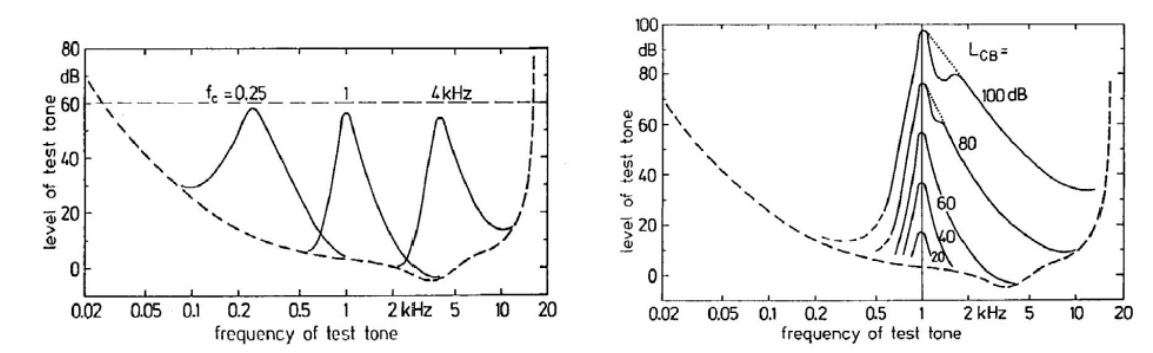

Figure 6: Simultaneous masking effect in varying frequency and energy (by Hellman [6]).

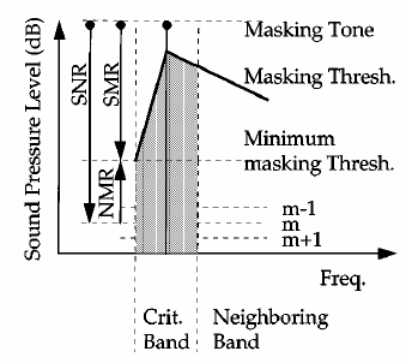

Figure 7: Illustration of masking effect (by Hellman [6]).

2. Nonsimultaneous masking

Nonsimultaneous masking is also called temporal masking shown in Figure 8, which is significantly different to simultaneous masking in occurrence of the maskee. Temporal masking happens prior to the masker or posterior to the masker. The first is named pre-masking [11] which lasts only few milliseconds about 1-2 ms and decays rapidly. The last, which is called post-masking, persists for more than 100 milliseconds after masker removal, depending upon the masker strength and duration. The violent transients of the audio signal will create the temporal masking either prior to the masker or after the masker which can lead the listener not to perceive signal beneath the masking threshold produced by the masker. State-of-the-art audio coding [12][13] algorithm have used the temporal masking. Pre-masking particularly has been utilized in conjunction with

adaptive block switch between long and short block to compensate for pre-echo distortions.

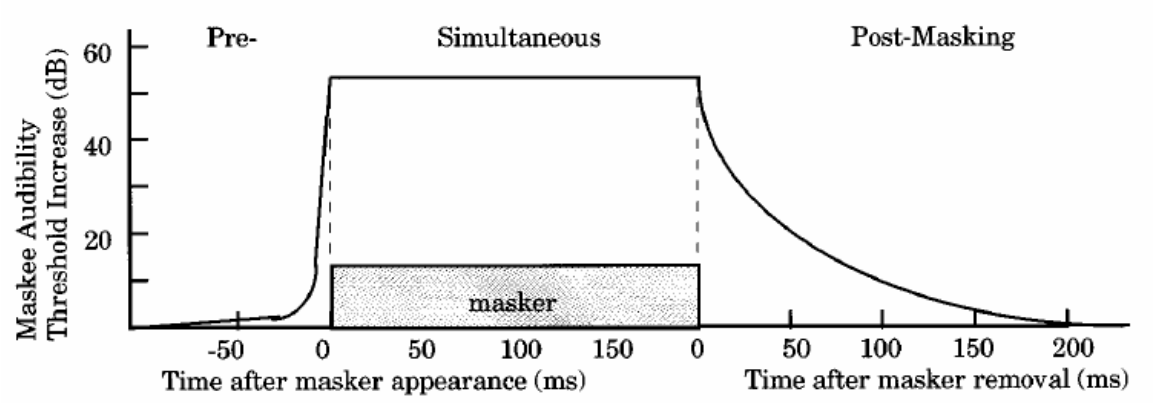

Figure 8: Illustration of the temporal masking (by Moore [11]).

3. The masking spreading effect

The audio coding algorithms use the TMN and NMT to estimate the simultaneous masking. As mentioned before, all TMN and NMT are not band-limited to within the boundaries of a single critical band. However, inter-band masking also happens. A masker centered within a critical band also masks other critical bands. This phenomenon, known as spread of the masking, is modeled in audio coding applications by a spreading function that is given by:

$$SF_{dB}(x) = 15.81 + 7.5(x + 0.474) - 17.5\sqrt{1 + (x + 0.474)^2}\, dB. \quad (4)$$

In fact, each frame in the audio coding scenario has not only tones but also noises such that it has two types of the masker. Finally, the coding algorithms separate the two different types from this frame and individually calculate the masking threshold. Furthermore, the two individual masking thresholds are combined to a global masking threshold for the quantization process in the perceptual audio coding.

2.1.4 Perceptual Entropy

Johnston, while at Bell Labs, first defined perceptual entropy. Perceptual entropy is a notion of psychoacoustic masking combined with signal quantization principles, which is a measure of perceptually relevant information contained in any audio coding. The PE estimation process is stated as follows：

The frequency-domain transformation is done with a Hanning window followed by a 2048-point Fast Fourier Transform (FFT).

$$P(x) = \mathrm{Re}(x)^2 + \mathrm{Im}(x)^2. \quad (5)$$

Perform critical band analysis with spreading.

$$B_b = \sum_{x=bl_b}^{bh_b} P(x)$$
$$E_b = \sum_{BB} B_{BB} * SF_{BBb} \quad . \tag{6}$$

Make a determination of the tonality of the signal.

$$SFM = \frac{\mu_g}{\mu_a}$$
$$\alpha_b = \min(\frac{SFM_{dB}}{-60}, 1) \quad . \tag{7}$$

Masking thresholds are obtained by applying the threshold rules for the signal and absolute hearing threshold.

$$T_b = 10^{-(\frac{O_b}{10})} * E_b \ \ and$$
$$PE = \sum_{b}^{25} \log(\frac{E_b}{T_b}) \tag{8}$$

The x means the spectral lines which are the time-domain pass by the FFT.

bh_b, bl_b are upper and lower bound of the $band_i$.

SF_{BBb} is the spreading function from band BB into b.

μ_a, μ_g are the arithmetic and geometry means.

α_b is the tonality of the band b.

T_b is the masking threshold for band b.

The signal is first windowed and transformed to the frequency domain. A masking threshold is then obtained using perceptual rules. Finally, a determination is made of the number of bits required to quantize the spectrum without injecting perceptible noise. PE represents a theoretical limit on the compressibility of a particular signal, expressed in bits per sample. PE measurements, reported in [14] and [15], suggest that a wide variety of CD-quality audio source material can be transparently compressed at approximately 2.1 bits per sample.

2.2 Psychoacoustic Model II

The psychoacoustic model II is most popularly used in perceptual audio coding defined in [16]. The model can be considered with the following steps:

Step 1 Input sample stream.

Two different window types are necessary for calculating the masking threshold in psychoacoustic model. The long window needs 2048 samples which consist of the 1024 samples at current frame and another 1024 samples at last frame and so short window does. In each frame, the coder needs shifting length 1024 for long window and length 128 for short window.

Step 2 Calculate the complex spectrum of the input signal.

First, input signal $s(i)$ from above step is windowed by a Hanning window:

$$sw(i) = s(i)(0.5 - 0.5\cos(\frac{(pi*(i+0.5))}{1024}). \tag{9}$$

Second, perform a forward Fast Fourier Transform (FFT) to $sw(i)$.

A FFT is an efficient algorithm to compute the discrete Fourier transform (DFT) and its inverse. The DFT is defined by the formula：

$$f(x) = \sum_{k=0}^{N-1} x_k e^{-\frac{2i\pi}{N}jk} \quad for \quad j = 0,...,N-1 \quad k = 0,...,N-1. \tag{10}$$

The Fast Fourier Transform usually adopted is derived from Cooley-Turkey. This is a divide and conquer algorithm that recursively breaks down a DFT of any composite size $n = n_1 n_2$ into many smaller DFTs of sizes n_1 and n_2.

Third, the result of the transform is obtained represented in polar form. $r(w)$ and $f(w)$ individually represent the magnitude and phase components of the transformed $sw(i)$.

Step 3 Estimate predicted values of the $r(w)$ and $f(w)$.

A predicted magnitude $r\_pred(w)$ and phase $f\_pred(w)$ are calculated from $r(w)$ and $f(w)$ of the preceding two frames and last frame.

$$\begin{aligned} r\_pred(w) &= 2.0r(t-1) - r(t-2), \quad and \\ f\_pred(w) &= 2.0f(t-1) - f(t-2) \end{aligned} \tag{11}$$

where t represents the current block number, t-1 indexes the previous block's data, and t-2 means the previous two block's data. This concept is using the median to predict the next value either magnitude or phase as below:

$$current(w) = \frac{next(w) + last(w)}{2} \Rightarrow next(w) = 2 \times current(w) - last(w). \tag{12}$$

Step 4 Calculate the unpredictability measure $c(w)$.

$$
\begin{aligned}
tmp\_\cos &= (r(w)\cos(f(w)) - r\_pred(w)\cos(f\_pred(w)))^2 \\
tmp\_\sin &= (r(w)\sin(f(w)) - r\_pred(w)\sin(f\_pred(w)))^2 \quad . \\
c(w) &= \frac{\sqrt{tmp\_\cos + tmp\_\sin}}{r(w) + abs(r\_pred(w))}
\end{aligned}
\tag{13}
$$

$\sqrt{tmp\_\cos e + tmp\_\sin}$ means the difference between the real spectral line and predicted spectral line and then divided by $r(w) + abs(r\_pred(w))$ in order to let the $c(w)$ range between 0 and 1, also called normalization.

This formula is used for all the short blocks with short FFT, but for long blocks the unpredictability measure is calculated from the long FFT for the first 6 lines, and for the remaining lines the minimum of the unpredictability of all short FFT's is used. If considering saving the calculation power, the unpredictability of the upper part of the spectrum can set to 0.4.

Step 5 Calculate the energy and unpredictability in the threshold calculation partition band.

The energy in each partition $e(b)$ is:

$$
e(b) = \sum_{lower\ \ index_b}^{upper\ \ index_b} r(w)^2 \ . \tag{14}
$$

And the weighted unpredictability $c(b)$ is:

$$
e(b) = \sum_{lower\ \ index_b}^{upper\ \ index_b} r(w)^2 c(w) \ . \tag{15}
$$

The upper index means the highest frequency line in the partition band, and respectively the lower index means lowest line in the partition band.

The threshold calculation partitions provide a resolution of approximately either one FFT lines or 1/3 critical band, whichever is wider. At low frequencies, a single line of the FFT will be likely to constitute a calculation partition band. However, many lines will be combined into one calculation partition band at high frequencies.

Step 6 Convolve the partitioned energy and unpredictability with the spreading function as:

$$
ecb(b) = \sum_{for\ \ each\ \ partition\ \ band} e(bb)spreading(bval(bb), bval(b)) \tag{16}
$$

$$
ct(b) = \sum_{for\ \ each\ \ partition\ \ band} c(bb)spreading(bval(bb), bval(b)) \ . \tag{17}
$$

Spreading Function:

Spreading function is calculated by the following step:

$$\begin{aligned} &Input = spreading(i, j) \\ &if \quad j \geq i \\ &\quad tmpx = 3.0(j - i) \\ &else \\ &\quad tmpx = 1.5(j - i) \end{aligned} \tag{18}$$

$$tmpz = 8 * \min((tmpx - 0.5)^2 - 2(tmpx - 0.5), 0) \tag{19}$$

$$tmpy = 15.811389 + 7.5(tmp + 0.474) - 17.5(1.0 + (tmpx + 0.474)^2)^{\frac{1}{2}} \tag{20}$$

$$\begin{aligned} &if \quad (tmpy < -100) \\ &\quad spreading(i, j) = 0 \\ &else \\ &\quad spreading(i, j) = 10^{\frac{(tmpz + tmpy)}{10}} \end{aligned} \tag{21}$$

where i is the Bark value of the signal being spread, and j is the Bark value of the band being spread into.

$bval(b)$ means the median bark of the partition band b.

Because $ct(b)$ is weighted by the signal energy, it must be renormalized to $cb(b)$ as

$$cb(b) = \frac{ct(b)}{ecb(b)}. \tag{22}$$

Similarly, due to the non-normalized nature of the spreading function, ecb_b should be renormalized and then normalized energy en_b is obtained:

$$en(b) = ecb(b) \times rnorm(b).$$

The normalization coefficient $rnorm(b)$ is:

$$\begin{aligned} &tmp(b) = \sum_{for \quad each \quad partition \quad band} spreading(bval(bb), bval(b)) \\ &rnorm(b) = \frac{1}{tmp(b)} \end{aligned} \quad . \tag{23}$$

Step 7 Convert $cb(b)$ *to* $tb(b)$ *the tonality index as*:

$$tb(b) = -0.299 - 0.43 \log_e(cb(b)).$$

Each $tb(b)$ is limited to the range $0 \le tb(b) \le 1$

Step 8 Calculate the required SNR in each partition band.

$NMT(b)$ = 6 dB for all b. $NMT(b)$ is the value for noise masking tone (in dB) for the partition band. $TMN(b)$ = 18 dB for all b. $TMN(b)$ is the value for tone masking noise (in dB) for the partition band.

The required signal to noise ratio SNR(b) is:

$$SNR(b) = tb(b) \times TMN(b) + (1 - tb(b)) \times NMT(b) . \quad (24)$$

Step 9 Calculate the power ratio.

The power ratio $bc(b)$ is:

$$bc(b) = 10^{\frac{-SNR}{10}} . \quad (25)$$

Step 10 Calculation of actual energy threshold $nb(b)$

$$nb(b) = en(b) \times bc(b) . \quad (26)$$

Step 11 Pre-echo control and threshold in quiet.

To avoid pre-echoes the pre-echo control is calculated for short and long FFT, the threshold in quite is also considered here:

$nb\_l(b)$ is the threshold of partition b for the last block, $qsthr(b)$ is the threshold in quite. The dB value must be converted into the energy domain after considering the FFT normalization actually used.

$$nb(b) = \max(qsthr(b), \min(nb(b), nb\_l(b) \times rpelev)) . \quad (27)$$

$rpelev$ is set to 0 for short blocks and 2 for long blocks.

Step 12 The PE is calculated for each block type from the ratio $e(b)/nb(b)$ *, where* $nb(b)$ *is the threshold and* $e(b)$ *is the energy for each threshold partition.*

$$PE = \sum_{for\ \ each\ \ partition\ \ band} -\log_{10}\left(\frac{nb(b)}{e(b)+1}\right) \times Bandwidth(b) . \quad (28)$$

$Bandwidth(b)$ represents the width of the partition band.

Step 13 The decision, whether long or short block type is used for encoding is made according to this pseudo code.

```
if PE for long block is greater than switch_pe then
    coding_block_type = short_block_type
else
    coding_block_type = long_block_type
end if
if (coding_block_type == short_block_type) and (last_coding_block_type == long_type) then
    last_coding_block_type = start_type
else
    last_coding_block_type = short_type
end if
```

The last four lines are necessary since there is no combined stop/start block type in AAC. Switch_pe is an implementation depended constant.

Step 14 Calculate the signal-to-masking ratios $SMR(n)$.

The index *swb* of the coder partition called scalefactor band which is the quantization unit. The offset of MDCT line for the scalefactor band is swb_offser_long/short_window Define the following variable:

$n = \text{swb}$

$w\_low(n) = \text{swb\_offset\_long/short\_window}(n)$

$w\_high(n) = \text{swb\_offset\_long/short\_window}(n+1) - 1$

The FFT energy in the scalefactor band epart(n) is:

$$epart(n) = \sum_{for\ \ each\ \ scalefactor\ \ band} r(w)^2 \tag{29}$$

the threshold for one line of the spectrum in the partition band is calculated according to:

$$thr(w\_low(b),...,w\_high(b)) = \frac{nb(b)}{w(high(b) - low(b) + 1)}. \tag{30}$$

the noise level in the scalefactor band on FFT level npart(n) is calculated:

$$\begin{aligned} npart(n) = \min(&thr(w\_low(n)),.....,\\ &thr(w\_high(n)))*(w\_high(n) - w\_low(n) + 1) \end{aligned} \tag{31}$$

$$SMR(n) = \frac{epart(n)}{npart(n)}. \tag{32}$$

The output of the psychoacoustic model is a set of the Signal-to-Masking Ratios, delayed time domain data used by filterbank, and an estimation of how many bits should be used

for encoding in addition to the average available bits. Filterbank should use the delayed data because if the switch decision algorithm detects an attack, short blocks have to be used for the actual frame, the long block before short block has to be patched to start block type. The psychoacoustic model II flow chart is shown in Figure 9:

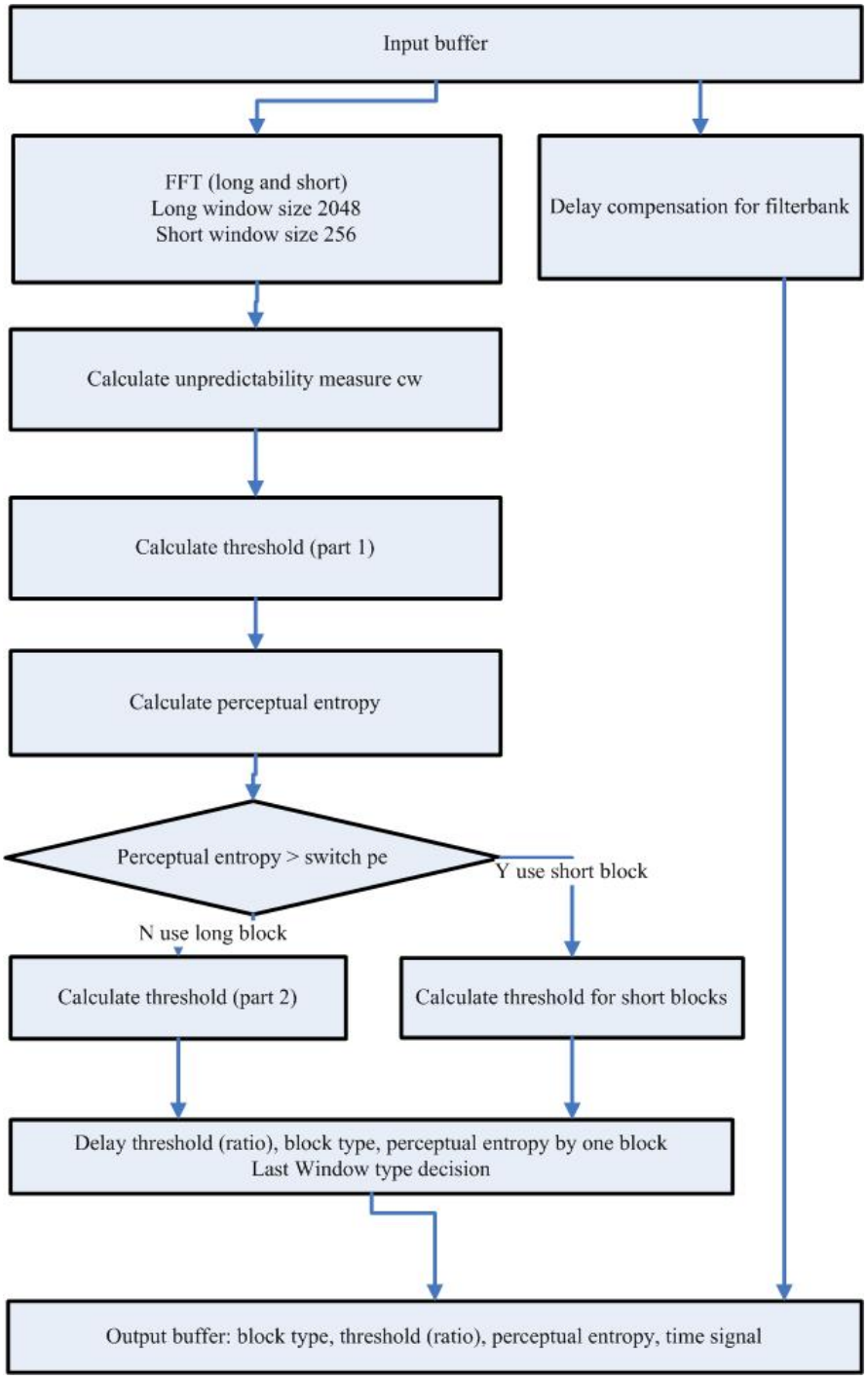

Figure 9: The flow chart of the psychoacoustic model II.

Chapter 3 Filterbank

3.1 Filterbank Concept

The time-frequency analysis block is essential to the all audio coders. It extracts from time-domain input signal some information which is responsible for the encoding according to perceptual distortion metrics. The filterbank is the component most usually used for this mapping. And it is a parallel bank of bandpass filters which covers the entire spectrum. The signal spectrum is separated by the filterbank into frequency subbands. And then the filterbank generates a time-indexed series of coefficients representing the local frequency signal power in each band. When used in combination with a psychoacoustic model, the filterbank is an important element to provide explicit information about the distribution of signal in order to identify the perceptual irrelevancies and masking threshold over the time-frequency plane. Simultaneously, the filterbank generates the time-frequency parameters that provide a signal mapping. However, the mapping is appropriately manipulated to shape the coding distortion in order to match the time-frequency distribution of masking power. That is to say, the filterbank eases psychoacoustic analysis as well as perceptual noise shaping. And then the filterbank also aids in the diminution of the redundancies by separating the signal into its constituent frequency components. A suitable filterbank is crucial to the success of a perceptual audio codecs. Efficient coding performance relies greatly upon sufficiently matching the properties of the analysis filterbank to the characteristics of the input signal [17]. When selecting a filterbank structure [18], the audio coding algorithm designers meet an important and complicated tradeoff between time and frequency resolution. Failure to choose an appropriate filterbank can cause perceptible artifacts like pre-echoes. And the failure also can result in impractically low coding gain and attendant high bit rates. There is no single optimal resolution tradeoff for all signals. In the strong harmonic signal like pitch pipe, the most appropriate filterbank must have fine frequency resolution and coarse time resolution because of the localized frequency masking threshold. On the contrary, the fast attacks signal like castanets creates highly time-localized masking thresholds such that the filterbank must have sufficient time resolution.

Actually, there is highly non-stationary and contains significant tonal and atonal energy in most audio source material, as well as both steady-state and transient intervals. Signal models [19] are usually deposed to last constant for long periods and then change abruptly. Therefore, in accordance with the time-frequency signal composition the ideal coder should make adaptive decisions. Moreover, the ideal analysis filterbank would have

content-varying resolutions in both the time and frequency domains. Many audio coding algorithm designers have been impelled by this fact to experiment with switching decisions occurring on the criterion of the changing signal properties, with switched hybrid filterbank structures. Filterbanks competes with the analysis characteristics of the human auditory system, and the most important one of these properties is non-uniform "critical bandwidth" subbands. Nowadays, filterbanks have proven highly effective in the coding of highly transient signals such as the castanets or glockenspiel. On the other hand, the "critical band" filterbanks have been not proper because of their reduced coding gain relative to filterbanks with a large number of subbands for dense harmonically structured signals such as the harpsichord or pitch pipe. Thus, signals containing very little irrelevancy such as the harpsichord particularly need good channel separation and stopband attenuation. Furthermore, for purposes of maintaining high quality at low bit rates for these signals, maximum redundancy removal is considerably necessary. Time-varying filter banks that have blocking artifacts can result in audible distortion of the reconstruction. Therefore, there are three filterbank types, Pseudo-QMF Filterbank, Perfect Reconstruction (PR) Cosine Modulated Filterbank, and Pseudo QMF in conjunction with PR Cosine Modulated Filterbank. The PQMF bank has played an important role in the evolution of modern audio codecs. The ISO IS11172-3 and IS13818-3 algorithms (MPEG-1 [20] and MPEG-2 BC/LSF [21]) employ a 32-channel PQMF bank for spectral decomposition in layers I–II. The PQMF in conjunction with PR cosine modulated filterbank, which is also called hybrid filterbank, is used in the MPEG-1 Layer III (MP3). The MPEG-2 AAC and MPEG-4 T/F filterbank use the PR cosine modulated filterbank. Princen and Bradley [22] first proved the PR in time-domain to develop the time-domain aliasing cancellation (TDAC) filter bank. Later, the modulated lapped transform (MLT), which restricts attention to a particular prototype filter and formulates the filter bank as a lapped orthogonal block transform, is developed by Malvar[23]. Lately, the modified discrete cosine transform (MDCT) has derived from the lapped block transform interpretation of this special-case filter bank in the audio coding literature.

3.2 Filterbank in AAC

MPEG-2 AAC uses the MDCT filterbank as shown in Figure 10 to transform the input signal from the time domain to frequency domain. The MDCT is a linear orthogonal lapped transform, derived from the foregoing TDAC [22][24]. The concept is using the overlap-add (OA) procedure that a single block after the IMDCT does not correspond to the original block that the MDCT is performed. However, the subsequent

blocks of inverse transformed data is added, the errors introduced by the transform would cancel out. The direct MDCT and inverse MDCT are defined as:

$$\alpha_r = \sum_{k=0}^{2N-1} \tilde{a}_k \cos\{\pi \frac{[k+(N+1)/2](r+1/2)}{N}\}, \quad \text{r} = 0,\ldots,\text{N}-1 \tag{33}$$

$$\hat{a}_k = \frac{2}{N} \sum_{r=0}^{2N-1} \alpha_r \cos\{\pi \frac{[k+(N+1)/2](r+1/2)}{N}\}, \quad k = 0,...,2N-1 \tag{34}$$

and

$$\tilde{a}_k = a_k \times h_k, \tag{35}$$

where the $a[n]$ is a input time domain signal of 2N samples and the h_n is a window function satisfying the constraints of perfect reconstruction as:

$$h_n = h_{2N-1-n} \tag{36}$$

and

$$h_n^{\ 2} + h_{n+N}^{\ 2} = 1. \tag{37}$$

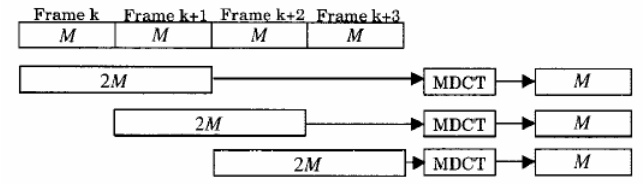

Figure 10: Illustration of the forward MDCT filterbank (by Princen [22]).

MDCT filterbank contain four window types for different demands. Long window is designed for the high frequency resolution, and oppositely short window is for the high time resolution. Start window and stop is designed for the transition between the long and short windows in order to the PR property. Finally according to the result of the psychoacoustic model, MDCT filterbank can obtain a window type to perform MDCT. In the MPEG AAC, the sine function is most popularly used for the window function as：

$$h_n = \sin(\pi \frac{k+1/2}{2N}) \quad for\ \mathrm{k} = 0,...,2\mathrm{N}-1 . \tag{38}$$

Another choice is the Kaiser-Bessel derived (KBD) window [16][21] which achieves considerably better stopband attenuation than sine window. KBD window is defined as

$$W_{KBD\_LEFT,N}(n) = \sqrt{\frac{\sum_{p=0}^{n}[w'(n,\alpha)]}{\sum_{p=0}^{N/2}[w'(p,\alpha)]}}, \quad for\ 0 \le n < \frac{N}{2} \tag{39}$$

$$W_{KBD\_RIGHT,N}(n) = \sqrt{\frac{\sum_{p=0}^{n-n}[w'(n,\alpha)]}{\sum_{p=0}^{N/2}[w'(p,\alpha)]}}, \quad for\ \frac{N}{2} \le n < N \tag{40}$$

and

$$W'(n) = \frac{I_0[\pi\alpha\sqrt{1.0-(\frac{n-N/4}{N/4})^2}]}{I_0[\pi\alpha]} . \quad for\ 0 \le n \le \frac{N}{2} \tag{41}$$

where

$$I_0[x] = \sum_{k=0}^{\infty}\left[\frac{\left(\frac{x}{2}\right)^k}{k!}\right]^2 \tag{42}$$

α = kernel window alpha factor, α = 4 for N=2048 and α = 6 for N=256.

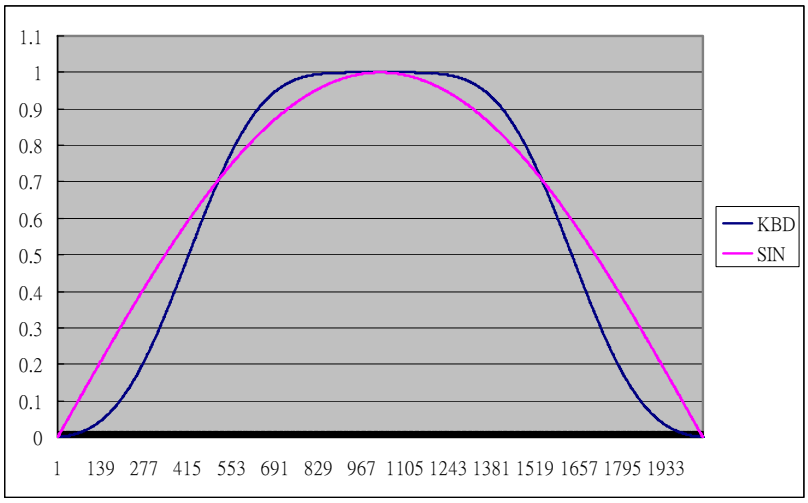

Figure 11: Examples of SIN and KBD windows.

Figure 11 shows the different curves of SIN and KBD windows in a long block type.

3.3 Filterbank in MP3

In MP3 as illustrated in Figure 12, the encoder system uses the hybrid filterbank composed of the polyphase filterbank and MDCT filterbank

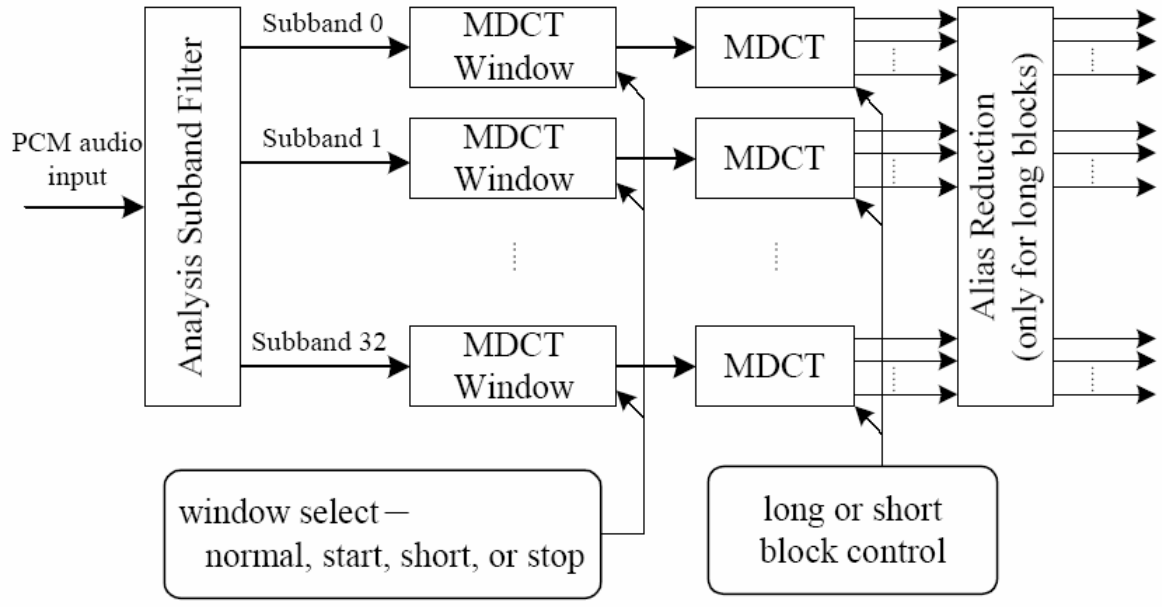

Figure 12: Illustration of the hybrid filterbank.

The polyphase filterbank is used to perform the analysis subband filter which will

transform the input signal to the 32 equally spaced subbands fs/32, where fs is the sampling frequency. The polyphase filterbank can be illustrated though the flow chart illustrated in Figure 13. Shift the 32 new input samples into the buffer $x[i]$ with length 512.

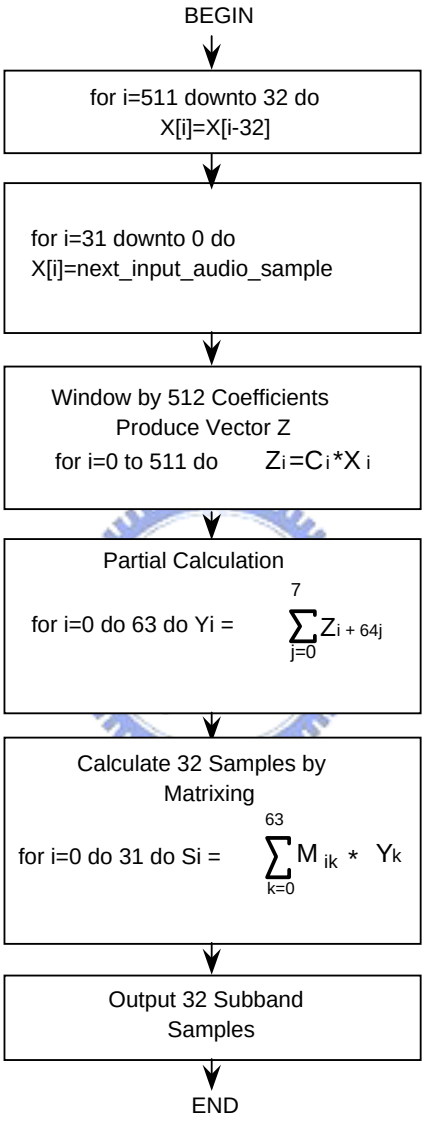

Figure 13: Analysis subband filter diagram.

The total process can be combined into the following formula:

$$s_i[i] = \sum_{k=0}^{63} \sum_{i=0}^{7} M[i][k] \times \left(C[k+64i] \times x[k+64i] \right) \tag{43}$$

and

$$M[i][k] = \cos[\frac{(2i+1)\times(k-16)\times\pi}{64}]. \tag{44}$$

After the polyphase filterbank transforming the PCM data into 32 equally spaced subbands, 18 consecutive output values of one granule and 18 output values of the granule before are assembled to one block of 36 samples which will pass through the MDCT filterbank in order to promote the frequency resolution. As MPEG-AAC, the MDCT filterbank has same properties except supporting the KBD window. Due to properties of the polyphase filterbank, the each neighbor subband has obvious overlap area which will affect the two subbands. For purpose of the reducing this aliasing, the spectral lines in the overlap area will need some modifications as below Figure 14:

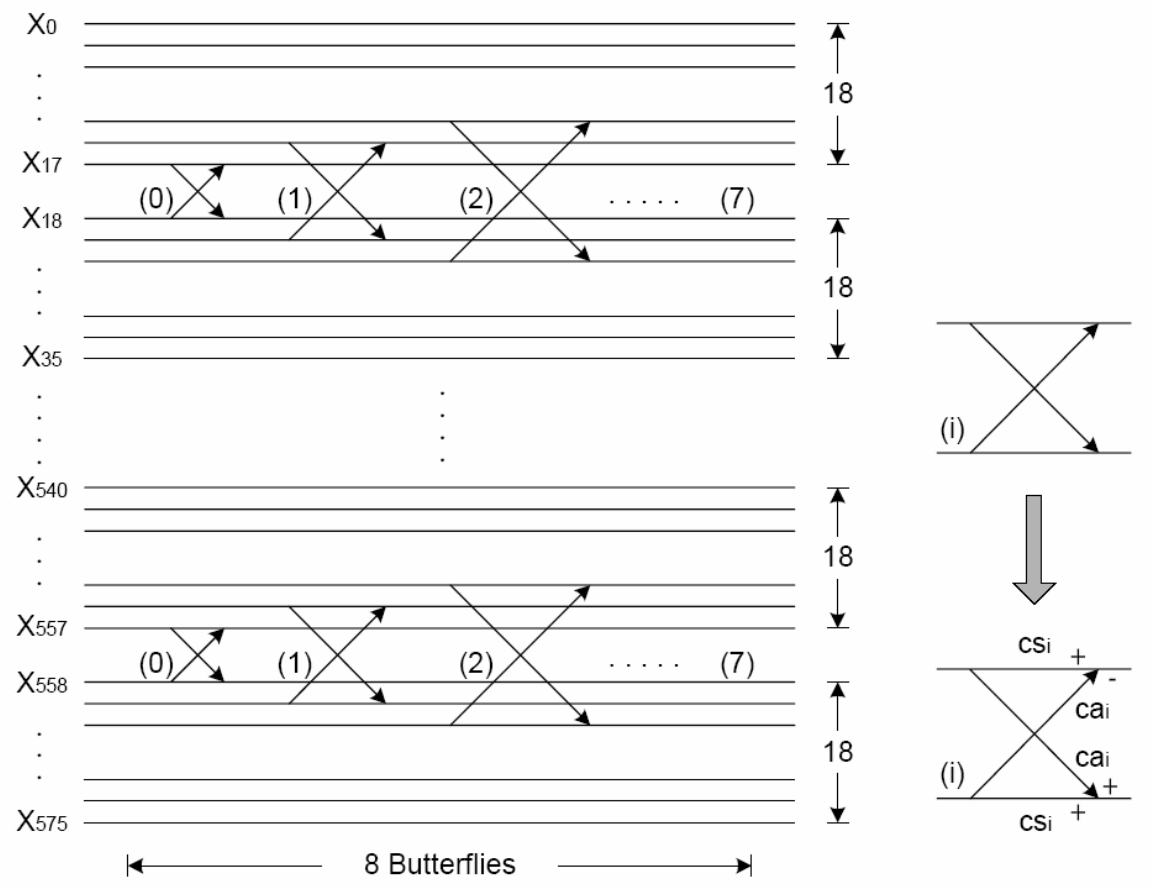

Figure 14: Aliasing butterfly.

The ca_i and cs_i are also defined in MP3 standard [10].

Chapter 4 Efficient Psychoacoustic Model

4.1 Efficiency psychoacoustic model based on the filterbank

Psychoacoustic model II uses the FFT to obtain the spectrum and estimate the masking threshold. However, the MDCT transform in AAC is used to get another spectrum for bit allocation and quantization. The two spectrum analysis for FFT and MDCT in AAC cause computation redundancy and require energy calibration of masking threshold used in bit allocation. Moreover, the inter-frame unpredictability to compute the tonality of each band in psychoacoustic model II requires high computing effort. The efficient psychoacoustic model proposed in this thesis directly uses the coefficients of the MDCT to get the spectrum and hence leads to the merits in complexity. Furthermore, the complicated unpredictability method in conventional psychoacoustic model is replaced by the flatness method to reduce both complexity and memory.

4.1.1 MDCT Psychoacoustic Model

In MPEG-4 AAC illustrated in Figure 15, the psychoacoustic model can obtain a copy of time signal, and then perform FFT in order to get the spectral information to calculate masking thresholds. Last, the masking thresholds will pass to the other encoding components like M/S coding, bit reservoir, and bit allocation.

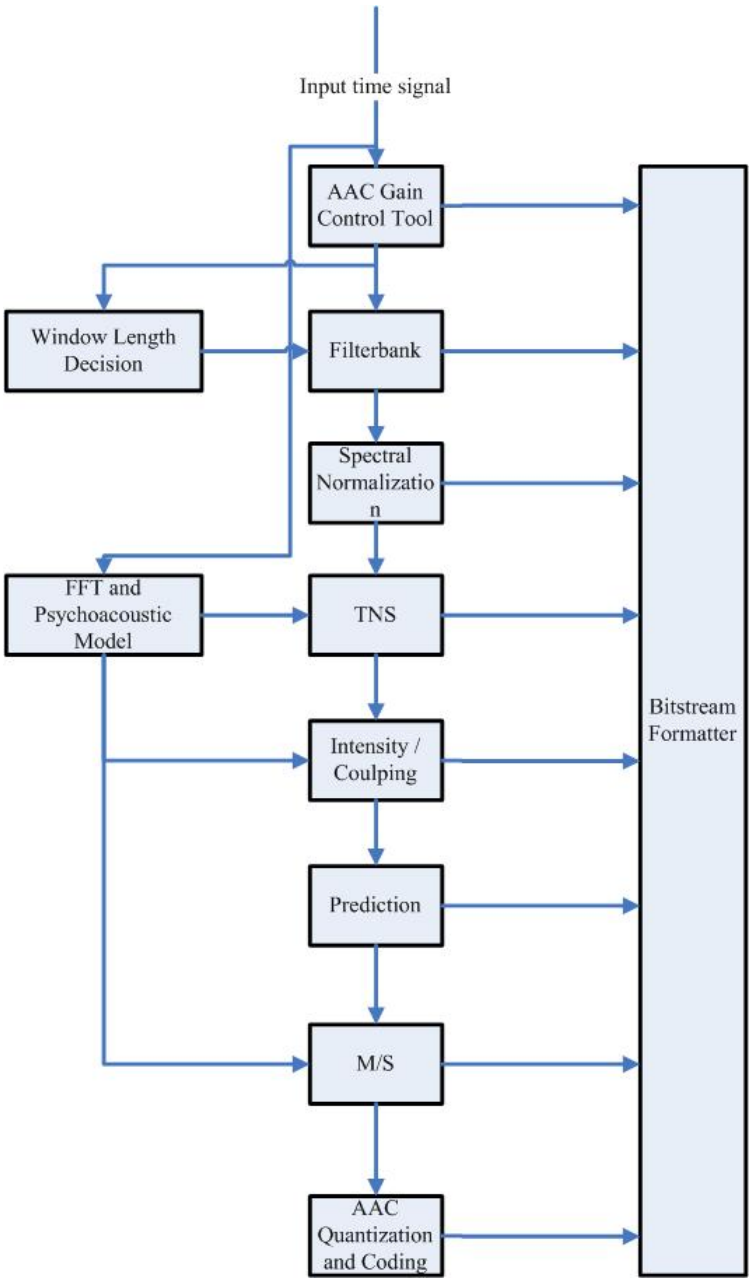

Figure 15: MPEG-4 AAC diagram.

However, this process will cost considerable computation. Thus, the psychoacoustic model based on the filterbank is addressed in Figure 16.

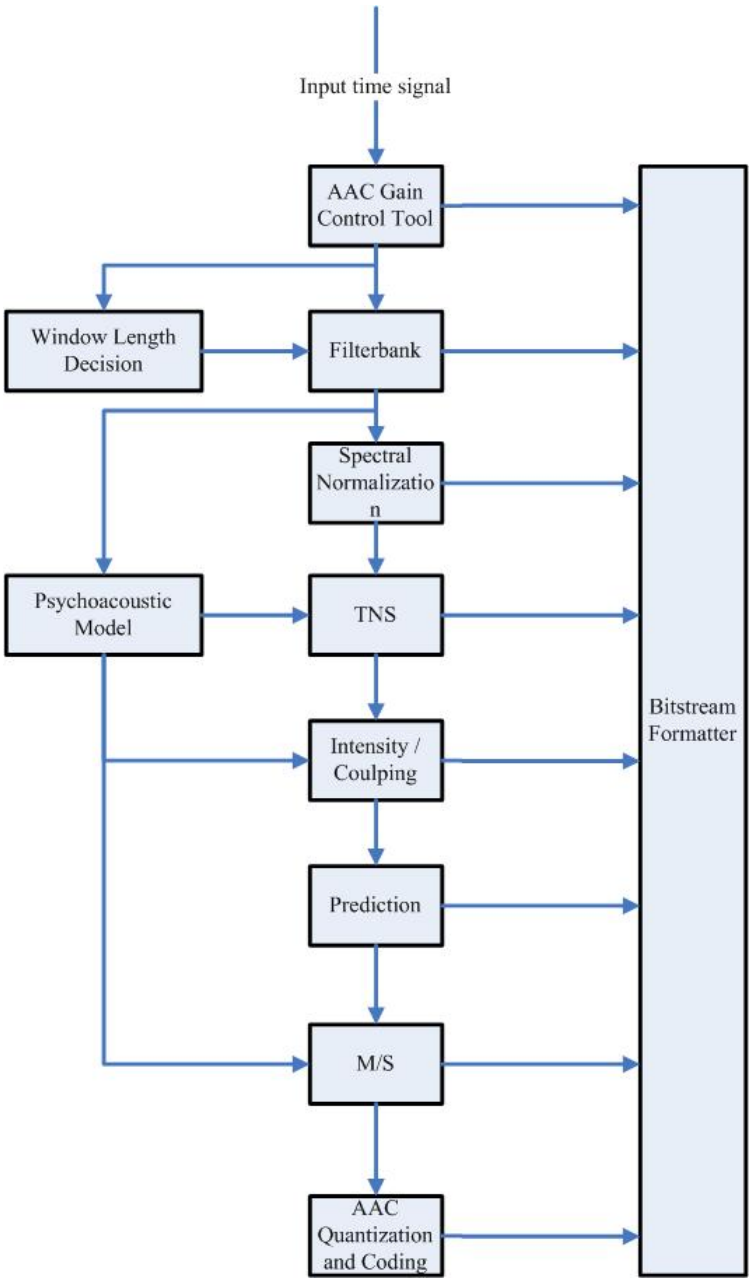

Figure 16: Efficient Psychoacoustic Model Diagram.

Figure 16 shows that the psychoacoustic model can use the spectrum from MDCT instead of the FFT. Nevertheless, using the MDCT spectrum instead of the FFT spectrum must confirm that the two output values have the same meaning as shown in Figure 17.

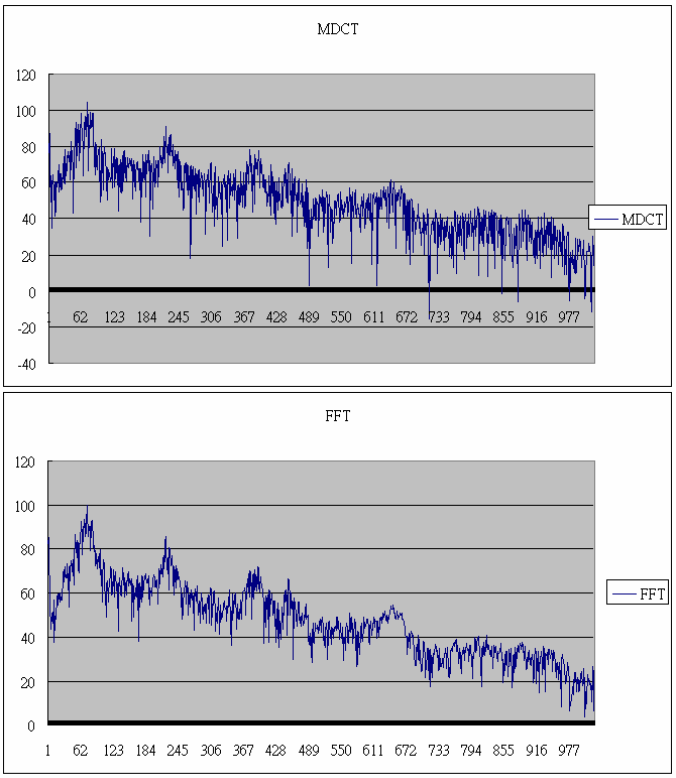

Figure 17: Illustration of two transform result where horizontal axis means the 1024 spectral lines and vertical axis means the magnitude in dB domain.

In [25], a relationship between MDCT and DFT via shifted discrete Fourier transform (SDFT) is established. The SDFT is a generalization of the DFT allowing a possible arbitrary shift in position of the samples in the time and frequency domains with respect to the signal and its spectrum coordinate system, which is defined as:

$$SDFT_{u,v} = \alpha_r^{u,v} = \sum_{k=0}^{2N-1} a_k \exp[i2\pi \frac{(k+u)(r+v)}{2N}], \tag{45}$$

$$ISDFT_{u,v} = a_r^{u,v} = \frac{1}{2N} \sum_{k=0}^{2N-1} \alpha_k{}^{u,v} \exp[-i2\pi \frac{(k+u)(r+v)}{2N}]. \tag{46}$$

where u, v represent arbitrary time- and frequency- domain shifts. And it provides a possible fast implementation of MDCT employing a fast Fourier transform routine. And it has proven that the MDCT is equivalent to the SDFT of a modified input signal as:

$$\hat{a}_k = \begin{cases} \tilde{a}_k - \tilde{a}_{N-k-1} & \text{k} = 0,...,\text{N}-1 \\ \tilde{a}_k + \tilde{a}_{3N-1-k} & \text{k} = \text{N},...,\text{N}-1 \end{cases}, \tag{47}$$

$\hat{a}_k$ areIMDCTcoefficients of α_r, and $\tilde{a}_k = h_k a_k$, a_k *is* original time signal. Thus, the SDFT will be

$$\alpha_r = \frac{1}{2}\sum_{k=0}^{2N-1} \hat{a}_k \exp\left\{ i\pi \frac{[k+(N+1)/2](r+1/2)}{N} \right\} \tag{48}$$

Therefore, the MDCT coefficients α_r can be represented as:

$$\alpha_r = real\{SDFT_{(N+1)/2,1/2}(\tilde{a}_k)\}. \tag{49}$$

And then, the $SDFT_{(N+1)/2,1/2}$ can be expressed by means of the conventional DFT as:

$$\begin{aligned} &\sum_{k=0}^{2N-1} \hat{a}_k \exp\{i2\pi \frac{[k+(N+1)/2](r+1/2)}{2N}\} \\ &= \{\sum_{k=0}^{2N-1} [\hat{a}_k \exp(i2\pi \frac{k}{4N})] \exp(i2\pi \frac{kr}{2N})\} \times \exp[i2\pi \frac{(N+1)R}{N}] \exp(i\pi \frac{N+1}{4N}) \end{aligned} \tag{50}$$

Last, $SDFT_{(N+1)/2,1/2}$ is the conventional DFT of this signal shifted in the time domain by $(N+1)/2$ of the sampling interval and evaluated with the shift of one-half the frequency-sampling interval. Although the MDCT filterbank is considerably similar to FFT, the output of MDCT filterbank still has a problem that is the lost of the imaginary information. In the foregoing psychoacoustic model II, the unpredictability measure needs the two information including the magnitude and phase. Therefore, the spectral flatness measure is appropriately applied to the MDCT psychoacoustic model due to the lost imaginary information.

4.1.2 SFM Tonality Decision

The unpredictability in conventional psychoacoustic model II needs great space to store old information in order to estimate the tonality. Replacing by the spectral flatness measure can save the space for storing and computational time of per spectral line calculation in unpredictability because the psychoacoustic model uses the unit in terms of the partition band rather than spectral line. The SFM is defined as:

$$flatness_b = \frac{GM_b}{AM_b} \text{ , } \mathrm{GM_b} = \prod_{i=0}^{N-1} x_i^{\frac{1}{N}} \text{ , } \mathrm{AM_b} = \frac{1}{N}\sum_{i=0}^{N-1} x_i \tag{51}$$

and the constrain $0 \le flatness_b \le 1$.

If the $flatness_b$ equal to 1, this means that all the equivalent x_i representing noise as shown in (a). Oppositely, if the $flatness_b$ approaches to 0, x_i varies much more as (b), which is so-called tone.

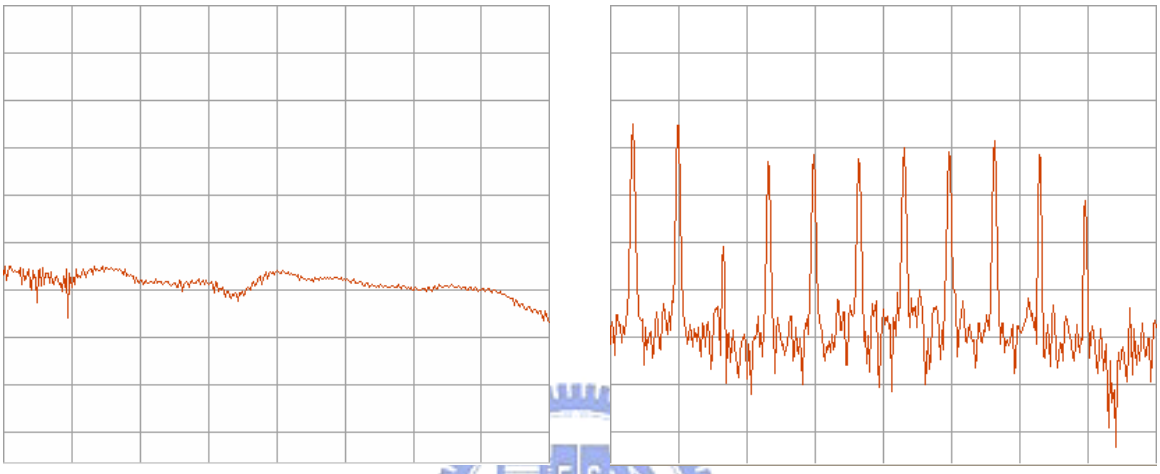

Figure 18: (a) Illustration of the noise signal. (b) Illustration of the harmonic signal.

Thus, using SFM instead of the unpredictability can speed up the efficiency and save storing space.

In the original SFM in [26], the tonality is that flatness is divided by a constant either bigger or smaller. And, the tonality will be one if the flatness is bigger than the constant. This thesis uses thresholds to separate the flatness into different intervals. And then, different intervals will be divided by the different constants for purposes of enhancement of the characteristic of the flatness. This is because if the flatness approaches to zero more, it tends to represent noise. On the contrary, if the flatness approaches to one more, it tends to represent tone.

4.1.3 Calculate SMR

The offset stated in Chapter 2 has divided into two parts：TMN, NMT which are used to form the signal-to-masking ratio (SMR) in the current coders. However in the [8] [9], the two offsets still keep constant for all bands. But the high frequency is insensitive for human auditory system such that the masking effect is stronger than the low frequency. Moreover, the bandwidth is narrower in the low frequency such that the wider tone will be ignored. Therefore, a non-fixed masking offset is addressed, which depends upon the bandwidth to modify the offset as shown in Figure 19.

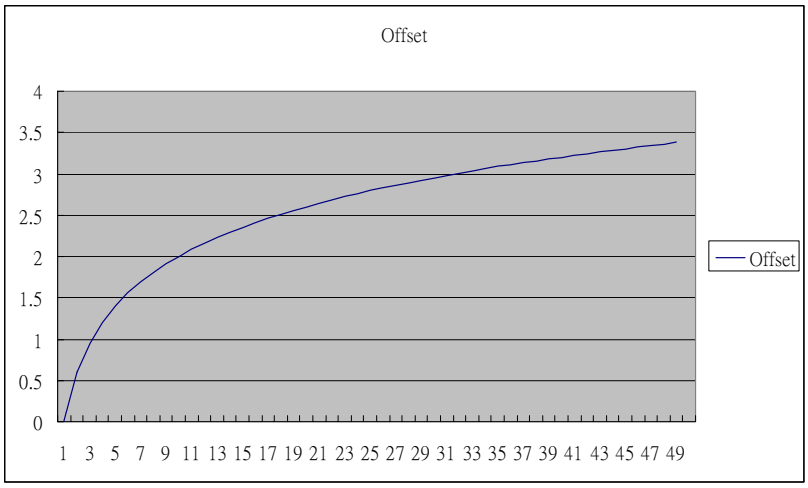

Figure 19: Adaptive offset control. Horizontal axis means band number and vertical axis represents the modification offset in dB.

4.2 Detection of Tonal Signal

Time domain attacks can be detected in the conventional psychoacoustic model. However, the frequency domain attacks are also essential for the audio coder. First, the tonal signal can make window switch avoid error switching due to poor frequency resolution. Second, in the strong harmonic signals the tonal signal also can make bit allocation use less scale factor bits in order to allocate more bits for Huffman coding.

4.2.1 Detection of Tonal Attack Band

In Subsec4.1.2, the tonality represents the degree of tone in this band. If detecting the tonality over a threshold, the band is deemed as a tonal attack band.

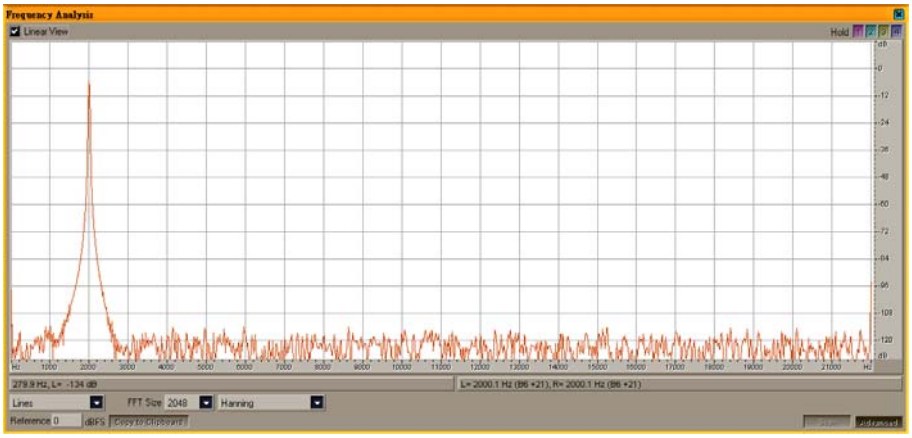

Figure 20: Illustration of peak signal at 1k.

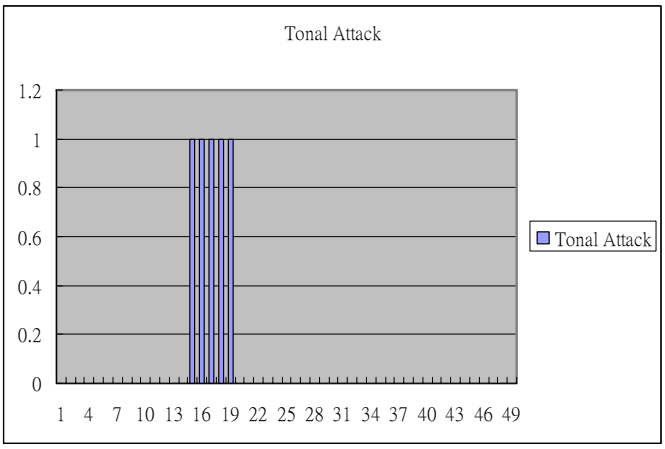

Figure 21: Illustration of the result of the tonal attack detection. x-axis means the quantization band and y-axis means the tonal attack band flag.

In the tonal attack band detection, a peak signal at 1k as shown in Figure 20 can be detected as tonal band in the corresponding band as shown in Figure 21.

4.2.2 Detection of Tone-Rich Signal

Tonal attack is local detection for peak signal in a band; however, the tone-rich signal is a global view for the spectrum. If the number of the tonal attack bands in a frame achieves a threshold, the signal is called tone-rich signal.

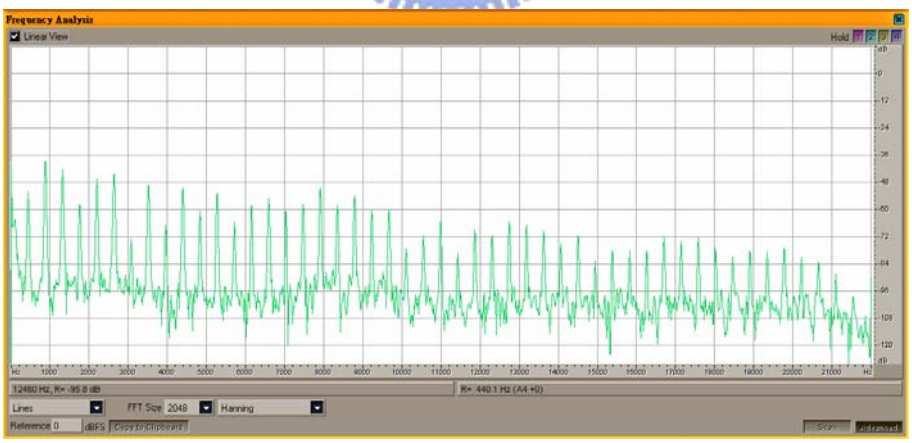

Figure 22: Illustration of the tone-rich signal.

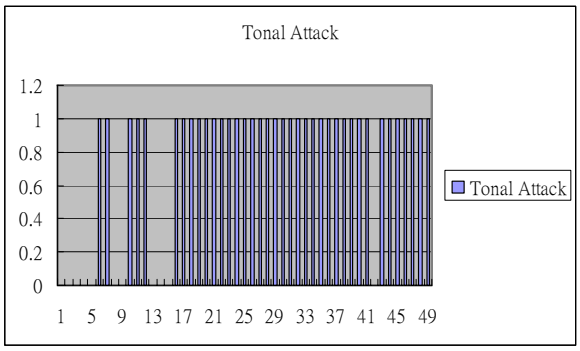

Figure 23: Example of tone-rich signal. x-axis means the quantization band and y-axis means the tonal attack band flag.

Figure 22 shows the frame has strong harmonic signals which can result in many tonal attack bands. Therefore as concerned as the whole spectrum, it has considerable tonal attack bands representing tone-rich signal as shown in Figure 23.

4.3 Experiments

The experiments can be separated into two parts according to the foregoing sections. First, the efficiency is to test the computation time of psychoacoustic model and the encoding time in different psychoacoustic models. Second, the quality test is always the critical issue in audio research on the experiments to prove the quality improvement. This thesis conducts the experiments on three hundred critical tracks and checks the possible risk through the Objective Difference Grade (ODG) developed by Recommendation ITU-R BS.1387 [27] in addition to the subjective measure. The result of ODG ranges from 0 to -4, where the value 0 corresponds to an imperceptible degradation and -4 to a degradation judged as very annoying. The result value is negative, because the quality of the Signal Under Test (SUT) is assumed to be worse than Reference Signal (RS). Also, the new efficient model has been extensively tested on the various coding combination like M/S coding, TNS coding, and bit rates. In the following test results, we use P4 representing the proposed efficient psychoacoustic model, and P1 representing the conventional psychoacoustic model II in Chapter 2.

First, we use a general performance testing tool Intel vTune 7.0 to test the psychoacoustic computational time.

Table 2: The psychoacoustic computational time in NCTU-AAC.

	1	2	3	4	5	Average	Speedup (%)
P1	30.240	29.660	29.750	29.960	27.750	29.472	***72.58***

P4	8.570	8.940	8.000	7.310	7.590	8.082	

The Table 2 is running 5 times each different psychoacoustic model incorporating with other encoding components. Obviously, the proposed model can speed up the coding efficiency 72% more than P1. In conclusion, the proposed model can dramatically improve the coding efficiency.

The encoding time is shown in Table 3 testing in NCTU-AAC.

Table 3: Encoding time for NCTU-AAC.

NCTU-AAC		Encoding time (s)		Speedup for P4
FileName	Length	P1	P4	Percentage (%)
es01	02:51	26	19	26.92
es02	02:17	19	14	26.32
es03	04:03	36	27	25.00
sc01	02:55	22	18	18.18
sc02	03:23	28	23	17.86
sc03	03:04	27	23	14.81
si01	04:47	39	36	7.69
si02	03:05	30	26	13.33
si03	05:34	49	45	8.16
sm01	04:27	38	35	7.89
sm02	02:01	18	16	11.11
sm03	04:11	38	34	10.53
Average		***30.83***	***26.33***	***14.59***

This proposed model can speed up the total encoding time by 14.59% compared with that based on P1 model. Moreover, Table 4 summarizes the encoding time using the different model incorporating M/S coding, window switching, TNS coding, and Bit Reservoir.

Table 4: The encoding time of encoder incorporating M/S coding, window switching, TNS coding, and bit reservoir.

NCTU-AAC		Encoding time (s)		Speedup for P4
FileName	Length	P1	P4	Percentage (%)

es01	02:51	23	17	26.09
es02	02:17	14	10	28.57
es03	04:03	27	19	29.63
sc01	02:55	23	19	17.39
sc02	03:23	29	24	17.24
sc03	03:04	28	25	10.71
si01	04:47	42	38	9.52
si02	03:05	29	25	13.79
si03	05:34	54	50	7.41
sm01	04:27	42	37	11.90
sm02	02:01	18	16	11.11
sm03	04:11	42	37	11.90
Average		***30.92***	***26.42***	***14.56***

The proposed psychoacoustic model also provides a complexity gains by 14.56% for P1. Figure 24 shows the NCTU-AAC in P4 provides a complexity gains by 20% for QuickTime 6.3 [28] and 18.37% for Nero 6 [29].

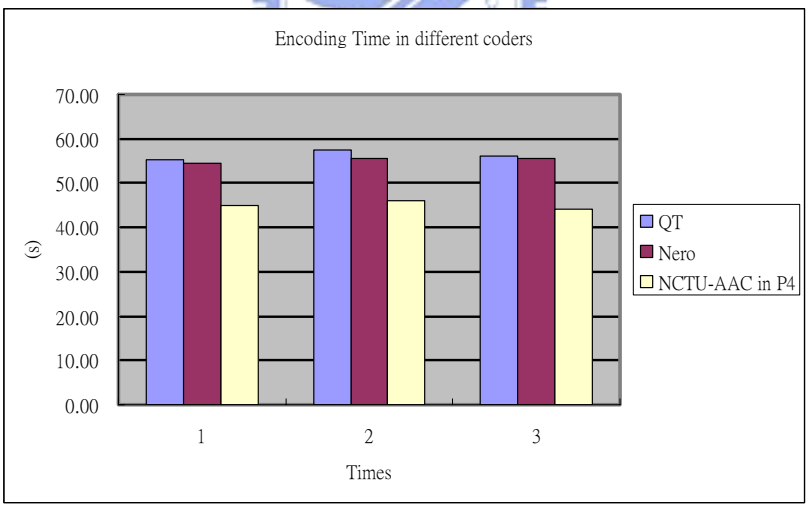

Figure 24: Illustration of the encoding time in different coders.

Second, testing tracks are most generally used to test the audio coding quality is the MPEG44100 set bitstream is the 44100 Hz version of MPEG set bitstream.

Table 5: MPEG12 44100 Test songs.

Track	Time	Signal description		
1	10	es01	vocal (Suzan Vega)	Speech signal

2	8	es02	German speech	
3	7	es03	English speech	
4	10	sc01	Trumpet solo and orchestra	
5	12	sc02	Orchestral piece	
6	11	sc03	Contemporary pop music	Complex sound mixtures
7	7	si01	Harpsichord	
8	7	si02	Castanets	
9	27	si03	pitch pipe	Single instruments
10	11	sm01	Bagpipes	
11	10	sm02	Glockenspiel	
12	13	sm03	Plucked strings	Simple sound mixtures

First, at 128kbps bit rate the quality test result is shown in Figure 25:

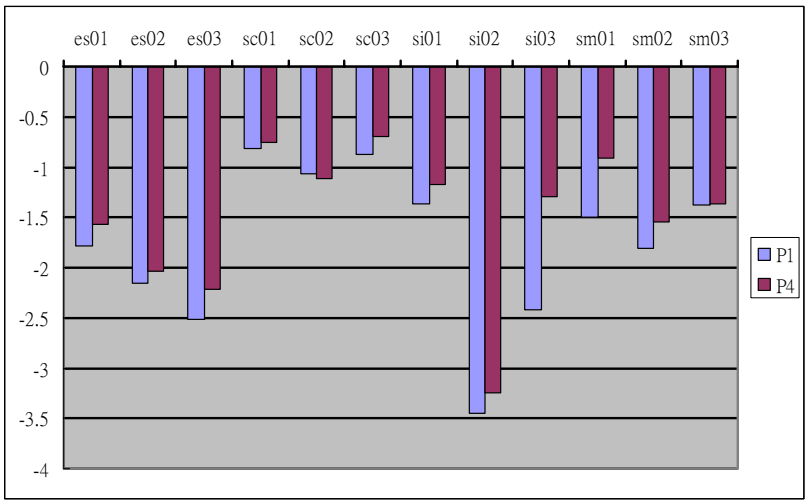

Figure 25: ODG at 128 kbps.

The P4 can get better quality than the conventional psychoacoustic models in the speech signal and single instrument and simple sound mixtures. Nevertheless, in the complex sound mixtures the ODG quality is equal to the P1 model. Second, at bit rate 112kbps result is Figure 26.

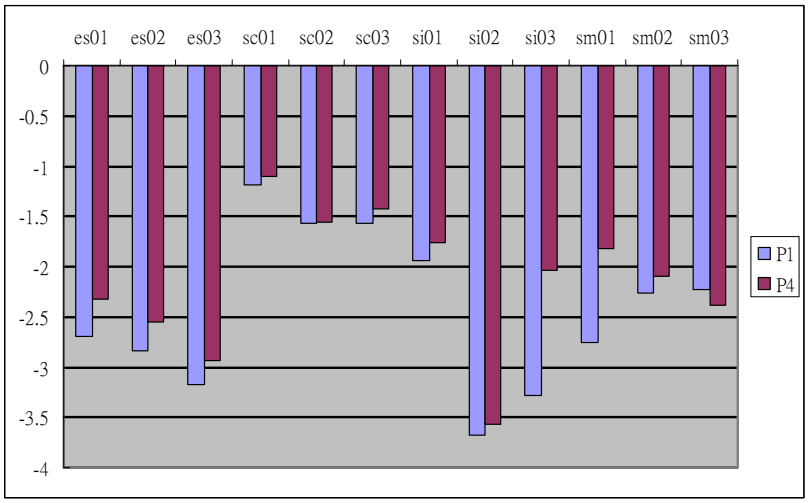

Figure 26: ODG at 112k.

The same result as that at 128kbps, P4 at 112kbps is much better than another model especially in speech signals and single instrument environments. At low bit rate 96k is as below Figure 27, the total quality degrades seriously in the low bit rates 96k but P4 obtains better quality than others even in complex sound mixtures.

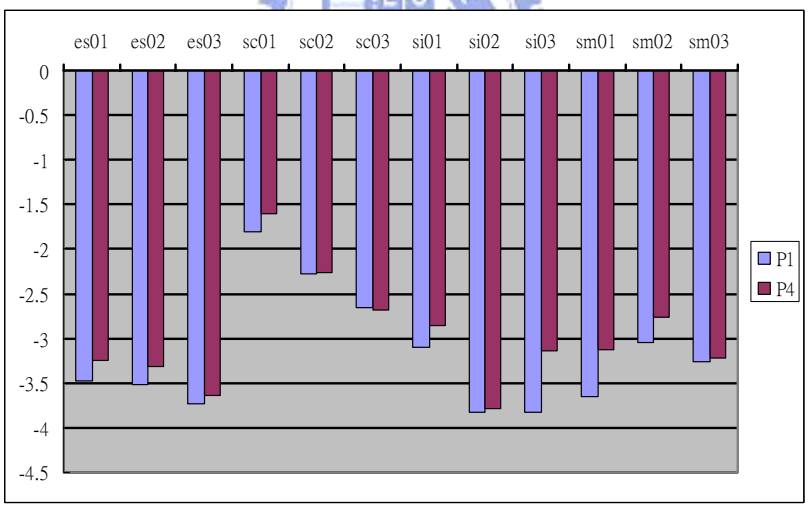

Figure 27: ODG at 96k.

The average, best, and worst of the above tests is shown in Figure 28. Consequently, P4 in different bit rate can also obtain the better grades. Moreover, P4 can enhance the quality 0.30 than P1 in the 112k bit rate.

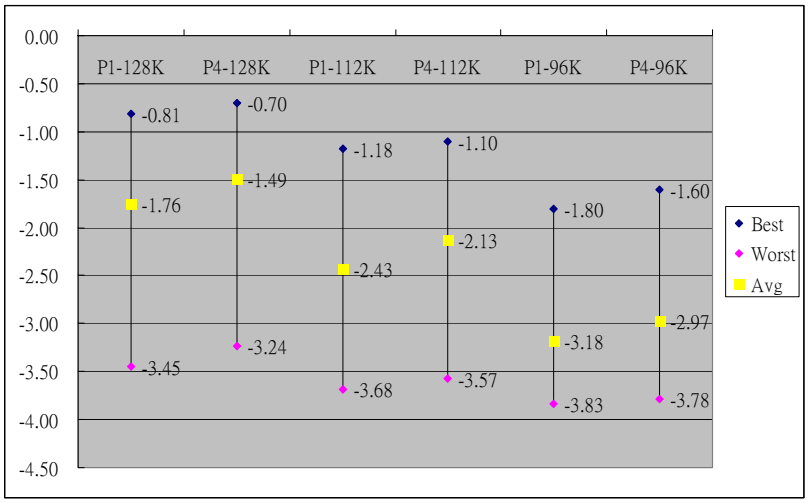

Figure 28: Illustration of the results in different bit rate in different model.

Besides the above MPEG 12 songs, we also test the three hundred critical tracks as below Table 6.

Table 6: Three hundred critical tracks.

	Categories		Remark
1	ff123	103	Killer bitstream collection from ff123
2	gpsycho	24	LAME quality test bitstream collection
3	HA128KTestV2	12	64 Kbps test bitstream for multi-format in HA forum
4	HA64KTest	39	128 Kbps test bitstream for multi-format in HA forum
5	horrible_song	16	Collections of killer songs among all bitstream in PSPLab
6	ingets1	5	Bitstream collection from the test of OGG Vorbis pre 1.0 listening test
7	Mono	3	Mono test bitstream
8	MPEG	12	MPEG test bitstream set for 48KHz
9	MPEG44100	12	MPEG test bitstream set for 44100 Hz

10	Phong	8	Test bistream collection from Phong
11	PSPLab	37	Collections of bitstream from early age of PSPLab. Some are good as killer.
12	sjeng	3	Small bitstream collection by sjeng
13	SQAM	16	Sound quality assessment material recordings for subjective tests
14	TestingSong14	14	Test bitstream collection from rshong
15	TonalSignals	15	Artificial bitstream that contain sin wave etc speciailly made bitstream to probe quality of encoder
16	VORBIS_TESTS	8	

The different psychoacoustic models are tested for above tracks in NCTU-AAC.

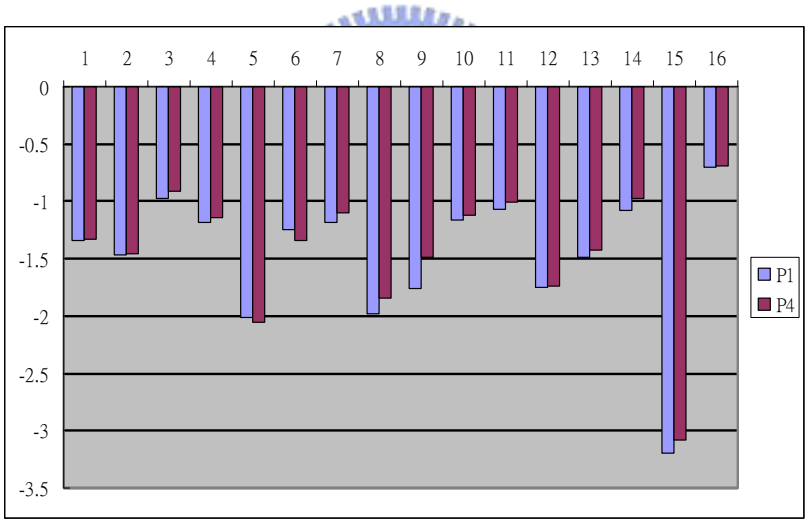

Figure 29: Different psychoacoustic models in three hundred critical tracks.

The proposed model in whole averages ODG can gain 0.04 than P1.

Chapter 5 Psychoacoustic Model based on Energy Floor

5.1 Masking Threshold Alignment

Conventional psychoacoustic model uses two masking offsets NMT and TMN to create the signal-to-masking ratio. However, the TMN offset actually overestimate the tone masking effect especially for tone-rich signals which will result in so-called fishy noise or birdie noise as shown in Figure 30.

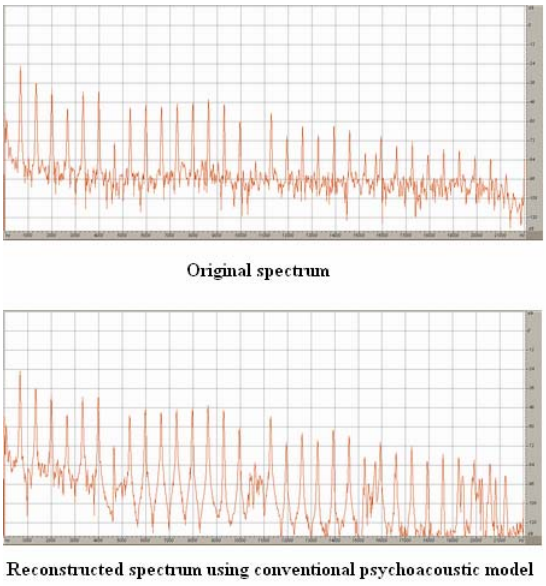

Figure 30: Illustration of fishy noise caused by the overestimation of masking threshold in conventional psychoacoustic model.

Figure 30 shows that the noise between two tones is disappear after reconstruction due to the overestimation of masking threshold. In fact, the noise is critical to human auditory system. Therefore, the proposed concept derives from another perspective of tradition energy calculation for the partition bands by the formula:

$$Masking\_Quantization_b = \frac{PM\_MaskingThreshold_b}{PM\_Energy_b} \times Energy\_Quantization_b \qquad (52)$$

However, the formula will result in the overestimation masking in the case of harmonic-rich signals as shown in Figure 31.

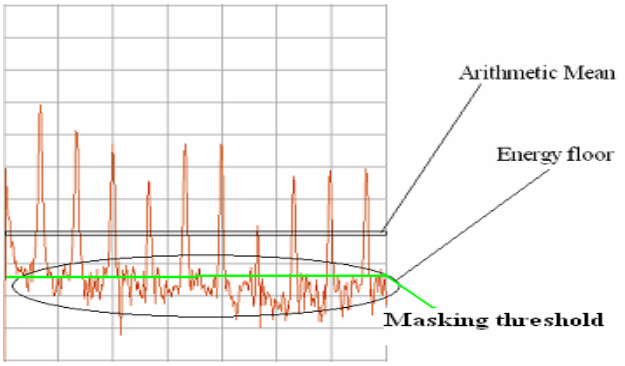

Figure 31: Illustration of the energy floor definition.

Because $PM\_Energy_b = AM_b \times Bandwidth_b$ and the masking effect of the noise is much stronger than tone, using AM_b to calculate the masking threshold will cause the overestimation of the energy which leads to an overestimation on the masking threshold due to the alignment between the signal in filterbank domain and the psychoacoustic model through the energy. The overestimation leads to noise generally referred to as the fishy noise or birdie noise [30].

Therefore from the viewpoint of energy floor, the masking threshold can be described as:

$$Masking\_Quantization_b = MSR_b \times Energy\_Quantization_b, \tag{53}$$

$$Masking\_Quantization_b = Energyfloor_b \times Bandwidth_b \times NoiseMasking_b, \tag{54}$$

$$MSR_b = 10^{\frac{-6}{10}} \times \frac{Energyfloor_b}{ArithmeticMean_b}. \tag{55}$$

Nevertheless in the low frequency, the bandwidth is very narrow such that the energy floor can result in error estimation as shown in Figure 32. So, the masking must have a constraint on maximum noise masking -6dB.

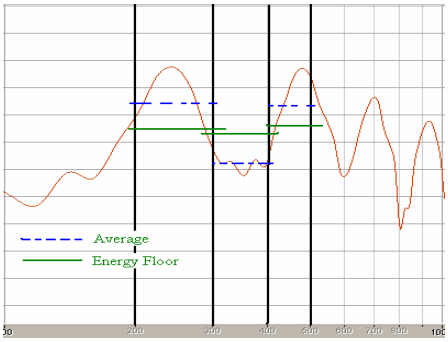

Figure 32: Illustration of the energy floor problem.

5.2 Energy floor

In Section 5.1, just employing the masking threshold of energy floor promises that the noise can not be masked. Therefore, this section focuses on the estimation of energy floor.

5.2.1 Smoothing

The thesis proposes an efficient energy floor estimation based on the smoothing and average as below:

$$\hat{x}_i = \frac{1}{Smooth\_Length} \times \sum_{k=i-Smooth\_Length/2}^{i+Smooth\_Length/2-1} x_k \tag{56}$$

where $Smooth\_Length$ means the length of the smoothing process and x_i means the i_{th} spectral line. And then,

$$Energyfloor_b = \frac{1}{Bandwidth_b} \times \sum_{for \text{ each partition band}} \hat{x}_i \quad . \tag{57}$$

As a result of the smoothing, the each spectral line will be smooth with neighbor lines. For example, a peak located in noise after the process of smoothing will be lower such that attendant average is more meaningful to represent the energy floor.

5.2.2 Recursive filter

[31] proposes a simple first-order recursive filter which is able to estimate the energy floor. A simple first-order recursive filter is designed as:

$$\hat{x}_i = \alpha \times \hat{x}_{i-1} + (1-\alpha) \times x_i \, . \tag{58}$$

And then,

$$Energyfloor_b = \frac{1}{Bandwidth_b} \times \sum_{for \text{ each partition band}} \hat{x}_i \quad . \tag{59}$$

5.2.3 Geometry Mean filter

$$Energyfloor_b = \sqrt[\frac{1}{Bandwidth_b}]{\prod_{i=0}^{N-1} x_i} \tag{60}$$

is a conservative estimation of energy floor which can validly degrade the strong peak signal. For example, Figure 33 shows the different methods for estimation of the energy floor.

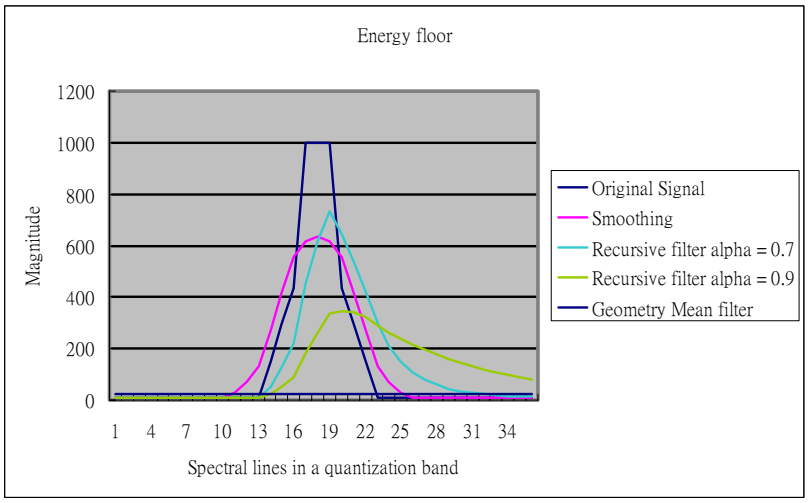

Figure 33: Illustration of estimations in the energy floor.

5.3 Detection of Tonal signal

Based on the energy floor, the psychoacoustic model also can detect attacks on the frequency domain as Sec4.2.

5.3.1 Detection of Tonal Attack Band

$$SMR_b = \frac{Signal_b}{Masking_b} \tag{61}$$

The SMR_b represents the degree of tone in the band. Therefore if this value is greater than a threshold, the band can be deemed as a tonal attack band as Subsec4.2.1.

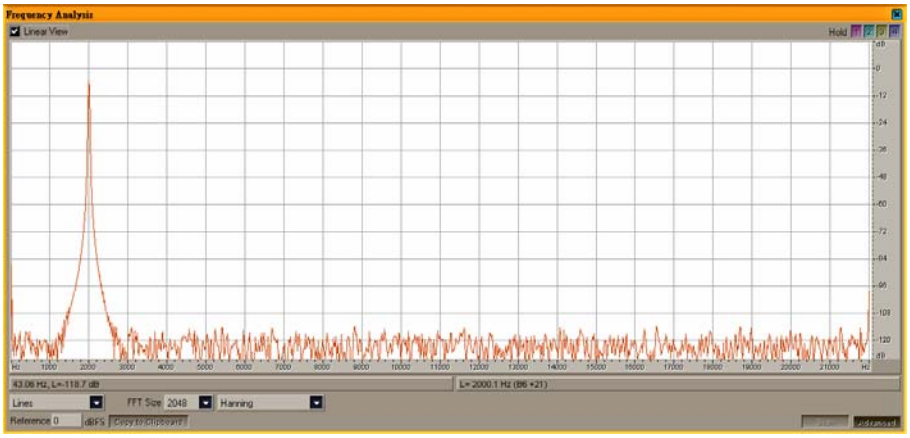

Figure 34: The peak signal at 2k.

Figure 34 shows the peak signal at 2k and detection result of tonal attack band is Figure 35 .

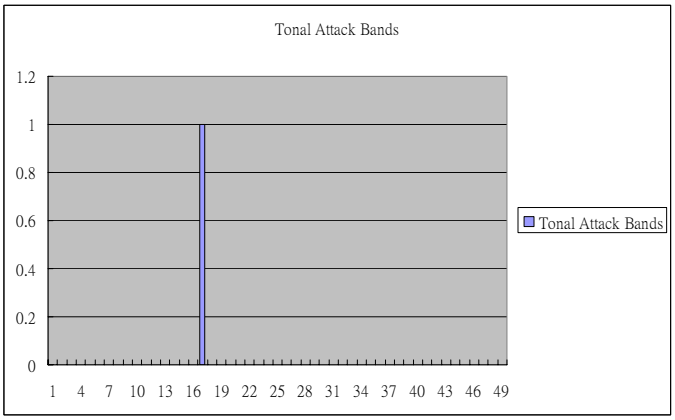

Figure 35: The detection of tonal attack band. x-axis means the quantization band and y-axis means the tonal attack band flag.

5.3.2 Detection of Tone-Rich Signal

Detection of tone-rich signals is similar to Subsec4.2.2. If the number of tonal attack bands in whole spectrum is over a threshold, the signal is deemed as a tone-rich signal.

5.4 Experiment

We use P1 representing conventional psychoacoustic model II, P4 representing proposed method in Chapter 4, and P5 representing psychoacoustic model based on energy floor. Tests in NCTU-MP3 is shown in the following Table:

Table 7: The computation time for NCTU-MP3.

NCTU-MP3	1	2	3	4	5	Average	Speedup over P1 (%)
P1	19.22	19.58	19.39	19.21	19.55	19.39	
P4	14.5	14.91	12.53	13.08	13.32	13.668	***29.51%***
P5	6.77	6.97	6.25	6.65	6.61	6.65	***65.70%***

Table 7 shows the result of this proposed psychoacoustic model applying to NCTU-MP3 compared with P1 and P4. From the table, the P5 can lead to complexity gain 65.7% over P1 and 51.35% over P4. We also test the encoding time in the NCTU-MP3 as

Table 8: Encoding time for the NCTU-MP3.

NCTU-MP3	Length	Encoding time (s)			P4 Speedup	P5 Speedup
File Name		P1	P4	P5	Percentage (%)	Percentage (%)
Es01	02:51	16	15	14	6.25	12.50
Es02	02:17	12	11	10	8.33	16.67
Es03	04:03	24	22	19	8.33	20.83
Sc01	02:55	17	16	13	5.88	23.53
Sc02	03:23	20	19	16	5.00	20.00
Sc03	03:04	19	18	15	5.26	21.05
si01	04:47	32	28	24	12.50	25.00
si02	03:05	21	20	16	4.76	23.81
si03	05:34	39	36	30	7.69	23.08
Sm01	04:27	32	28	24	12.50	25.00
Sm02	02:01	14	13	11	7.14	21.43
Sm03	04:11	29	27	23	6.90	20.69
Average		***22.92***	***21.08***	***17.92***	***7.55***	***21.13***

The P4 can gain 7.55% over P1. Moreover, the P5 can gain 21.13% more than P1. In conclusion, the proposed method can dramatically speed up the psychoacoustic model calculation in the different coders.

We also test the quality of the proposed psychoacoustic model in the NCTU-MP3 as:

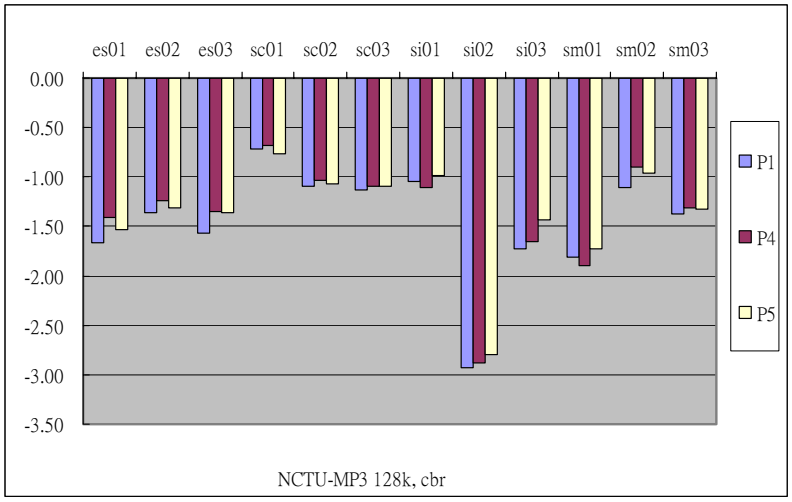

Figure 36: ODG test for the three psychoacoustic models under the NCTU-MP3.

The proposed psychoacoustic model still can get better quality. In Figure 36, the encoder based on P4 can have quality gain 0.08 over that based on P1, and furthermore the P5 can

have a gain 0.1 over P1.

We also test the proposed psychoacoustic model in NCTU-AAC as illustrated in Figure 37. the encoder based on P5 can gain 0.05 over that based on P1 in three hundred tracks.

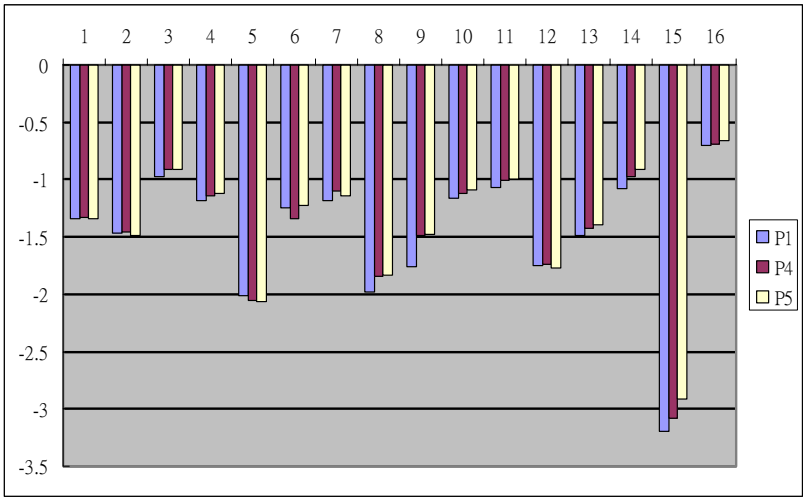

Figure 37: Three hundred tracks tested in NCTU-AAC.

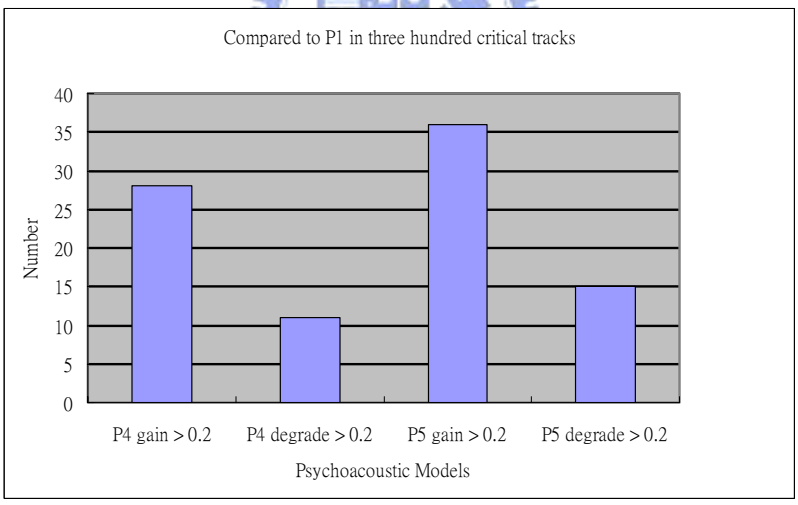

Figure 38: Compared to P1 in three hundred tracks.

x-axis in Figure 38 means the number of the model is 0.2 better or worse than P1. y-axis means the different psychoacoustic models.

Chapter 6 Conclusion

This thesis has proposed an efficient psychoacoustic model which can reduce the computation complexity by replacing FFT with filterbank and using SFM tonality decision. However, the thesis also addressed the detection of tonal signal. Moreover from aspect of the energy floor, this thesis only uses noise masking effect to calculate threshold which can effectively reduce the fishy noise problems. Finally, we have implemented this proposed psychoacoustic model in NCTU-AAC and NCTU-MP3 integrated with M/S coding, TNS coding, window switching, and bit reservoir. And, the speedup of the psychoacoustic model can achieve 70% in AAC and 65% in MP3. The quality has also improved by 0.2 in AAC and 0.1 in MP3 compared to the conventional psychoacoustic model.

References

[1] T. Nomura and Y. Takamizawa. Processor efficient implementation of a high quality MPEG-2 AAC encoder. In *AES 110th Convention*, 2001.

[2] Antonio S. Pena et al. ARCO (adaptive resolution COdec): a hybrid approach to perceptual audio coding. In *AES 100th Convention*, 1996.

[3] H. Fletcher, Auditory patterns, Rev. Mod. Phys., pp. 47–65, Jan. 1940.

[4] E. Terhardt, Calculating virtual pitch, Hearing Res., vol. 1, pp. 155–182, 1979.

[5] E. Zwicker and H. Fastl, Psychoacoustics Facts and Models. Berlin, Germany: Springer-Verlag, 1990.

[6] R. Hellman, Asymmetry of Masking Between Noise and Tone, Percep. and Psychphys., pp. 241-246, vol.11, 1972.

[7] B. Scharf, Critical bands, in Foundations of Modern Auditory Theory. New York: Academic, 1970.

[8] ISO/IEC, Coding of Moving Pictures and Audio— IS 13818-7 (MPEG-2 Advanced Audio Coding, AAC), Doc. ISO/IEC JTC1/SC29/WG11 n1650, Apr. 1997.

[9] ISO/IEC, Information Technology- Coding of audiovisual objects,—.ISO/IEC.D 14496 (Part 3, Audio), 1998.

[10] ISO/IEC, JTC1/SC29/WG11 MPEG, Information technology— Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s—Part 3: Audio, IS11172-3 1992 ("MPEG-1").

[11] B. C. J. Moore, Masking in the human auditory system, in Collected Papers on Digital Audio Bit-Rate Reduction, N. Gilchrist and C. Grewin, Eds., pp. 9–19, 1996.

[12] K. Brandenburg et al., ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio, J. Audio Eng. Soc., pp. 780-792, Oct. 1994.

[13] E. Scheirer and L. Ray, Algorithmic and Wavetable Synthesis in the MPEG-4 Multimedia Standard, in Proc. 105th AES Convention, Sep 1998.

[14] J. Johnston, Estimation of Perceptual Entropy Using Noise Masking Criteria, in Proc. ICASSP-88, pp. 2524-2527, May 1988.

[15] J. Johnston, Transform Coding of Audio Signals Using Perceptual Noise Criteria, IEEE J. Sel. Areas in Comm., pp. 314-323, Feb. 1988.

[16] ISO/IEC, Information Technology- Coding of audiovisual objects,—.ISO/IEC.D 14496 (Part 3, Audio), 1998.

[17] J. Johnston, S. Quackenbush, G. Davidson, K. Brandenburg, and J. Herre, MPEG audio coding, in Wavelet, Subband, and Block Transforms in Communications and Multimedia, A. Akansu and M. Medley, Eds. Boston, MA: Kluwer Academic, ch. 7, 1999.

[18] K. Brandenburg, E. Eberlein, J. Herre, and B. Edler, Comparison of filter banks for high quality audio coding, in Proc. IEEE ISCAS, pp. 1336–1339, 1992.
[19] J. Johnston, Audio coding with filter banks, in Subband and Wavelet Transforms, A. Akansu and M. J. T. Smith, Eds: Kluwer Academic, pp. 287–307, 1996.
[20] ISO/IEC, JTC1/SC29/WG11 MPEG, Information technology— Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s—Part 3: Audio, IS11172-3 1992 ("MPEG-1").
[21] ISO/IEC, JTC1/SC29/WG11 MPEG, Information technology—Generic coding of moving pictures and associated audio—Part 3: Audio, IS13818-3 1994 ("MPEG-2").
[22] Princen J, Bradley A, Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation, IEEE Transactions, ASSP-34, No.5, Oct 1986, pp. 1153-1161.
[23] H. Malvar, Lapped transforms for efficient transform/subband coding, IEEE Trans. Acoust., Speech, Signal Processing, vol. 38, pp. 969–978, June 1990.
[24] Princen J, Johnson A, Bradley, A, Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation, Proc. of the ICASSP, pp 2161-2164, 1987.
[25] Y. Wang and M. Vilermo, Modified Discrete Cosine Transform-Its Implications for Audio Coding and Error Concealment, AES 22th International Conference, June 2002.
[26] N. Jayant and P. Noll, Digital Coding of Waveforms Principles and Applications to Speech and Video. Englewood Cliffs, NJ: Prentice- Hall, 1984.
[27] ITU Radiocommunication Study Group 6, DRAFT REVISION TO RECOMMENDATION ITU-R BS.1387 - Method for objective measurements of perceived audio quality.
[28] QuickTime 6.3 http://www.apple.com/quicktime/
[29] Nero 6 http://www.nero.com/en/index.html
[30] H. W. Hsu, C. M. Liu, W. C. Lee, Audio Patch Method in Audio Decoders—MP3 and AAC, Audio Engineering Society 116th Convention, Berlin, Germany, May 8~11, 2004
[31] J. Johnston, Perceptual Coding of Audio Signals Employing Envelope Uncertainty, United States Patent 6466912 B1, 2002.