國立交通大學

資訊工程學系

碩士論文

MPEG-4 AAC 中的 PNS 模組之設計與 M/S 編碼技術之改良

Design of Perceptual Noise Substitution and M/S Coding Enhancement in MPEG-4 Advanced Audio Coding

研究生：蘇明堂

指導教授：劉啟民　教授

李文傑　博士

中華民國 九十四 年 六 月

MPEG-4 AAC 中的 PNS 模組之設計與 M/S 編碼技術之改良

Design of Perceptual Noise Substitution and M/S Coding Enhancement in MPEG-4 Advanced Audio Coding

研究生：蘇明堂　　Student：Ming-Ton Su

指導教授：劉啟民　　Advisor：Dr. Chi-Min Liu

李文傑　　Dr. Wen-Chieh Lee

國立交通大學

資訊工程系

碩士論文

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 2005
HsinChu, Taiwan, Republic of China

中華民國 九十四 年 六 月

MPEG-4 AAC中的PNS模組之設計
與MS編碼技術之改良

學生：蘇明堂　　　　　　　　　　　　　　　　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

在一般音訊壓縮所要處理的訊號中，常存在著許多的噪音訊號，可是一般的音訊壓縮器(如:mp3)，並沒有特別針對噪音設計一個模組來壓縮噪音，導致噪音訊號仍然被當成一般訊號來壓縮，消耗可觀的位元，讓一些對聽覺比較重要的訊號無法被更精確的保留下來。在新一代的MPEG-4所定的音訊壓縮標準中(AAC, Advanced Audio Coding)，特別提出了PNS(Perceptual Noise Substitution)模組，專門用來取代訊號中的噪音，以促進位元的有效利用，來達到更高的音訊壓縮品質。此論文提出了數種關於PNS模組的設計以及應用，不僅使用PNS來做噪音取代，也將PNS模組的功能做延伸，用來做高頻訊號的取代以及消滅在一般音訊壓縮器中常出現的zero-band現象，進而彌補壓縮上的缺陷。另外，此論文會根據我們實驗室先前所提出的M/S編碼技術再提出改進的方法，以消除在某些情形下會產生異音的現象，進而增進編碼品質。最後，此論文會以我們實驗室的AAC編碼器爲平台，將上述的方法加入到此編碼器上，其結果不論主觀測試(聽覺測試)或者客觀測試(ODG)，皆有相當的改善，能達到一定的聽覺品質進步。

Design of Perceptual Noise Substitution and M/S Coding Enhancement in MPEG-4 Advanced Audio Coding

Student: Ming-Ton Su　　　　Advisor: Dr. Chi-Min Liu

Dr. Wen-Chieh Lee

Institute of Computer Science and Information Engineering
National ChiaoTung University

Abstract

Perceptual Noise Substitution (PNS) is an efficient tool in ISO/MPEG-4 Advanced Audio Coding (AAC) to achieve better coding quality at low bit rates. The principle of PNS is to replace the noise-like signals in bands by random noise to reduce the bits required. On the design of PNS, this paper considers two different and complementary approaches: noise substitution approach and zero-band dithering approach. Noise substitution approach is to replace the noise-like bands before quantization process to increase the efficiency of bit usage and retain the signals in high frequency for the graceful bit allocation. Zero-band dithering approach is to replace zero-bands with proper energy after quantization process to eliminate birdie artifacts and enhance the coder output. This thesis discusses the design issues of noise substitution approach and zero-band dithering approach, including the noise-band detection mechanism, the replacement method, and the cooperation of two approaches. Furthermore, this thesis proposes an enhancing mechanism in M/S coding based on our previous work to eliminate some artifacts caused by the improper inter channel bit allocation. Finally, in the experiments, extensive experiments have been conducted to verify the possible risk and the robustness of propose methods based on the codec developed in our laboratory through both subjective listening test and objective quality measurement.

致謝

不知不覺，我已經在交通大學看過兩次梅竹賽，這意味著我在交大的七百多個日子要結束了，感謝兩年來指導、幫助過我的師長以及實驗室同學們，感謝你們在我有所疑問時，能適時地提供寶貴的意見，幫助我尋求解答進而促成此論文的誕生，也讓我兩年的碩士生涯，除了在專業知識的建立與培養外，也對於做研究的精神與方法獲益良多。另外，我要感謝我的父母以及家人，感謝他們對我長久以來的支持與照顧，讓我能專注於學業及研究上，不至於分心。最後，我要對所有曾經幫助、協助過我的每一個人，致上深深的感謝之意。

Contents

List of Figures

List of Tables

Chapter 1 Introduction

For the last twenty years, the great advances in audio coding have changed the way of audio signal appearance in our daily life and facilitated the development of numerous applications in audio storage, broadcasting, and transmission. With these various applications, a trend has been established to satisfy the demand for high quality digital audio delivered at low bit rates. In the mid-eighties, the compact disk (CD) was introduced by SONY and Philips and then became an international standard [1]. In the CD format, the audio signal is digitally represented as a stereo signal sampled at 44.1 kHz with a 16-bit resolution. This format results the data rate of CD to 1.41Mbps. Although the data rate is so high, the CD became very popular and essentially replaced the analog audio in 1990's. Since the success of CD, the application of digital audio signal increases widely. Unfortunately, the high data rate of CD is not acceptable in some applications. In order to reduce the bit rate of digital audio coding, the perceptual audio coding technique is introduced. The perceptual audio coding encodes an audio signal transparently by using models of human auditory perception. That means the perceptual audio encoder can remove the perceptually irrelevant signals and shape the noise introduced by quantization process to be inaudible. In some researches, perceptually encoded signal can have a SNR around 15 dB and be indistinguishable from the original signal, while the unshaped one have as high as 60 dB [2]. Nowadays, it is feasible to compress CD-quality stereo audio signals at 128 kbps via a perceptual audio encoder.

Figure 1 illustrates the basic architecture of the generic perceptual audio encoder. Typically, the inputs of the audio signals are in PCM format rather than the analog inputs. The input signals are segmented into overlapped blocks and transformed into frequency domain through the time-to-frequency mapping routine. The frequency domain data is then quantized and coded with the parameters decided by bit allocation algorithm. The psychoacoustic model analyzes the input signals and evaluates the perceptual information required including signal-to-masking ratio (SMR). The bit allocation routine allocates a limited number of bits to the frequency domain data to minimize the distortion for human hearing. Finally, the MUX routine packs all the coded information into the standard format.

At the end of eighties, the International Organization for Standardization (ISO) formed the Motion Picture Experts Group (MPEG) to develop standards for high quality video and audio coding. The MPEG group has established several existing standards. MPEG-1 standard was the first milestone achieved by this committee in November 1992. MPEG-1 is an international compression standard that addresses the compression of synchronized video and audio at a total data rate of 1.5Mbit/s [3]. In November 1994, the MPEG-2 standard was finalized in order to extend the features of MPEG-1 including multichannel extension and the audio coding systems at lower sampling rates.

Followed the success of MPEG-1 and MPEG-2, the MPEG-4 became an ISO/IEC final draft international standard, FDIS, on October 1998 (ISO/IEC 14496 version 1). The second version of MPEG-4 was finalized in December 1999. The MPEG-4 standard targets at a wide range of applications including wired, wireless, streaming, digital broadcasting, interactive multimedia and high quality audio/video. For above various applications, the MPEG-4 audio coding standard extends MPEG-2 with several tools and each tool can be used as a component to making coding more efficient and more flexible.

The target of this thesis focuses on the Perceptual Noise Substitution (PNS) in MPEG-4 AAC. The PNS is a tool which is extended by MPEG-4 to replace the noise components without any obvious distortion in human perception. This thesis proposes an efficient approach to substitute the noise-like components with PNS parameters. Furthermore, this thesis also extends the PNS to substitute the high frequency signals to increase the efficiency of bit usage and to dither zero band after quantization process to eliminate artifacts. The thesis is organized as follows. In Chapter 2, basic concepts of audio coding are described including psychoacoustic principles and audio quality measurements. In Chapter 3, the PNS tool in MPEG-4 AAC is introduced. Then the design issues of proposed PNS approaches are discussed in Chapter 4. In Chapter 5, the enhanced psychoacoustic model and M/S coding based on our previous works are proposed. In Chapter 6, extensive experiments are made to prove the improvement of the proposed PNS approaches, modified psychoacoustic model and M/S coding. Finally, the conclusion and future work are discussed in Chapter 7.

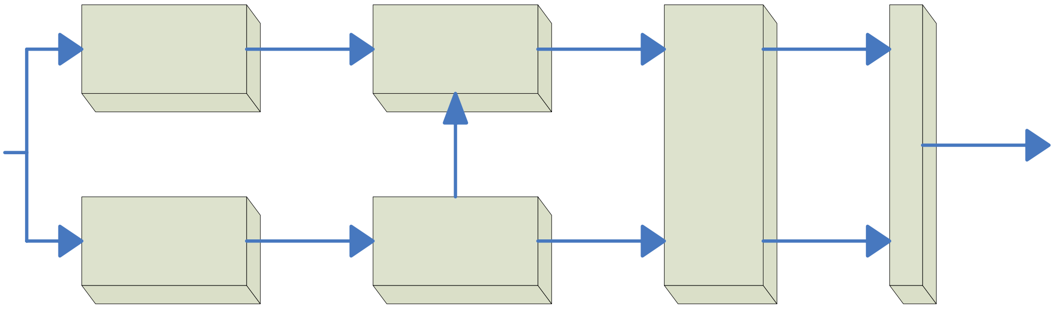


Figure 1: Generic perceptual audio encoder

Chapter 2 Basic Concepts of Audio Coding

2.1 Psychoacoustic Model

The goal of audio coding is to use minimum bits to represent the audio signals with a perceptually lossless quality for human hearing. For an ideal audio coding system, humans can not tell the differences between the original and coded signals. To achieve this goal, the psychoacoustic model is introduced to simulate the human auditory system. The human auditory system has some interesting properties. Human hearing has a dynamic frequency range from about 20 to 20000 Hz, and hears sounds with intensity varying over many magnitudes. The hearing system may thus seem to be a very wide-range instrument, which is not altogether true. In current audio coding, only some critical properties of human auditory system are known and used since there is no model that can simulate the human hearing process precisely now. Even though, the listening quality of the coded audio signals is significantly affected by these psychoacoustic properties. Involving these critical psychoacoustic principles in audio coding, the audio developers can remove the perceptually irrelevant information, shape the quantization noise to be inaudible, and improve the listening quality significantly. In this section, we will briefly introduce these psychoacoustic principles including absolute hearing threshold, critical bands, and masking effects.

2.1.1 Absolute Threshold of Hearing

In a noiseless environment, the amount of energy needed in a pure tone such that it can be detected by a listener is called the absolute threshold of hearing (ATH). The absolute threshold of hearing is varying from frequency to frequency. In 1979, Terhardt [4] proposed a well approximated nonlinear function:

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

f is the frequency index, $T_q(f)$ is the absolute threshold of hearing in frequency *f*. The curve of above function was shown in Figure 2. In general, the absolute threshold of hearing can be measured by increasing the sound pressure level (SPL) of a test tone to the listeners. Then, the absolute threshold of hearing in all frequencies can be measured by increasing the frequency of the test tone from low to high. Since the signals lower than the absolute threshold of hearing will be inaudible, those signals can be removed from the input samples to improve the coding efficiency. Furthermore, $T_q(f)$ can also be considered as the maximum allowable energy level of coding distortion. That is, the quantization error will be inaudible and can be omitted if it is lower than the absolute threshold of hearing.

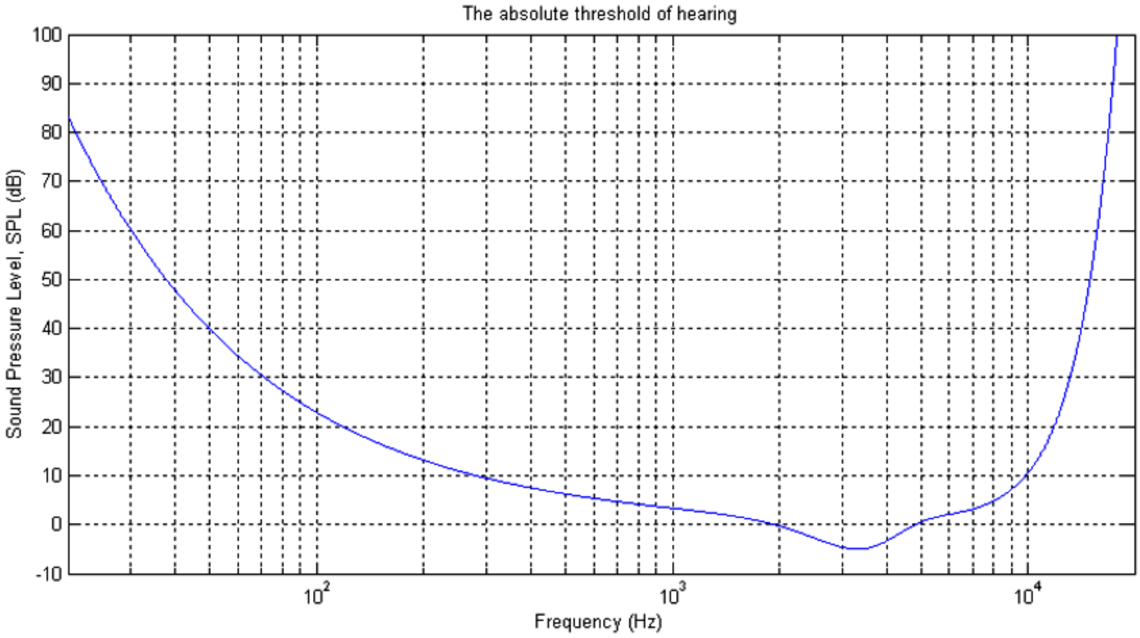


Figure 2: The absolute threshold of hearing in quiet [4]

2.1.2 Critical Bands

In perceptual audio coding, the absolute threshold of hearing provides only the basic utility to shape the coding distortions. To understand and simulate the human auditory system, the critical band must be involved. The critical band structure and the related analysis can be used to describe the behavior of the auditory system in many aspects. A basic definition of the critical band is “the bandwidth at which subjective response changes abruptly” [5]. That means the human perception of signals within the same critical band will be similar and not change rapidly. The basic unit of critical band rate is Bark. The length of one Bark on the basilar membrane is about 1.3 mm [6]. Experiments revealed that 25 critical bands exist over the frequency range of human hearing and the bandwidth of critical band can be approximated by

$$\Delta f_G = 25 + 75(1 + 1.4 f^2)^{0.69} \text{ (Hz)}, \qquad (2)$$

where f is frequency, expressed in kHz. The center and edge frequencies of these 25 critical bands are shown in Table 1 [7].

For a given frequency, the critical band is the smallest band of frequencies around it which activate the same part of the basilar membrane. Whereas the different threshold is the just noticeable difference of a single frequency, the critical bandwidth represents the ear’s resolving power for simultaneous tone or partials. In a complex tone, the critical bandwidth corresponds to the smallest frequency difference between two partials such that each can still be heard separately. It may also be measured by taking a sine tone barely masked by a band of white noise around it. When the noise band is narrowed until the point where the sine tone becomes audible, its width at that point is the critical bandwidth. Simultaneous tones lying within a critical bandwidth do not give any increase in perceived loudness over that of the single tone and provided the sound pressure level remains constant. For tones lying more than critical bandwidth apart, their combination results in increased loudness. When two tones are close together in frequency, the resulting tone is a confusion of the two frequencies. If the frequency difference increases, the roughness in the tones appears. The phenomenon appears because both frequencies are activating the same part of the basilar membrane. Further apart, the two frequencies can be discriminated separately, whereas the roughness only occurs at a frequency separation equal to the critical bandwidth.

Table 1 Critical Band Center and Edge Frequencies [7]

Band (Bark)	Lower (Hz)	Center (Hz)	Upper (Hz)	Band (Bark)	Lower (Hz)	Center (Hz)	Upper (Hz)
1	0	50	100	14	2000	2150	2320
2	100	150	200	15	2320	2500	2700
3	200	250	300	16	2700	2900	3150
4	300	350	400	17	3150	3400	3700
5	400	450	510	18	3700	4000	4400
6	510	570	630	19	4400	4800	5300
7	630	700	770	20	5300	5800	6400
8	770	840	920	21	6400	7000	7700
9	920	1000	1080	22	7700	8500	9500
10	1080	1170	1270	23	9500	10500	12000
11	1270	1370	1480	24	12000	13500	15500
12	1480	1600	1720	25	15500	19500	
13	1720	1850	2000				

2.1.3 Simultaneous Masking Effect

Masking effect is one of the most important concepts for perceptual audio coding to hide the coding distortions to be inaudible. The masking effect is a phenomenon of human auditory system that the threshold of audibility of one sound is raised by the presence of another sound. Masking effect occurs in both frequency and time domains, and the masking effect of a signal depends on the frequency, structure, and the energy level of both masker and maskee.

Simultaneous masking effect is the masking effect which occurs in frequency domain. Simultaneous masking occurs when two stimuli are simultaneously presented to the auditory system and one of them is made inaudible by the other. Physiological evidence reveals that the simultaneous masking is caused due to the function of the basilar membrane and the hair cells. Many researchers believe that the masker produces a great amount of activities on the basilar membrane such that any activity caused by the weaker signal may become undetectable. In physiology, the hair cells detect the strongest vibration in any critical band along the basilar membrane. Simultaneous masking is typically determined for a noise masking a noise (NMN), a tone masking a tone (TMT), a noise masking a tone (NMT), or a tone masking a noise (TMN). However, in practice, only TMN and NMT are often involved to determine the simultaneous masking effect in order to reduce the complexity. Figure 3 shows the noise-masking-tone effect and the energy level of a test tone that just masked by narrow band noise [8]. The tone-masking-noise effect is shown in Figure 4 for a narrow band noise and a tonal masker [9].

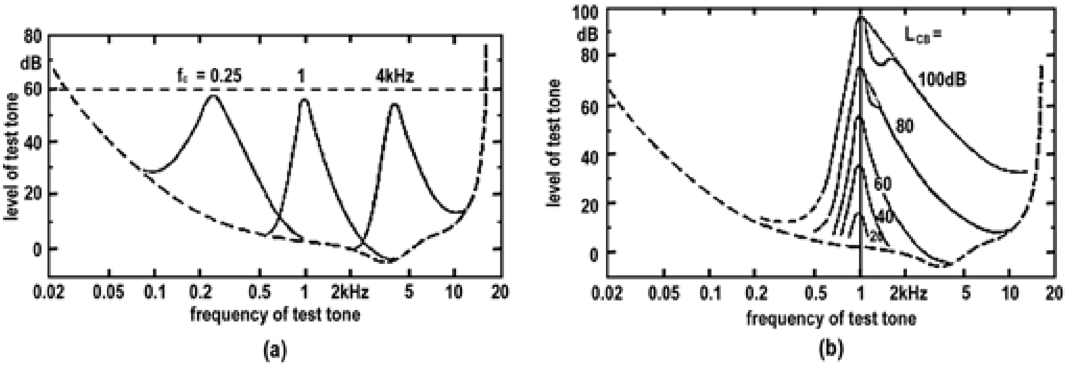


Figure 3: The broken line is the absolute threshold of hearing [8]. (a) The noise level is 60 dB and the center frequency varies. (b) The center frequency of the noise is 1 kHz and the energy level varies

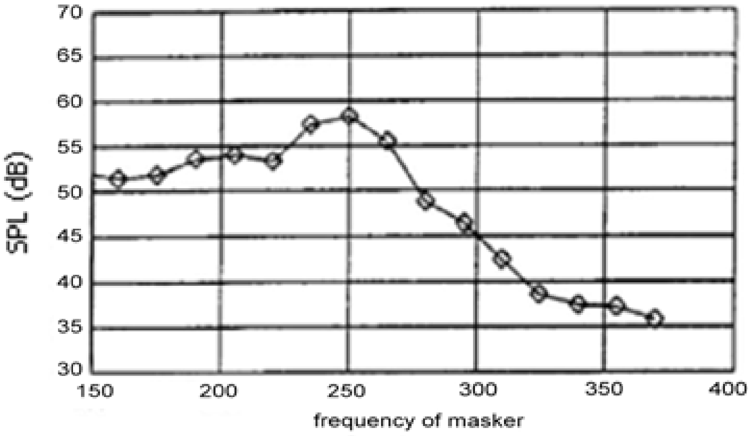


Figure 4: The energy level required for a narrow band noise to be auditable in a tonal masker [9]. The center frequency of the noise is 250Hz and the energy level of the tone is 80 dB

2.1.4 Non-simultaneous Masking Effect

Non-simultaneous masking effect is the masking effect which occurs in time domain. Non-simultaneous masking occurs when the masker and the maskee are not presented to the hearing system at the same time. Sometimes a signal can be masked by a sound preceding it, called pre-masking, or even by a sound following it, called post-masking. The characteristic of non-simultaneous masking for human hearing system is asymmetric, meaning that the pre-masking effect is much less than the post-masking. Pre-masking occurs before the presence of the masker and lasts approximately 20 milliseconds. However, the significant pre-masking tends to last only 1-2 milliseconds. Post-masking occurs after the vanishment of the masker and lasts more than 100 milliseconds [6]. Figure 5 illustrates the non-simultaneous masking effect including pre-masking and post-masking. For the purpose s of perceptual audio coding, abrupt audio signal transients (e.g., the onset of a percussive musical instrument) create pre-masking and post-masking regions in time. During these time slots, a listener will not perceive the signal which is beneath the raised audibility threshold produced by the masker. In fact, non-simultaneous masking has been used in several audio coding algorithms [10]-[14]. Pre-masking in particular has been exploited in conjunction with adaptive block size transform coding to compensate for pre-echo distortions.

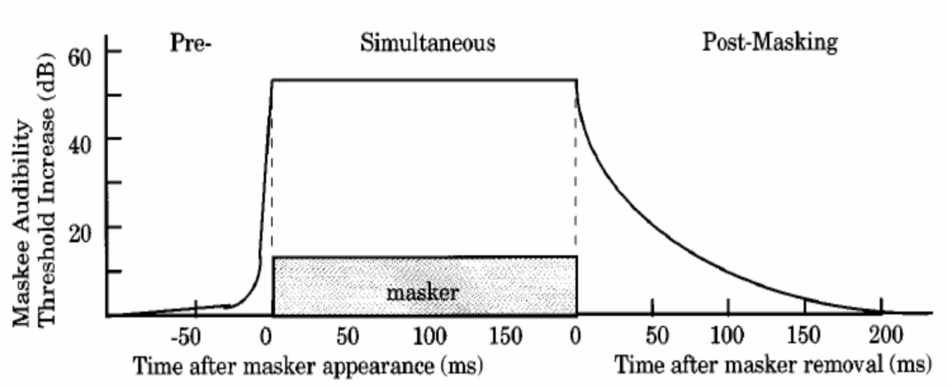


Figure 5: Illustration of the non-simultaneous masking effect [6]

2.2 Quality Measurement

The audio quality of a coding system can be described as the perceived difference between the output of a system and the original signal. The most trivial way to assess the audio quality is to perform a listening test, also called subjective quality measurement. Subjective quality measurement is evaluating the quality by human ears directly. However, performing a subjective quality measurement is very expensive and time consuming. Another way to assess the audio quality is to perform objective quality measurement. Objective quality measurement is to predict the basic audio quality by using objective measurements incorporating psychoacoustic principles. The objective quality measurement is cheaper and more efficient to assess the quality of an audio coding system. However, the objective way is not accurate enough to replace the subjective one and not generally accepted now. In this section, both the subjective and objective quality measurements are introduced below.

2.2.1 Subjective Quality Measurement

The ITU-R BS.1116 [15] is a standard of subjective audio quality measurement and very effective in evaluating high quality audio system with small impairments. The grading scale used in BS.1116 listening test is based on the five-grade impairment scale as defined by ITU-R BS.562-3 [16] and shown in Figure 6. According to BS.562-3, any perceived difference between the reference signal and the test signal should be matched to one of the discrete five scales based on the degree of the impairment. In BS.1116, the ratings are represented on a continuous scale between 1.0~5.0. Scale “1.0” stands for highly annoying impairment and “5.0” for transparent coding.

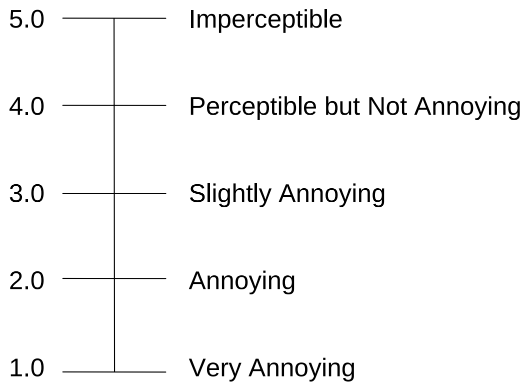


Figure 6: ITU-R five-grade impairment scale [16]

The test method most widely accepted for subjective listening test is the so-called "double-blind, triple-stimulus with hidden reference" method. In this method, the listener is presented with three signals: the reference signal "R" and then the test signals "A" and "B". Either A or B will be identical to the reference signal and the other will be the coded signal. The assignment of A and B will be done randomly so that none of the listeners could predict which signal is identical to the reference one. The listeners are asked to assess the impairment of A compared to R, and of B compared to R according to the grading scale in Figure 6. Since one of the test signals is actually the reference signal, one of them should receive a grade of 5.0 while the other may receive a grade that describes the listener's assessment of the impairment.

The double-blind, triple-stimulus with hidden reference method has been implemented in various ways. For example, the system under test can be a real-time hardware implementation or a software simulation of the system. The stimuli can be presented with a tape-based reproduction or with a playback system from computer hard disk. The listener is allowed to switch between R, A or B and to loop through the test. The inclusion of the hidden reference in each trial provides an easy mean to check that the listener does not consistently make mistakes and therefore provides a control condition on the expertise of the listener. The double-blind, triple-stimulus with hidden reference method has been employed worldwide for many formal listening tests of perceptual audio codecs. The consensus is that it provides a very sensitive, accurate, and stable way of assessing small impairments in audio systems. In general formal listening tests have shown very good reliability in the evaluation of audio coding systems and high correlation in their results, see for example [17]-[20].

2.2.2 Objective Quality Measurement

The purpose of objective quality measurement is to predict the basic audio quality by using objective measurements incorporating psychoacoustic principles. Objective quality measurements that incorporate perceptual models have been introduced since the late 70's [21]. More recently, psychoacoustic models have been exploited in the measurements of perceived quality of audio coding systems, see for example [23]-[26]. The effectiveness of objective quality measurements can only be assessed by comparison with corresponding scores obtained from subjective listening test. One of the first global opportunities for correlating the results of these different audio objective evaluations with informal subjective listening test results arose in 1995 in the early stages of the development of the MPEG-2 AAC codec. The need to test different reference models in the development of MPEG-2 AAC led to the study of objective tests as a supplement and as an alternative to listening tests. Unfortunately, none of the objective quality measuring techniques under the examination at that time showed reliable correlation with the results of the listening test [27]. The recent adoption by ITU-R of PEAQ in BS.1387 [28] came in conjunction with data that proved the correlation between PEAQ objective difference grades, ODGs, with the subjective difference grades, SDGs, obtained averaging the results of previous formal subjective listening test [29]. While PEAQ is based on a refinement of generally accepted psychoacoustic models, it also includes new cognitive components to account for high-level processes that come to play a role in the judgment of audio quality.

PEAQ was used to generate objective quality measurements for audio data previously utilized in formal listening tests of state-of-the-art perceptual audio codecs. The performance of PEAQ was evaluated in different ways. The objective and mean subjective ratings were compared for each critical audio item used in formal tests. Then the objective and subjective overall system quality measurements were compared by averaging codec quality measurements over critical items. The correlation between subjective and objective results proved very good and analysis if SDG and ODG showed no significant statistical differences [29]. The accuracy of the ODG demonstrated the capacity of PEAQ to correctly predict the outcome of the formal listening tests including the ranking of the codecs in terms of measured quality. PEAQ was also tested as a tool in aiding the selection of critical material for formal listening tests. On the basis of quality measurement, the PEAQ set of critical material included more than half the critical sequences used in the formal listening test under exam [29].

Chapter 3 Perceptual Noise Substitution in MPEG-4 Advanced Audio Coding

3.1 PNS Tool Description

While generic audio coding predominately employs methods to coding the waveform of the input signal, the other coding methods do not aim at preserving the waveform of the input signal but at reproducing a perceptually equivalent output signal at the decoder end. Relaxing the requirement for preserving waveform may lead to significant saving in bits when parts of the signals are reconstructed from a compact parametric representation of signal features. Most audio signals contain frequency regions in which the human auditory system cannot detect sinusoids due to the nonstationary nature of those regions. Generic audio coding algorithms assume that any frequency region can be analyzed by the human auditory system at any time. Even the white noise which is perceptually irrelevant for human ear is coded. Coding noisy signals in this way will cause high bit rates for transmission and large disk space for storage. If noise components of signals can be detected and coded with information about their energy levels, frequency range, and time range, the above problem will be decreased. Coding noise components with a few parameters will lead to a great reduction in the amount of data required to code these signals.

The Perceptual Noise Substitution (PNS) technique is based on the observation that the subjective perception stimulated by noise-like signals is based on its energy level, frequency range, and time range, not the actual waveform of the signals. From a historic perspective, this phenomenon is well-known and has been used widely e.g. in parametric speech coding where unvoiced speech segments are synthesized from a parametric representation of the noise signal rather than a description of its waveform [30]-[31]. The concept of PNS is straight forward and can be described as follows [32] (see Figure 7):

- In the encoder, noise-like components of the input signal are detected on a frequency band.
- The groups of spectral coefficients belonging to scalefactor bands containing noise-like components are not quantized and coded as usual but omitted from quantization and coding processes.
- Instead, only a noise substitution flag and the power of the substituted spectral coefficients are transmitted for each of these bands.
- In the decoder, pseudo random vectors with the desired total power are inserted for the substituted spectral coefficients.

The PNS approach will result in a highly compact representation of the noise-like spectral components since only the signaling and the energy information are transmitted once for a scalefactor band rather than a set of quantized and coded spectral coefficients.

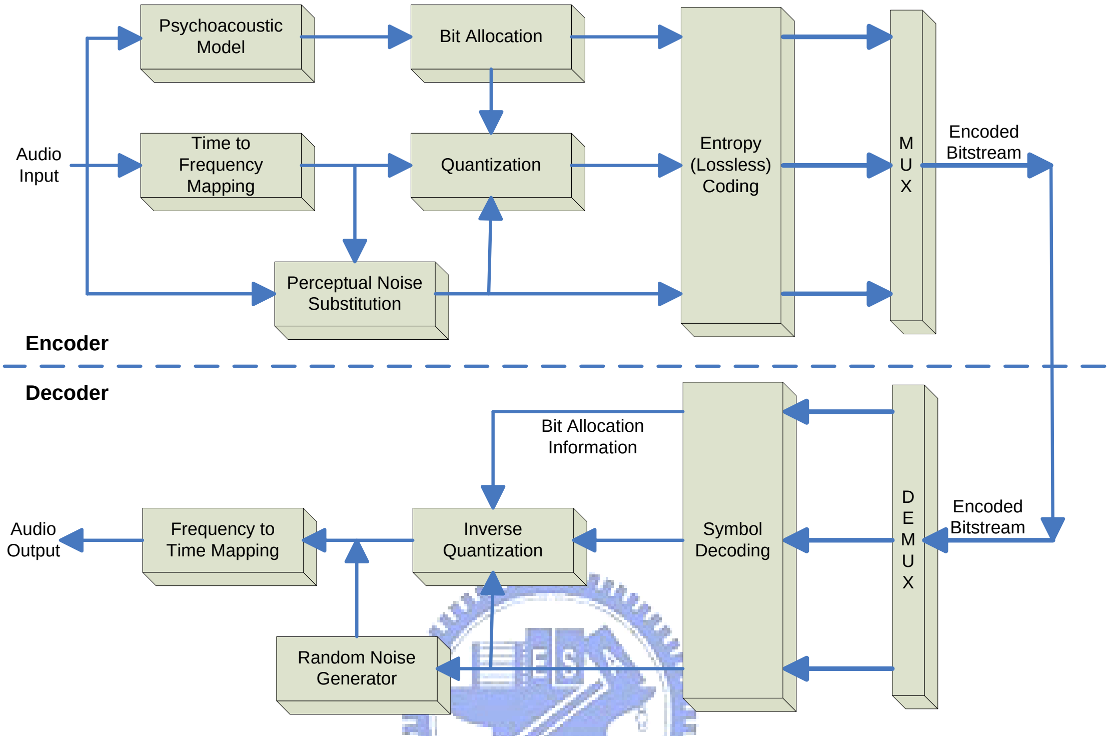


Figure 7: The principle of perceptual noise substitution

3.2 The Implementation of PNS in MPEG-4 AAC

In the MPEG-4 standard [33], the PNS tool is used to implement perceptual noise substitution coding within an individual channel stream (ICS), certain sets of spectral coefficients are derived from random vectors rather than from Huffman coded symbols and an inverse quantization process. This is done selectively on a scalefactor band and group basis when perceptual noise substitution is flagged as active. Figure 8 shows the block diagram of MPEG-4 general audio non-scalable encoder.

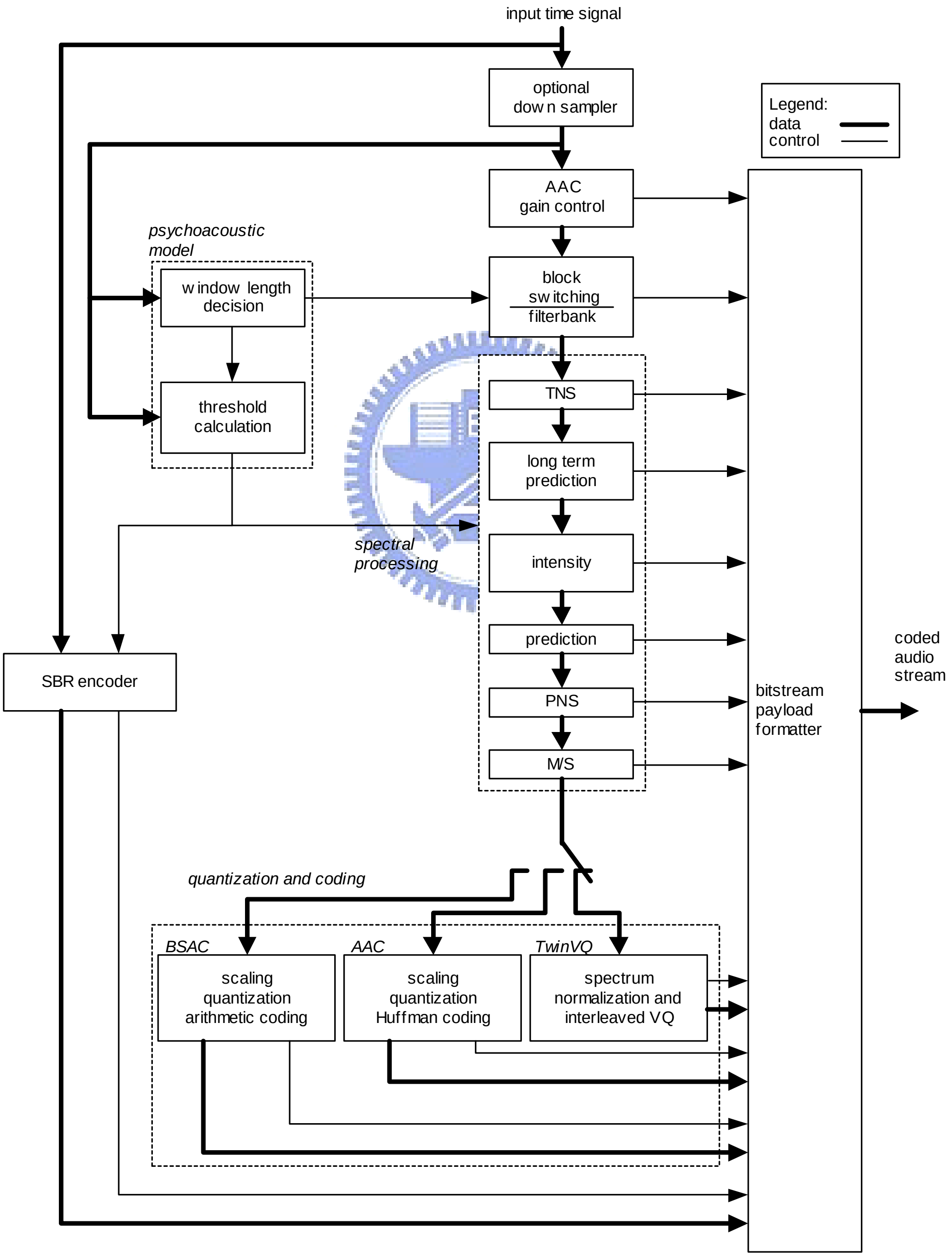


Figure 8: Block diagram of general audio non-scalable encoder [33]

3.2.1 PNS Decoding Process

Symbol Definitions:

hcod_sf[] Huffman codeword from the Huffman code table used for coding of scalefactors (see [12]).

dpcm_noise_nrg[][] Differentially encoded noise nergy.

noise_nrg[g][sfb] Noise energy for group g and scalefactor band sfb.

spec[] Array containing the channel spectrum of the respective channel.

ms_used[g][sfb] One-bit flag per scalefactor band indicating that M/S coding is being used in group g and scalefactor band sfb.

The used of perceptual noise substitution tool is signaled by the use of the pseudo Huffman codebook *NOISE_HCB* (13). Furthermore, if the same scalefactor band and group is coded by perceptual noise substitution in both channels of a channel pair, the correlation of the noise signal can be controlled by means of the *ms_used* field. While the default noise generation process works independently for each channel (separate generation of random vectors), the same random vector is used for both channels if the *ms_used* flag is set for a particular scalefactor band and group. In this case, no M/S stereo coding is carried out (because M/S stereo coding and noise substitution coding are mutually exclusive). The energy information for perceptual noise substitution decoding is represented by a “noise energy” value indicating the overall power of the substituted spectral coefficients in units of 1.5 dB. If noise substitution coding is active for a particular group and scalefactor band, a noise energy value is transmitted instead of the scalefactor of the respective channel. Noise energies are coded just like scalefactors, i.e. by Huffman coding of differential values:

- The start value of the DPCM decoding is given by *global_gain*.
- Differential decoding is done separately between scalefactors, intensity stereo positions and noise energies. In other words, the noise energy decoder ignores interposed scalefactors and intensity stereo position values.
- The same codebook is used for coding of noise energies as for scalefactors.

One pseudo function is defined for use in perceptual noise substitution decoding:

```
function is_noise(group, sfb) {
    1    for window group / scalefactor bands with codebook
         sfb_cb[group][sfb] == NOISE_HCB
    0    otherwise
}
```

The constant *NOISE_OFFSET* is used to adapt the range of average noise energy values to the usual range of scalefactor and has a value of 90.

The function gen_rand_vector(addr, size) generates a vector of length <size> with signed random values of average energy *MEAN_NRG* per random value. A suitable random number generator can be realized using one multiplication/accumulation per random value.

The noise substitution decoding process for one channel is defined by the following pseudo code:

```
nrg = global_gain – NOISE_OFFSET – 256;
for (g=0; g<num_window_groups; g++) {
    /* Decode noise energies for this group */
    for (sfb=0; sfb<max_sfb; sfb++)
        if (is_noise(g,sfb))
            noise_nrg[g][sfb] = nrg += dpcm_noise_nrg[g][sfb];
    /* Do perceptual noise substitution decoding */
    for (b=0; b<window_group_length[g]; b++) {
        for (sfb=0; sfb<max_sfb; sfb++) {
            if (is_noise(g,sfb)) {
                offs = swb_offset[sfb];
                size = swb_offset[sfb+1] – offs;

                /* Generate random vector */
                gen_rand_vector( &spec[g][b][sfb][0], size);
                scale = 1/(size * sqrt(MEAN_NRG));
                scale *= 2.0^(0.25*noise_nrg[g][sfb]);
                /* Scale random vector to desired target energy */
                for (i=0; i<len; i++)
                    spec[g][b][sfb][i] *= scale;
            }
        }
    }
}
```

3.2.2 Integration with Intra Channel Prediction Tools

For scalefactor bands coded using PNS, the corresponding predictors are switched to “off”, thus overriding the status specified by the *prediction_used* mask. In addition, for scalefactor bands coded by perceptual noise substitution, the predictors belonging to the corresponding spectral coefficients are reset [12]. The update of these predictors is done by feeding a value of zero as the “last quantized value” $x_{rec}(n-1)$. In Long Term Prediction (LTP), the scalefactor bands coded using PNS are not predicted.

3.2.3 Integration with other AAC Tools

The following interactions between the perceptual noise substitution tool and other AAC tools take place:

- During Huffman decoding of the quantized spectral coefficients, the Huffman codebook table *NOISE_HCB* is treated exactly like the zero codebook *ZERO_HCB*, i.e. no Huffman codewords are read for the corresponding scalefactor band and group.
- If the same scalefactor band and group is coded by perceptual noise substitution in both channels of a channel pair, no M/S stereo decoding is carried out for this scalefactor band and group.
- The pseudo noise components generated by the perceptual noise substitution tool are injected into the output spectrum prior to the temporal noise shaping (TNS) process step.

3.2.4 Integration into a Scalable AAC-based Coder

The following rules apply for usage of perceptual noise substitution tool in a scalable AAC-based coder:

- If a particular scalefactor band and group is coded by perceptual noise substitution, its contribution to the spectral components of the reconstructed output signal for the update of the intra channel predictor is omitted.
- If a particular scalefactor band and group is coded by perceptual noise substitution, its contribution to the spectral components of the output signal is omitted if spectral coefficients are transmitted for this scalefactor band and group in any of the higher layers by means of a non-zero codebook number.
- If a particular scalefactor band and group is coded by perceptual noise substitution in both channels of a channel pair, the higher layers may still use the M/S stereo flag *ms_used* to signal the use of M/S stereo decoding.

3.2.5 Suggested Encoding Procedure for PNS

In MPEG-4 AAC, the encoding procedure for perceptual noise substitution is similar to the coding procedure for intensity stereo and is performed as follows:

- For each scalefactor band containing spectral coefficients above a lower border frequency (e.g. 4 kHz) a noise detection procedure is carried out. The scalefactor band is classified as noise-like if the corresponding signal is neither tonal nor contains strong changes in energy over time. The tonality of the signal can be estimated by using the tonality values calculated in the psychoacoustic model. Similarly, changes in signal energy can be evaluated using the FFT energies calculated in the psychoacoustic model.
- From the detection procedure, a map, *noise_flag[sfb]*, is constructed such that noise-like scalefactor bands are flagged with a non-zero value.
- For each flagged scalefactor band, the energy of the corresponding spectral coefficients is calculated and mapped to a logarithmic representation with a resolution of 1.5 dB. An offset (*NOISE_OFFSET*=90) is added to the logarithmic noise energy values.
- For each flagged scalefactor band, the corresponding spectral coefficients are set to zero before quantization of the coefficients is carried out as usual.
- During the noiseless coding procedure, the pseudo codebook *NOISE_HCB* is set for all flagged scalefactor bands. Apart from this, the regular section / noiseless coding procedure is carried out on the quantized coefficient data.
- The logarithmic noise energy values are coded analogous to the regular scalefactors, i.e. with a differential encoding scheme starting with the *global_gain* value. They are transmitted in place of the scalefactors belonging to the flagged scalefactor bands.

3.3 Related Works

The related researches about perceptual noise substitution can be classified into two categories. One is detecting noise-like component in time domain while the other is detecting in the frequency domain. This section will discuss these two mechanisms and analyze the pros and cons of them.

3.3.1 Noise Detection in Time Domain

In time domain, the nonstationary signal components can be detected through various ways. For example, some audio developers have used a least-mean-square or a recursive-least-square transversal filter to detect the noise components [34]-[35]. Some use the zero-crossing rate to identify the noise components from the input signals [36]. Noise detection in time domain has less computing complexity than frequency domain since the original input signals are recorded in time domain. However, the PNS module in MPEG-4 AAC is operated in frequency domain and the PNS decoder generates pseudo random noise in frequency domain. Detecting noise components in time domain must have a mechanism to map the noise components into frequency bands (quantization bands). It's difficult to find a mapping mechanism with highly mapping accuracy and to determine the energy of the mapped frequency bands. Usually, noise detection in time domain decides only the index of start band and then uses PNS to substitute all bands higher than the start frequency band.

3.3.2 Noise Detection in Frequency Domain

In frequency domain, the spectral energies and phase relations of noise components are nonstationary over time. Based on this phenomenon, Serra [37] proposed a method to detect tonal components using a time-frequency transform and tracking stationary peaks in succeeding spectra. In this method, all signals are classified as noise components except the detected tonal signals. Furthermore, from the psychoacoustic perspective, the noise component has stronger masking ability and lower tonality than the tonal component. According to these two characteristics, PNS can detect noise bands directly in frequency domain based on the information calculated in psychoacoustic model [38]. Usually, noise detection in frequency domain has to use a time-frequency transform and increases the computing complexity. Since MPEG-4 AAC performs encoding in spectral domain mainly and PNS reconstructs signals in spectral domain, detecting PNS noise band in frequency domain increases only a little computing complexity and gains a high accuracy of detecting noise components.

Chapter 4 Proposed PNS Approaches

4.1 Signal Substitution before Quantization Process

As mentioned above, the perceptual noise substitution module can substitute the noise-like component in bands and reconstruct pseudo random noise with the same energy in the decoder end. According to the characteristic of noise, the perceived quality of noise-like component is mainly based on its energy, not the actual waveform. The perceived difference between the noise-like component that PNS substituted and the random noise reproduced in the decoder is small and can be neglected. Based on the motivation of PNS module, this thesis proposes a well-designed method to substitute the noise-like component including noise band detection, a frequency lower bound of PNS, and an adjustment in energy to remove the effects of tonal signals. Furthermore, this thesis extends the PNS module to retain the high frequency signals for the graceful bit allocation and to improve the overall quality. The design issues of retaining the high frequency signals are discussed in this section later including how to find a suitable start frequency at different bit rates and a limitation on bits that PNS uses to substitute the high frequency signals.

4.1.1 Noise Band Substitution

As illustrated in Figure 9, the noise-like component is substituted by PNS parameters and random noise is reconstructed in the decoder end. In the example, the original signals contain noise-like components from 9.5~22 kHz and the PNS substitutes these noise-like components at quantization band level. Since PNS recorded the frequency range and energy level of the substituted noise band, the perceived difference between the original signals and coded signals is very small. In this section, the design issues of proposed noise substitution method are described in detail and the effectiveness of the issues are illustrated with clear examples.

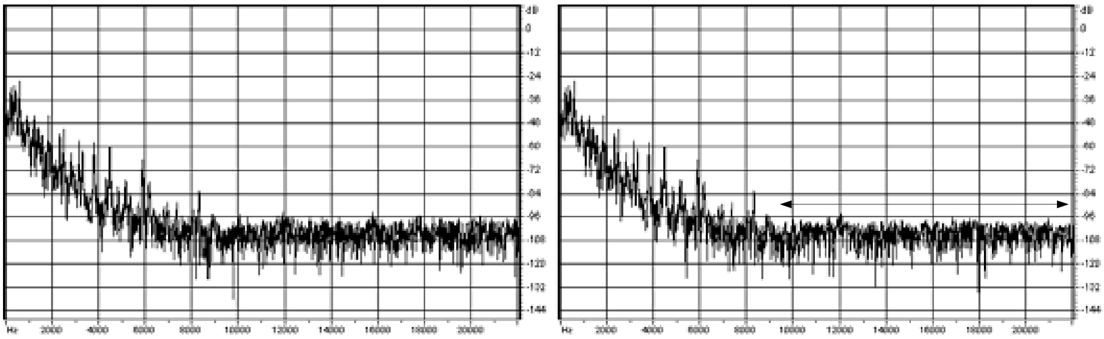


Figure 9: The original signal and the reconstructed signal that noise-like components are substituted

4.1.1.1 Noise Band Detection

The most important design issue of PNS is the noise band detection mechanism. Detecting noise component in frequency domain can rely on the characteristics of noise. The proposed noise band detection mechanism is based on the observation that the spectrum of noise component is much flatter than that of tone component. To detect this characteristic, a measurement of the spectral flatness must be consulted and compared to locate the noise band in the frequency domain. The flatness of the spectrum is defined by the following formula:

$$Flatness\ F_b = \frac{\sqrt[N]{\prod_{i=M(b)-N/2}^{M(b)+N/2} (X_i)^2}}{\frac{1}{N}\sum_{i=M(b)-N/2}^{M(b)+N/2} (X_i)^2} \tag{3}$$

where *M(b)* is the middle frequency bin of quantization band *b* and *N*=*max*(*L*,12). *L* is the length of quantization band *b*. The value of spectral flatness F_b calculated by formula (3) is located within the range of 0 to 1 since the mathematical law that the arithmetic mean is greater than or equal to the geometric mean. If a frequency band *b* consists of noise-like signals mainly, the flatness F_b will be close to 1 because the spectrum of noise is flat, and vice versa.

Figure 10 shows two different examples of a frequency band, one contains noise-like components and the other contains tone-like components. The difference between these two examples is only in the magnitude of the middle frequency bin while the other bins are the same. As shown in the examples, the spectral flatness measured by formula (3) reveals the effectiveness and accuracy of the proposed formula. The spectral flatness of the tone band is 0.001 since the presence of the tone in the middle frequency and the noise band gets a value of 0.923 in spectral flatness measurement.

Based on the spectral flatness measured above, the noise band detection mechanism can be defined by following equation:

$$is\_noise(b) = \begin{cases} true, if\ F_b \geq T_N \\ false, otherwise \end{cases} \tag{4}$$

where T_N is a threshold that implies a frequency band is noisy. The return value of function *is_noise(b)* indicates that the quantization band *b* is noisy or not. After extensive experiments on objective quality measurement and subjective listening test, this thesis suggests that the suitable value for T_N should be 0.85 for effective noise detection in the frequency domain.

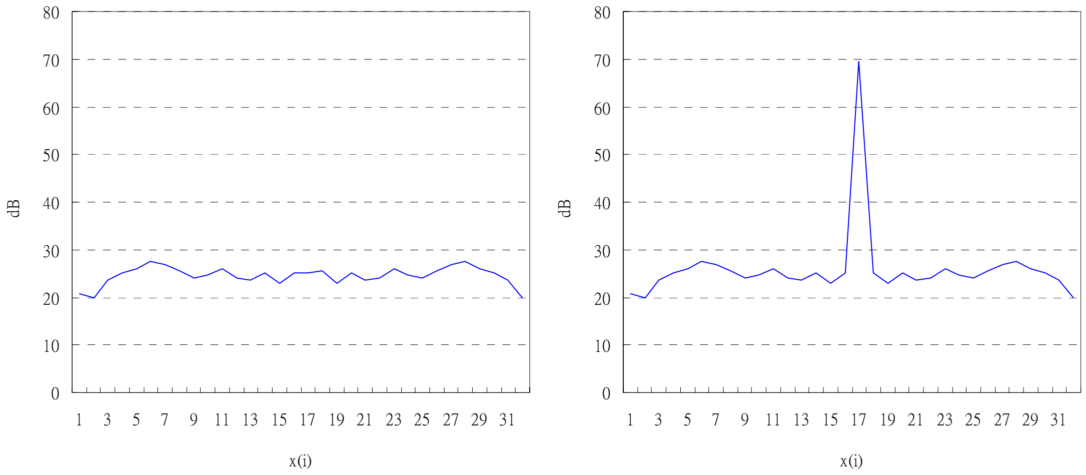


Figure 10: Examples of spectral flatness measurement in different frequency band types

4.1.1.2 Frequency Lower Bound

As mentioned in the MPEG-4 standard [33], there is a boundary of PNS in the frequency. The reason that perceptual noise substitution technique can function well is based on the phenomenon: the perceived difference between the original noisy signals and reconstructed signals is inaudible. In the low frequency, this phenomenon no longer exists since the finer resolution of human auditory system in this frequency region. The difference between the original signals and the reconstructed signals can be heard by the listener and the reconstructed signals become an outstanding noise in the perception. In the MPEG-4 standard, it suggests 4 kHz as the frequency lower bound of PNS. After extensive experiments on objective quality measurement and subjective listening test, this thesis adopts the value suggested in standard. However, the boundary of a PNS band may not match 4 kHz perfectly since the MPEG-4 AAC is performed in quantization band level. This thesis uses the nearest quantization band whose frequency boundary greater than or equal to 4 kHz. For example, the lower bound of PNS band is the 25th quantization band for long window at 44.1 kHz sample rate and the lower bound for short window is based on the group determination.

Flatness = 0.923

4.1.1.3 Noise Floor Estimation

The purpose of perceptual noise substitution is to substitute the noise-like components in encoder end and reproduce pseudo random noise in decoder end. However, the random noise generated in the decoder contains only noise-like components and such pure noise component appears not often in generic audio signals. Most generic noise-like components contain noisy signals mainly and tonal signals slightly. The energy of the band associated with PNS has to be adjusted according to the likeness of noisy signals to remove the effects of tonal signals. If PNS substituted the noise-like component that contains tonal signals slightly, directly accumulating the energy of each spectrum may be inappropriate in determining the energy of substituted signals. The calculated energy may be slightly affected by the tonal signals. Furthermore, for the other proposed approaches in the below sections the accumulated energy could be seriously influenced by the presence of the strong tonal signal and reproduce annoying noise in the decoder end.

This thesis proposes a method to reduce the effects of the tonal signals based on the spectral flatness when calculating the energy of substituted signals. The adjusted energy is called noise-floor energy and is defined by the following formula:

$$E'_b = F_b * E_b \tag{5}$$

where E'_b is the energy of noise floor, F_b is the spectral flatness of band b, and E_b is the summation of energy of each spectrum in band b. The modified energy E'_b will be similar to the noise floor of the substituted signals. Figure 11 illustrates the difference and importance of energy adjustment by two obvious examples. Without adjustment in energy, the floor of reconstructed noise is much higher than that in original signals. Such incorrect energy determination can induce annoying feeling in the coded signals for human perception. In contrast, with energy adjustment, the modified energy E'_b will be similar to the noise floor in the original signals and the effects of the tonal signal have been removed from the calculation of energy. In the following proposed PNS approaches, the energy of the band that PNS substituted always means to be the noise-floor energy. In the experiment part, this thesis proves the significance and importance of the noise floor estimation through both subjective listening test and objective quality measurement.

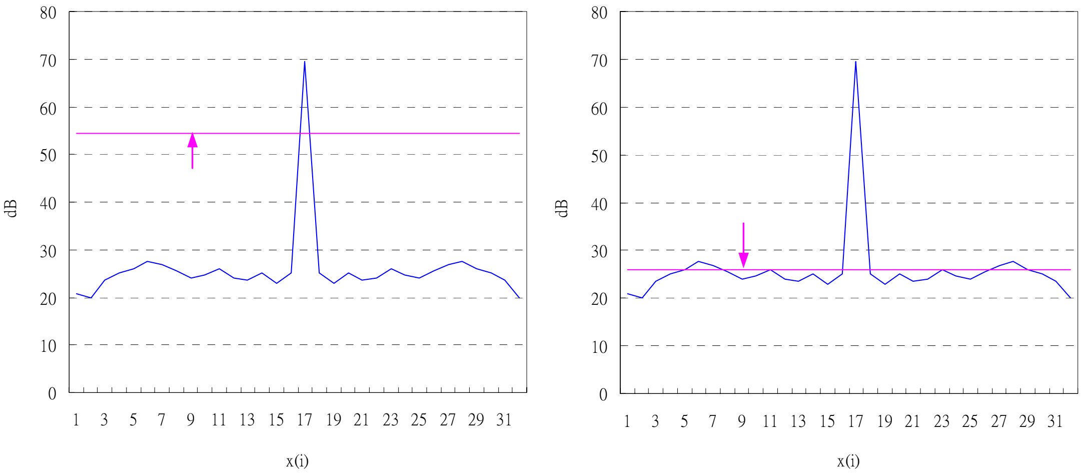
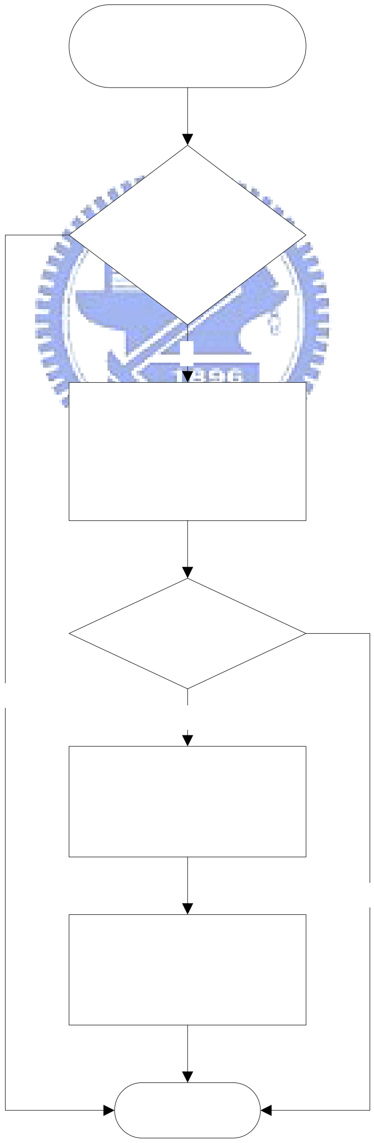


Figure 11: Examples of energy adjustment

Without En



The Floor of
Reconstructed Noise

Figure 12: The flowchart of noise band substitution approach

4.1.2 High Frequency Signal Substitution

The audio coding usually has the tradeoff between retaining all the frequency range and cutting off high frequency signals. As shown in Figure 13, a generic audio codec usually cuts off the signals above 16k Hz to maintain a better precision in low frequency range at 128kbps. This mechanism results in a less degree of loudness and clearness of the coded signals. However, in another example illustrated in Figure 13, the sparse bands will appear in high frequency range and the precision of coded signals in low frequency range will be lower if retaining all frequency range. This thesis shows that the tradeoff between retaining all frequency and cutting off high frequency bands needs to be reconsidered within perceptual noise substitution module. As illustrated in Figure 13, the signals in high frequency bands can be replaced by PNS information to protect better precision on the signals in low frequency range. Since the purpose of PNS is to substitute noise-like components, the bands with high tonality must be avoided during the high frequency signal replacement process to prevent great distortions in substituted bands. Furthermore, a limitation on the bit usage for PNS to replace high frequency signals must be conducted at low bit rates. At low bit rates, the PNS should not spend too many bits on retaining the high frequency signals and cause degradation in low frequency range since the bit rate is already low. In this section, the design issues of proposed high frequency signals substitution approach will be brought up and discussed in detail.

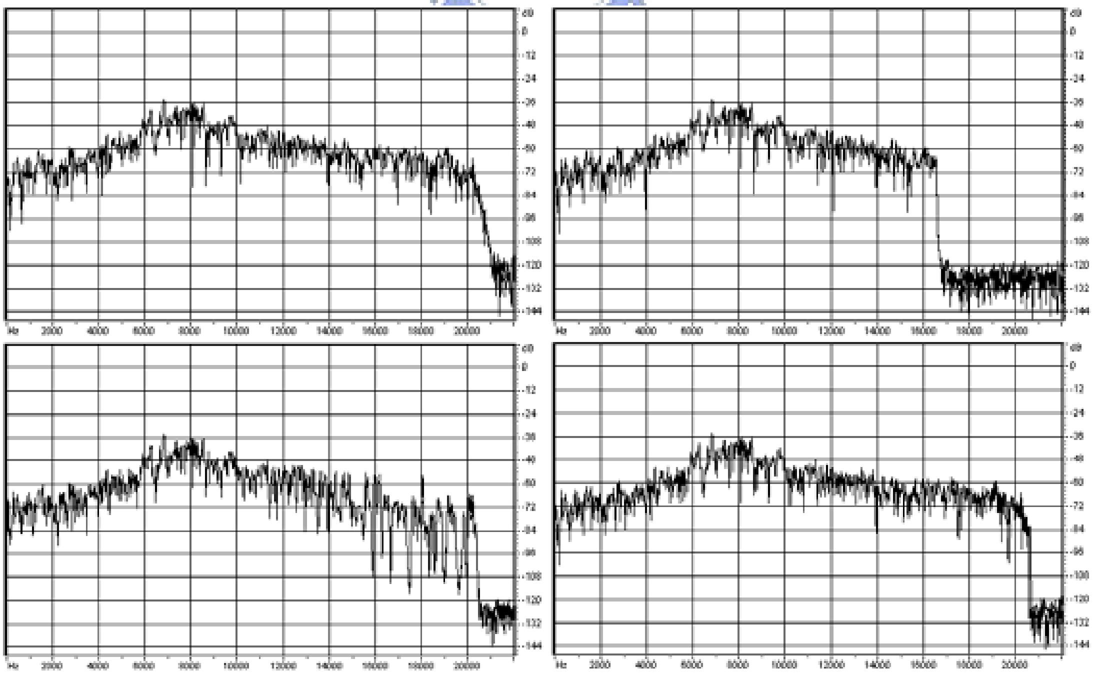


Figure 13: Example of high frequency signal substitution by PNS

4.1.2.1 Start Frequency Decision and Tonal Band Avoidance

Retaining high frequency signals by means of perceptual noise substitution can reserve the loudness of the original signals, avoid sparse bands in high frequency and gain a better precision on the signals in low frequency. The start frequency of high frequency signal substitution is according to the location that sparse band appears and the appearance of the sparse band is mainly based on the encoding bit rate. As illustrated in Figure 13, the suitable start frequency of high frequency signal substitution in the example is about 16k Hz since the sparse bands appeared in the range of 16k~20k Hz if retaining all frequency range at 128kbps. Table 2 shows the suitable start frequency proposed in this thesis at different bit rates. The suggested start frequency is conducted by the objective quality measurement in MPEG 12 tracks at different bit rates.

Table 2: Suggested start frequency of high frequency substitution by PNS

Bit rate (kbps)	128	112	96	80	64
Start frequency (Hz)	17875	16500	15125	13750	12375

In MPEG-4 Advanced Audio Coding (AAC), the bit rate can be arbitrarily specified. For the bit rate which is not belonged to any entry in Table 2, the start frequency can be determined by interpolation of the nearby entries in the table. Formula (6) is a mathematical expression of the method.

$$F_{start}(B) = S_i + \frac{S_{i+1} - S_i}{B_{i+1} - B_i} * (B - B_i), if\ B_i \le B < B_{i+1} \qquad (6)$$

where *B* is the value of specified bit rate, $F_{start}(B)$ is the calculated start frequency at bit rate B, B_i and B_{i+1} are the nearby bit rates found in Table 2, S_i and S_{i+1} are the corresponding start frequencies of bit rate B_i and B_{i+1}.

Figure 14 shows the signals that contain highly tonal components and the reconstructed signals that the high frequency signals are substituted by PNS parameters. Obviously, the tonal signals in high frequency are disappeared in the reconstructed signals. The human perception of a highly tonal band is mainly stimulated by the high tone, not the other signals. If enabling PNS module in a highly tonal band, the distortion between the original signals and reconstructed random noise can be perceived obviously even the substituted energy was adjusted to the noise floor. Therefore, for the signals which contain plenty of tonal components, the PNS module has to avoid being active in such signals. This thesis proposes a method to detect highly tonal bands and avoid activating PNS in these bands based on the spectral flatness measured in equation (3).

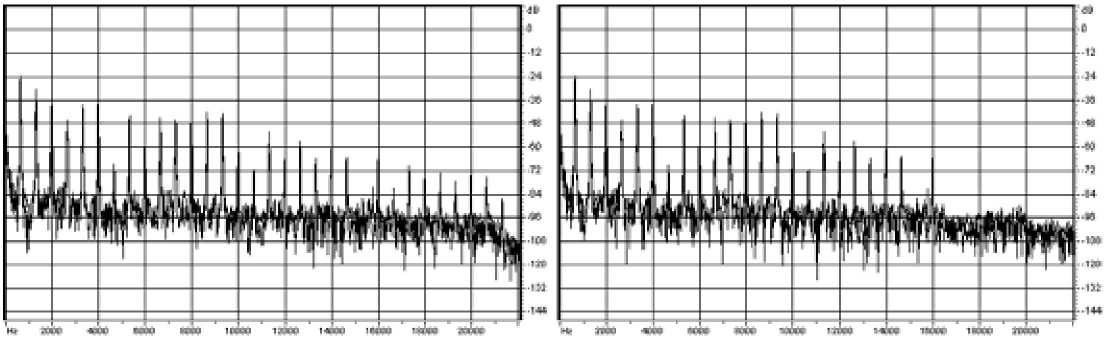


Figure 14: Example of tonal band avoidance in high frequency signal substitution

4.1.2.2 Limitation on Bit Usage

Retaining high frequency signals by perceptual noise substitution dose not always save bits. It depends on the bit usage of the band that PNS substituted. If the band that PNS substituted was allocated some bits, then PNS can save bits in this band since the cost of PNS is only a few bits. On the other hand, if the band was not allocated any bits, then PNS cannot save bits but spends costs in this band. Therefore, at low bit rates, using PNS technique to retain high frequency signals may lead to obvious degradation in low frequency range since the high frequency signals were not allocated any bit if PNS did not substitute them.

Figure 15 illustrates an example to reveal the difference. Without limitation on bit, the signals in low frequency range are sparser than the one with limitation on bit. Such distortion in low frequency can degrade the quality significantly although it retained the signals in higher frequency range. This thesis proposes a method to prevent such situation based on the bits allocated to the frame. The bits used for PNS to substitute high frequency signals will not exceed a ratio R of the available bits in a frame. After extensive experiments on objective quality measurement and subjective listening test, this thesis suggests the suitable value for R to be 15%. With such limitation on bit, the qualities of coded signals in low frequency range will not be influenced by retaining high frequency signals via perceptual noise substitution technique.

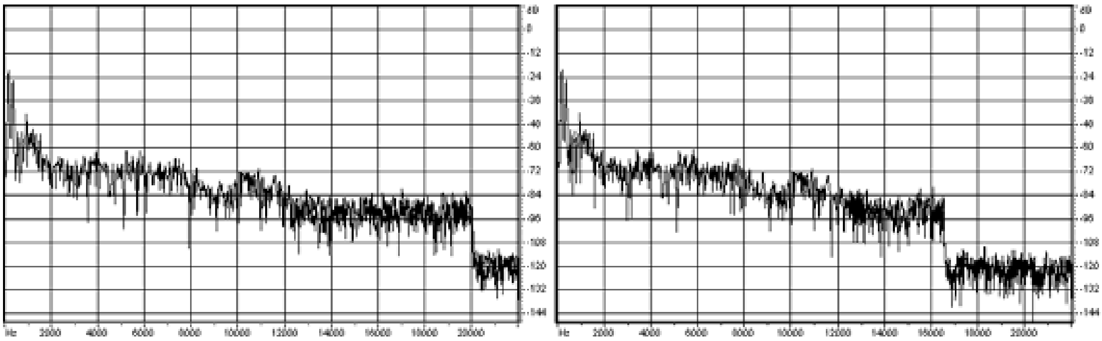


Figure 15: Limitation on bit usage in high frequency signal substitution

High frequency s
without bit li

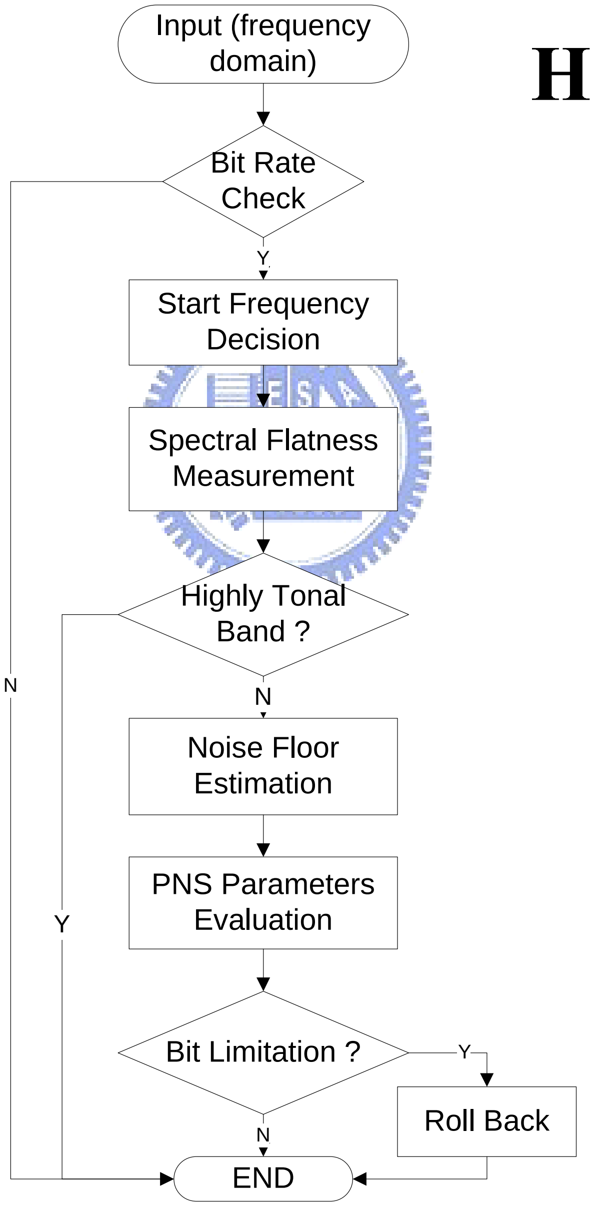


Figure 16: The flowchart of high frequency signal substitution

4.2 Zero-Band Dithering after Quantization Process

In general audio coding, the zero-bands are often caused by lacking of bits and will cause the birdie artifacts [39] for human auditory system. However, if zero-bands can be transformed into PNS bands properly, the birdie artifacts will be significantly reduced and hence improve the listening quality greatly, especially at low bit rates. As shown in Figure 17, the dithered spectrum is much similar to the original one and most of the zero-bands are replaced into bands which are consisted of random noise. The human perception stimulated by dithered signals is much better than the one without dithering since the birdie artifacts are eliminated. In the encoder end, zero-bands can be detected once the quantized values are available and the quantized values are available when quantization process is done. Activating perceptual noise substitution module after the quantization process can transform the zero-bands into PNS bands with a little increasing in complexity. Since the cost of PNS is not free, the bits used by PNS parameters will be charged from the bits saved in bit reservoir to fit the desired bit rate. Furthermore, dithering zero-bands by perceptual noise substitution technique can lead to inconsistency in the bandwidths of different audio channels if substituting zero-bands in a channel-by-channel manner. This thesis proposes a well-designed method to transform zero-bands into PNS bands to eliminating artifacts and improve the quality of coded signals.

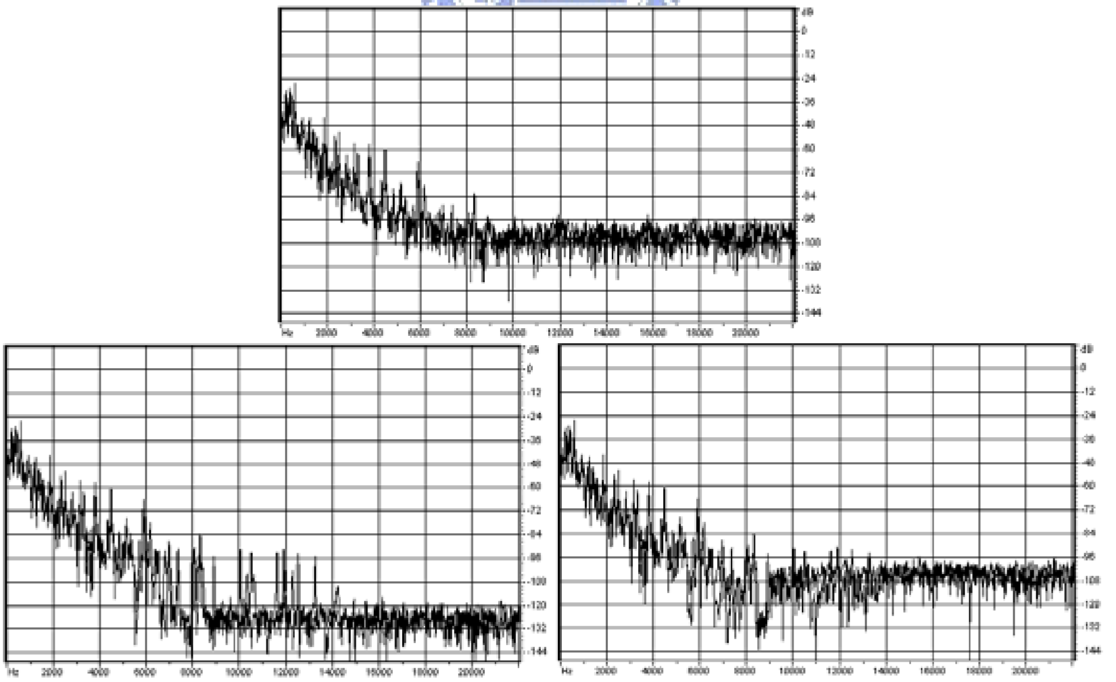


Figure 17: Example of zero-band dithering by PNS

4.2.1 Zero-Band

In most perceptual audio codecs, the non-uniform quantizer is introduced to handle the weight of distortion efficiently. An overall spectrum of a time frame is separated into several quantization bands with non-uniform bandwidths. Every quantization band has its own quantization step size Δ_q to fit different perceptually tolerable distortion allowed by psychoacoustic model. In MPEG-4 AAC standard [33], the non-uniform quantization model is given as follow:

$$S[k] = \operatorname{int}\left(\frac{X[k]^{\frac{4}{3}}}{\Delta_q} \right) \tag{7}$$

where *x[k]* is a frequency line, *s[k]* is the quantized value, and Δ_q is the quantization step size for quantization band *q*. Furthermore, the quantization step size is Δ_q defined as the following formula.

$$\Delta_q = 2^{\frac{3}{16}\cdot(g - s_q)} \tag{8}$$

where *g* is the global gain used for all quantization bands, and S_q is the scalefactor for quantization band *q*.

In the decoder, the coded frequency signal *x[k]* will be inversely quantized as $\tilde{X}[k]$ by the formula (9).

$$\tilde{X}[k] = \left(S[k] \cdot \Delta_q\right)^{\frac{4}{3}} \tag{9}$$

That is equivalent to (10) where $\tilde{\Delta}_q$ is defined as $\Delta_q^{\frac{4}{3}}$.

$$\tilde{X}[k] = S[k]^{\frac{4}{3}} \cdot \tilde{\Delta}_q \tag{10}$$

In fact, the original *x[k]* value should be given as:

$$X[k] = R[k]^{\frac{4}{3}} \cdot \tilde{\Delta}_q \tag{11}$$

where *R[k]* is a real number, and there is a relation between *R[k]* and *s[k]* as follows:

$$S[k] = \operatorname{int}(R[k]) \tag{12}$$

From the definition of zero-bands, the requantized frequency signal $\tilde{X}[k]$ in a zero-band must be zero. From (10), it implies that the relative *s[k]* must be also set to zero. Hence, from (12), it shows that $\left|R[k]\right|$ should be less than 1/2. Substituting the result to (11) illustrates the occurring of zero-bands is due to the following relation

$$\left|X[k]\right| < \left(\frac{1}{2}\right)^{\frac{4}{3}} \cdot \tilde{\Delta}_q \tag{13}$$

4.2.2 Cooperation with Bit Reservoir

Although transforming zero-bands into PNS bands after quantization process can eliminate the artifacts significantly, the cost of the transformation is not free. Since the expenditure of PNS parameters, PNS bands consume more bits than zero-bands after the transformation. To fit the desired bit rate precisely, the cost of PNS parameters must be charged. However, the cost of PNS cannot be determined in the quantization stage since PNS dithers zero-bands after quantization process.

In most popular audio codecs, like mp3 or aac, a tool called bit reservoir is often involved in the design of the codecs. The bit reservoir is a tool to control the allocation of bits among frames. The bits unused in the previous frames can be deposited to bit reservoir and bit reservoir can allocate the bits to the frames with demand to enhance the quality. Since the bits saved in bit reservoir can exist over frames, the cost of PNS parameters can be charged from bit reservoir. However, the bits saved in bit reservoir are varying from one frame to another. The saved bits may not be enough for PNS to dither all zero-bands in a frame. The PNS must dither zero-bands in a band-by-band manner and the cost of PNS must be less than or equal to the bits saved in bit reservoir. According the design of this method, the bits saved in bit reservoir limit the improvement of PNS to dither zero-bands. Figure 18 shows a clear example of zero-band dithering cooperating with bit reservoir.

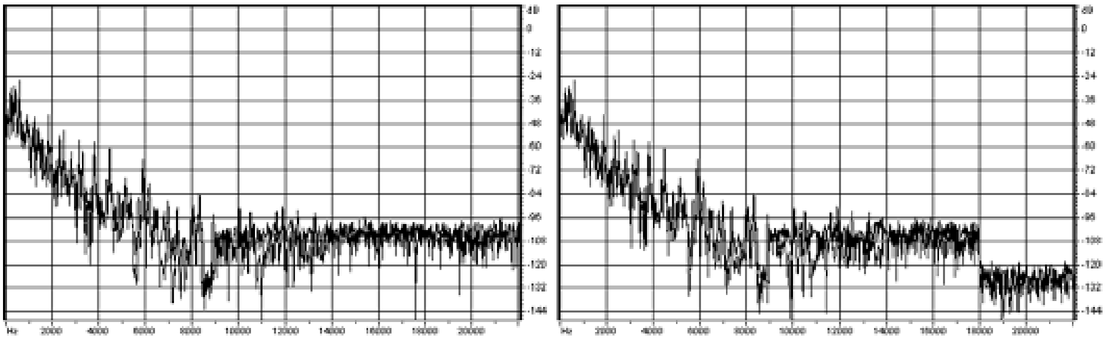


Figure 18: Example of zero-band dithering with/without bit limitation

4.2.3 Dithering All Channels to the Same Frequency Band

In generic audio coding, the audio signals are coded in a channel-by-channel manner. For example, in the audio signals that contains two channels. For a coding frame, the left channel can be encoded first and then the right channel is encoded. However, the proposed zero-band dithering method can cause the inconsistent bandwidths of different channels if PNS dithered in a channel-by-channel manner. The inconsistency on the bandwidths of different channels is caused by the bit reservoir. The amount of bits saved in bit reservoir is usually updated in frame level. If PNS dithers zero-bands in a channel-by-channel manner, the bits saved in bit reservoir will be allocated to the channel which is encoded first. Then, the bits left in the bit reservoir after dithering first channel can be allocated to the second channel. If the amount of bits saved in bit reservoir is not enough for PNS to dither all the zero-bands in all channels, the bandwidths of different channels will be inconsistent. Figure 19 illustrates an obvious example. The left channel was dithered first and then the right channel is dithered. Since the amount of bits saved in bit reservoir is limited, the zero-bands in right channel are not all dithered and cause the inconsistency on the bandwidths of left and right channels.

To eliminate this inconsistency, PNS has to dither zero-bands in a band-by-band manner. For quantization band *b*, both left and right channels must be dithered by PNS at the same pass and check the bits used after dithering both channels. If the amount of bits saved in bit reservoir is not enough to dither both channels, then both channels are not dithered and PNS stop dithering zero-bands in this frame. With such a little modification, the proposed PNS approach can eliminate zero-bands and keep the bandwidths of different channels consistent.

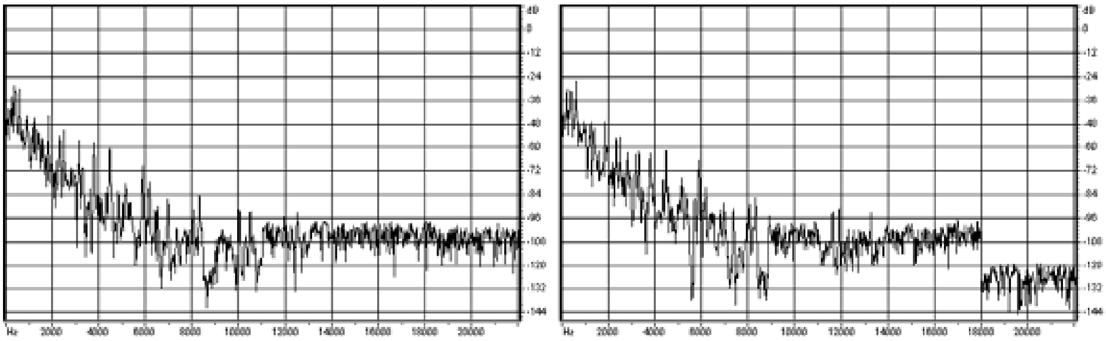


Figure 19: The inconsistency on the bandwidths of different channels

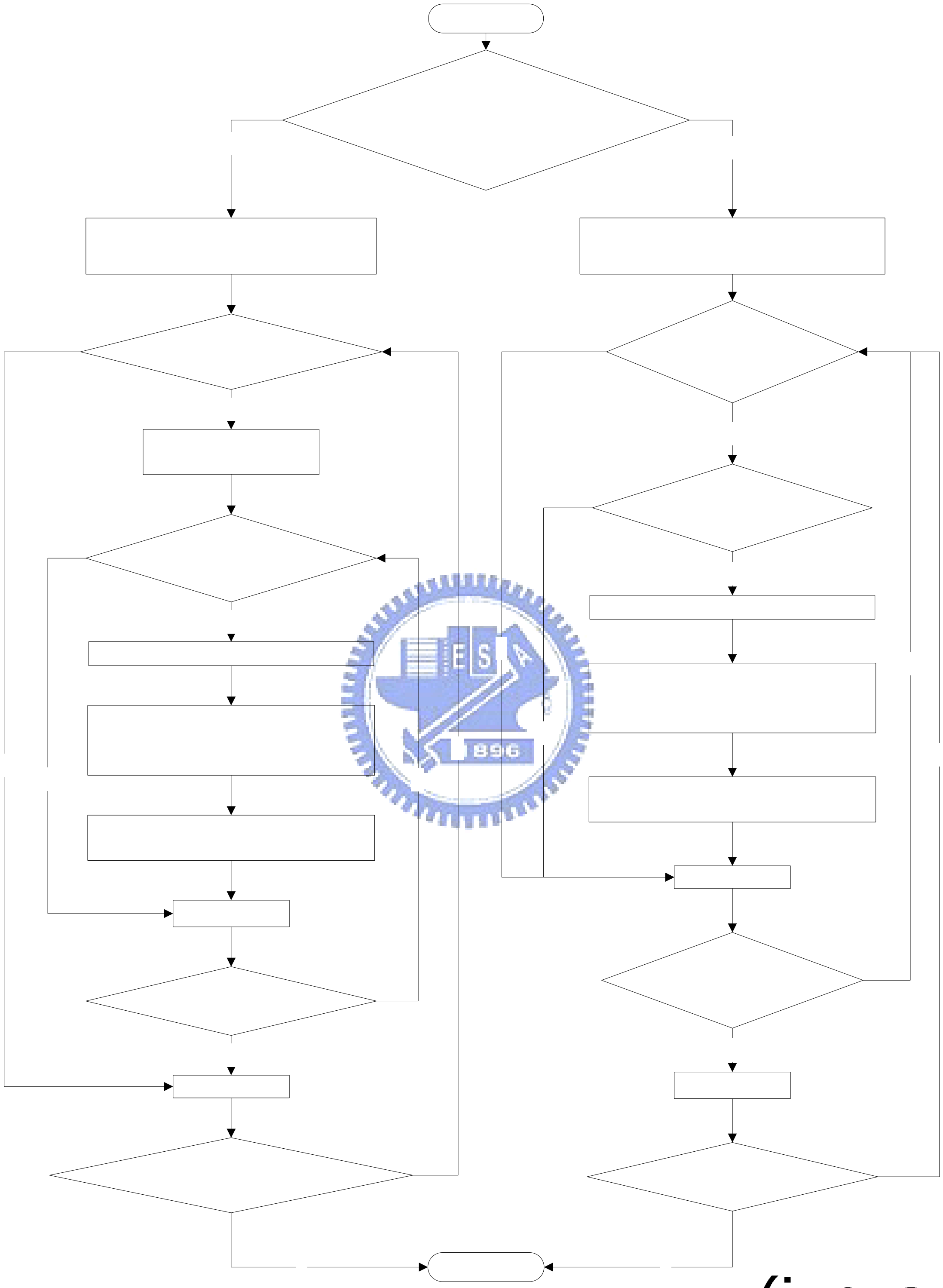


Figure 20: The flowchart of proposed zero-band dithering approach

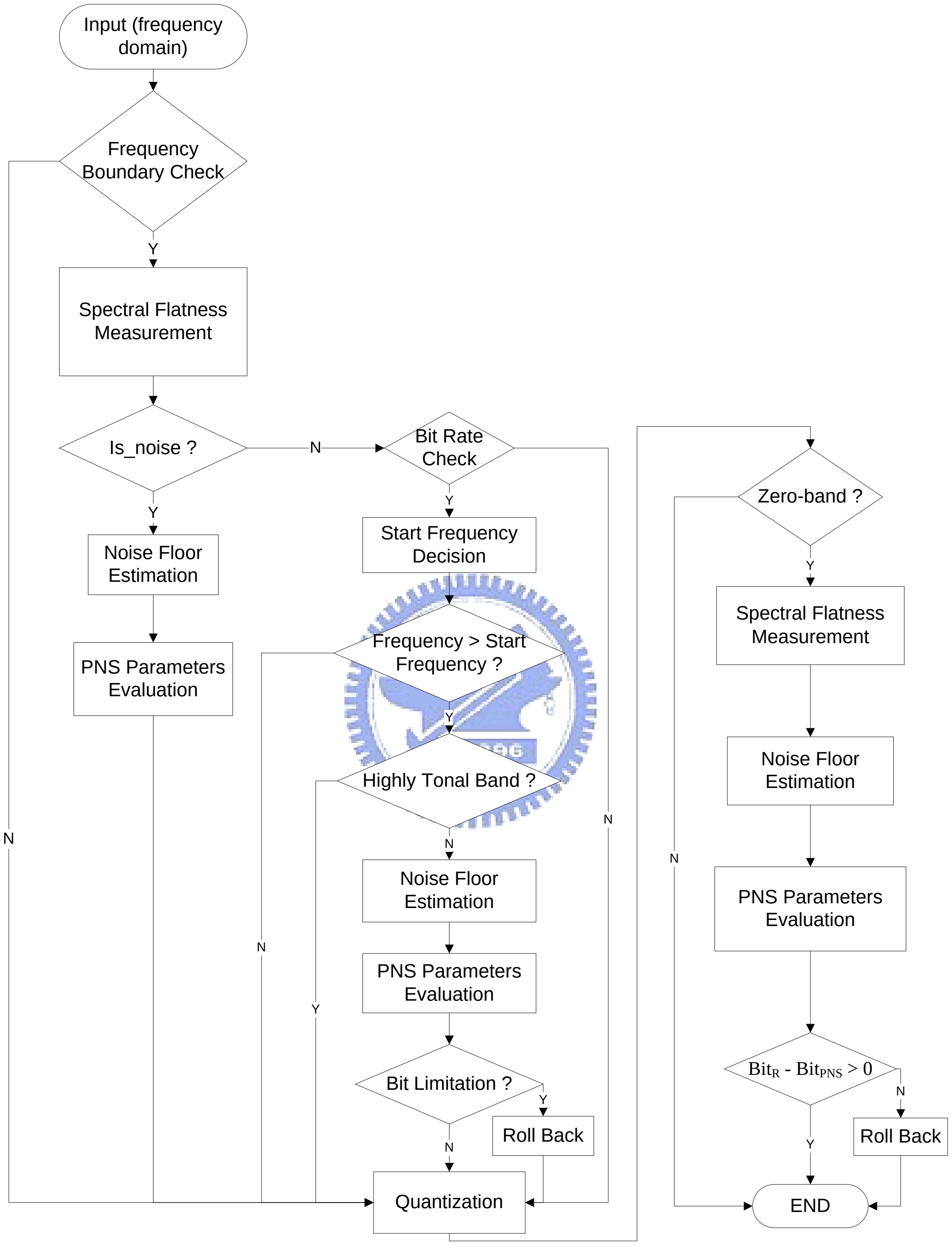


Figure 21: The flowchart of combined approach: noise substitution, high frequency substitution, and zero0band dithering by PNS

Chapter 5 Enhanced Psychoacoustic Model and M/S Coding in Advanced Audio Coding

In this chapter, we extend our previous works to enhance the performance of our AAC codec “NCTU-AAC” [40]. The NCTU-AAC codec is a MPEG 2/4 advanced audio coding encoder developed in our laboratory [41] based on the theories that we have proposed in [42]-[54]. In the following sections, this thesis proposes the enhancing mechanisms in psychoacoustic model and M/S coding based on our previous works [46] [55].

5.1 Adaptive Masking Offset in Psychoacoustic Model

As discussed in chapter 2, simultaneous masking effects can be classified into four types: noise-masking-noise (NMN), noise-masking-tone (NMT), tone-masking-tone (TMT), and tone-masking-noise (TMN). Usually, in general audio coding, only NMT and TMN are involved in the design of psychoacoustic model and the values of NMT and TMN are not always constant. In fact, the values of TMN and NMT are varying from codec to codec. In the MPEG-4 AAC standard [33], it defines NMT(b) = 6dB and TMN(b) = 18dB for all band b. In the MPEG-1 Layer III [3], it defines NMT(b) = 6dB and TMN(b) = 29dB for all partition band b. However, according to the nature of the masker, the offsets of different bands should be different values. The human auditory system is less sensitive to the signals in high frequency and the masking effect of signals in high frequency is stronger than low frequency. In our previous work of psychoacoustic model [55], the masking offsets in tone-masking-noise and noise-masking-tone are fixed for all bands, just like the MPEG standard [33]. Keeping the offsets of all bands being the same can not reveal the difference of masking effect between low and high frequency. Therefore, adaptive masking offsets in TMN and NMT need to be consulted to improve the performance of the codec.

As mentioned in [56], fine-tuning the offsets of all bands can optimize the coder output when developing and designing the codec. Based on the characteristics of human auditory system, the offsets of bands in low frequency range should be higher than the offsets in high frequency range since the difference in masking ability. This thesis tunes the two offsets, TMN and NMT, to optimize the output performance of the codec. In the fine-tuning stage, the offsets of all bands are tuned based on the result of objective quality measurement in the twelve tracks recommended by MPEG. The fine-tuned offsets and the difference between the fixed offsets and the fine-tuned offsets are shown bellow.

Table 3: Adaptive tone-masking-noise offset

Bands	TMN (dB)	Adaptive TMN (dB)	Band	TMN (dB)	Adaptive TMN (dB)
1	18	28.16	26	18	18.79
2	18	27.03	27	18	18.56
3	18	25.9	28	18	18.56
4	18	24.77	29	18	18.34
5	18	23.98	30	18	18.23
6	18	23.42	31	18	18.11
7	18	22.74	32	18	18
8	18	22.29	33	18	18
9	18	21.84	34	18	17.89
10	18	21.39	35	18	17.89
11	18	21.05	36	18	17.77
12	18	20.82	37	18	17.66
13	18	20.6	38	18	17.55
14	18	20.48	39	18	17.44
15	18	20.37	40	18	17.32
16	18	20.26	41	18	17.21
17	18	20.14	42	18	17.1
18	18	20.03	43	18	16.98
19	18	19.92	44	18	16.87
20	18	19.69	45	18	16.76
21	18	19.47	46	18	16.65
22	18	19.35	47	18	16.53
23	18	19.13	48	18	16.42
24	18	19.13	49	18	16.31
25	18	18.9			

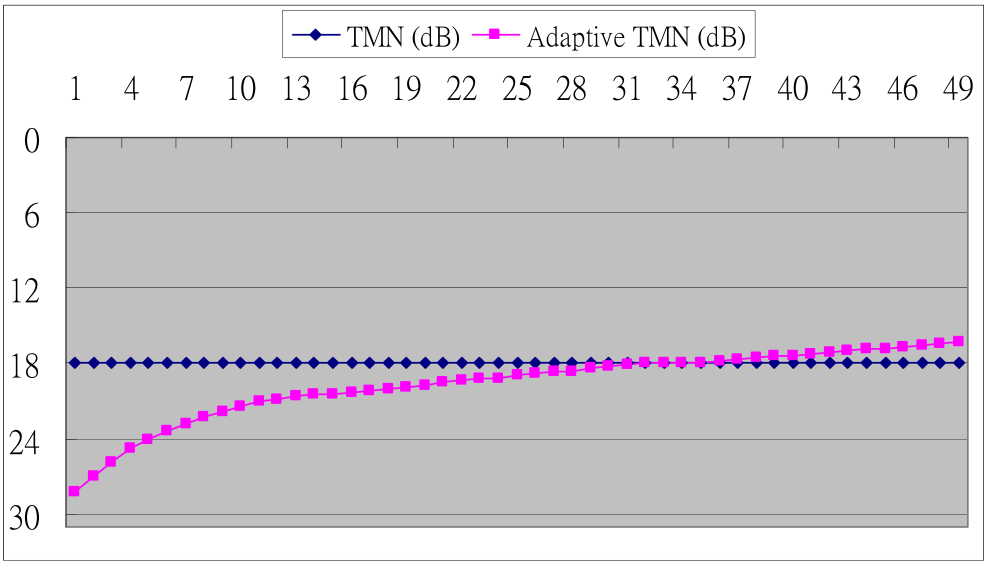


Figure 22: The difference between fixed TMN and proposed TMN

Table 4: Adaptive noise-masking-tone offset

Bands	NMT (dB)	Adaptive NMT (dB)	Band	NMT (dB)	Adaptive NMT (dB)
1	6	16.16	26	6	6.79
2	6	15.03	27	6	6.56
3	6	13.9	28	6	6.56
4	6	12.77	29	6	6.34
5	6	11.98	30	6	6.23
6	6	11.42	31	6	6.11
7	6	10.74	32	6	6
8	6	10.29	33	6	6
9	6	9.84	34	6	5.89
10	6	9.39	35	6	5.89
11	6	9.05	36	6	5.77
12	6	8.82	37	6	5.66
13	6	8.6	38	6	5.55
14	6	8.48	39	6	5.44
15	6	8.37	40	6	5.32
16	6	8.26	41	6	5.21
17	6	8.14	42	6	5.1
18	6	8.03	43	6	4.98
19	6	7.92	44	6	4.87
20	6	7.69	45	6	4.76
21	6	7.47	46	6	4.65
22	6	7.35	47	6	4.53
23	6	7.13	48	6	4.42
24	6	7.13	49	6	4.31
25	6	6.9			

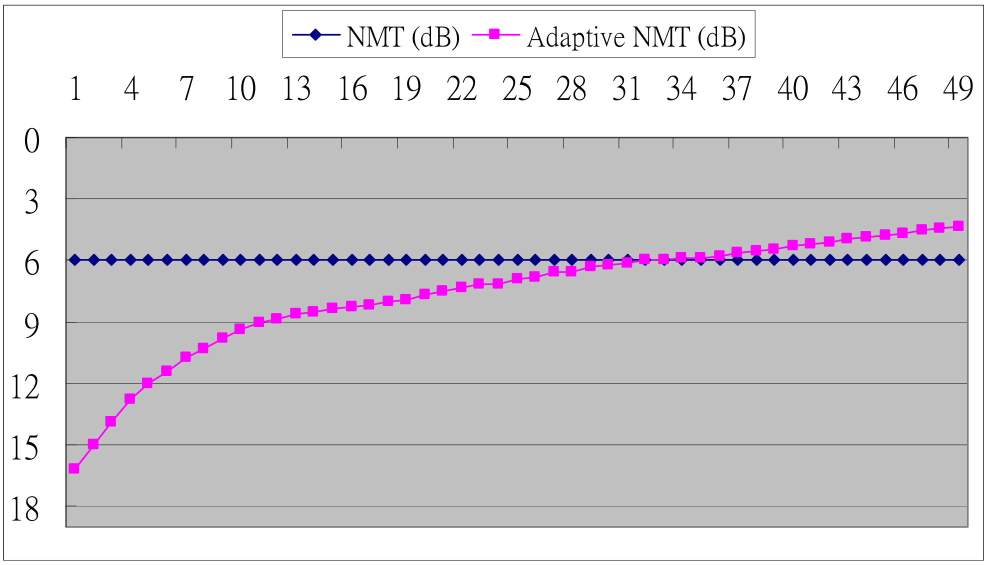


Figure 23: The difference between fixed NMT and proposed NMT

5.2 Enhanced M/S Coding in AAC

M/S coding [57]-[58] is an extended perceptual audio coding which transforms L/R signals into M/S signals. Based on the transformation, M/S coding provides an effective method to remove the irrelevant and redundant information from stereo channels. In MPEG-4 AAC [33], M/S coding is derived from the L/R pair signals in the frequency domain. For each band, the M/S signals can be represented as follows:

$$M_i[k] = \frac{L_i[k] + R_i[k]}{2} \tag{14}$$

and

$$S_i[k] = \frac{L_i[k] - R_i[k]}{2} \tag{15}$$

where $L_i[k]$, $R_i[k]$, $M_i[k]$ and $S_i[k]$ are the frequency lines of each coding state in quantization band *i*. In the decoder end, the L/R signals are reconstructed by the following formulas:

$$L'_i[k] = M'_i[k] + S'_i[k] \tag{16}$$

and

$$R'_i[k] = M'_i[k] - S'_i[k] \tag{17}$$

where $L'_i[k]$ and $R'_i[k]$ are the reconstructed spectral lines from the $M'_i[k]$ and $S'_i[k]$ of quantization band *i*. Such alternative representation of signals leads to the design issues of M/S coding including M/S band decision, psychoacoustic model for the M/S bands, bits allocated between two channels which contain L/R and M/S signals, and bit allocation in the channel pair. In this section, this thesis proposes an enhancing mechanism on the inter channel bit allocation based on our previous work in M/S coding [46].

Based on our previous work in M/S coding, the bits of M and S channels are allocated according to the allocation entropy (AE) of each channel. The allocation entropy is different from the perceptual entropy [59]-[60] which indicates the minimum number of bits required for transparent coding quality. The perceptual entropy does not reflect the weights of signals in different frequency when the amount of bits is limited. The allocation entropy can reflect the weights of signals in different frequency when transparent coding quality can not be achieved. The derivation of allocation entropy is based on the bandwidth proportional noise-shaping criterion [45].

The definition of allocation entropy is defined in formula (18).

$$AE_i = W_i * \log(R_i + 1) \tag{18}$$

and

$$R_i = \begin{cases} \dfrac{E_i}{T_i * B_i} & if\ (E_i \geq T_i * B_i) \\ 0 & if\ (E_i < T_i * B_i) \end{cases} \tag{19}$$

where *i* is the band index, W_i, E_i, T_i, and B_i are the four attributes of band *i*: bandwidth, energy, masking, and the effective bandwidth. The effective bandwidth is derived from the critical band with quarter critical bandwidth.

After introducing the allocation entropy criterion, the bit allocation mechanism between M and S channels is proposed as follows:

$$Bit_M = \frac{AE_M}{AE_M + AE_S} B \tag{20}$$

and

$$Bit_S = \frac{AE_S}{AE_M + AE_S} B \tag{21}$$

where *B* is the total available bits of a frame, Bit_M and Bit_S represent the amount of bits allocated to M and S channels, AE_M and AE_S represent the allocation entropies of M and S channels. Based on such method, the available bits can be allocated to M and S channels according to their allocation entropies to achieve a better coding quality. However, after extensive tracks tests, the bit allocation mechanism above can not function well in the tracks which have middle similarity between L and R channels. In these tracks, the amount of bits allocated to S channel is not enough to preserve the difference between L and R channels and it causes an obvious artifact in the coded signals. The difference between L and R channels in the track with middle channel similarity is more important than that in the track with high channel similarity. Allocating bits to S channel based on the ratio of allocation entropy can not reflect the importance of the content in S channel and will cause serious distortion in the coded signals. As shown in Figure 24, the difference between L and R channels is lost in frequency range 10k~16k Hz in the coded signals. In the remarked region, the signals in L and R channels are almost the same and such distortion can cause an obvious artifact for human perception.

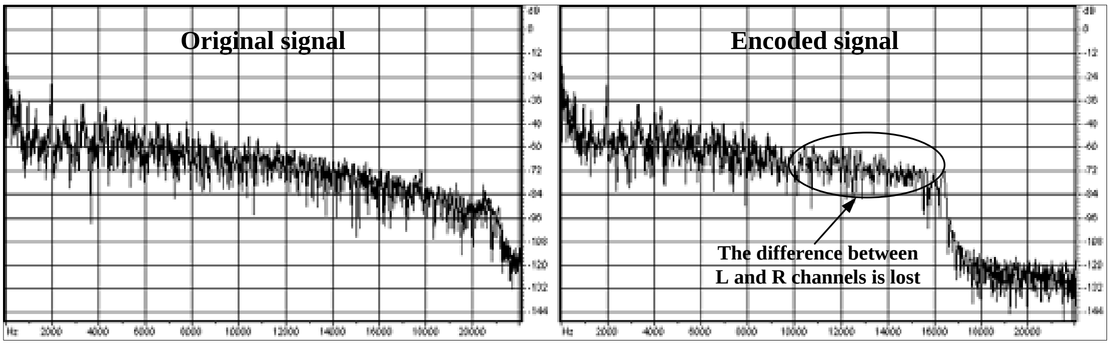


Figure 24: The artifact caused by AE-based inter channel bit allocation in M/S coding

To eliminate the artifacts and enhance the coder output, this thesis proposes a conservative method based on our previous work first. The bits allocated to M and S channels are determined by the following formula.

$$Bit_M = \left\{ \frac{1}{4} B + \frac{AE_M}{2(AE_M + AE_S)} B \right. \tag{22}$$

and

$$Bit_S = \left\{ \frac{1}{4} B + \frac{AE_S}{2(AE_M + AE_S)} B \right. \tag{23}$$

The conservative method reserves one half of total available bits *B* and distributes them to M and S channels equally. The bits allocated to M and S channel are be guaranteed to be 0.25*B* at least. Such conservative method can really reduce the artifact that caused by lacking of bits in S channel. However, such method allocates too many bits to S channel when the similarity between L and R channels is high and causes a poorer performance than the old method in such kind of tracks. See the experiment result shown in Table 5. The conservative method gets poorer ODG in es01, es02, and es03. The signals contained in L and R channels in these three tracks are similar and the content in S channel is close to be empty after M/S transformation. In the conservative method, the bits allocated to S channel are too many to coding the signals in S channel and cause degradation in ODG in these tracks. Therefore, an adaptive method must be conducted to combine the advantages of the old method and conservative method and avoid the disadvantages occurred in above two methods.

Table 5: ODG of conservative method in MPEG 12 tracks

Codec	NCTU-AAC	
Bit Rate	128 kbps	
Tracks	Original M/S	Conservative method
es01	-0.23	-0.34
es02	-0.05	-0.18
es03	-0.09	-0.26
sc01	-0.43	-0.42
sc02	-0.72	-0.73
sc03	-0.59	-0.59
si01	-0.51	-0.5
si02	-0.62	-0.63
si03	-0.93	-0.92
sm01	-0.55	-0.54
sm02	-0.43	-0.44
sm03	-0.74	-0.74
Average	-0.491	-0.524

Before discussing the adaptive method, a measurement of the similarity between L and R channels must be defined. The similarity is defined as follows:

$$Similarity\ F = \frac{AE_M - AE_S}{AE_M + AE_S} \tag{24}$$

where AE_M and AE_S represent the allocation entropies of M and S channels. If the signals contained in L and R channels are almost the same, then AE_S will be close to zero and the similarity F will be close one. If the signals contained in L and R channels are unlike, then AE_M will be close to AE_S and the similarity F will be close zero. Based on the similarity measured above, this thesis proposes an adaptive mechanism for inter channel bit allocation in M/S coding. The amounts of bits allocated to M and S channels are defined by the following formula (25)-(26).

$$Bit_M = \begin{cases} \dfrac{AE_M}{AE_M + AE_S} B & , if\ \ F \geq 0.8 \\ \dfrac{1}{4} B + \dfrac{AE_M}{2(AE_M + AE_S)} B & , otherwise \end{cases} \tag{25}$$

and

$$Bit_S = \begin{cases} \dfrac{AE_S}{AE_M + AE_S} B & , if \;\; F \geq 0.8 \\ \dfrac{1}{4} B + \dfrac{AE_S}{2(AE_M + AE_S)} B & , otherwise \end{cases} \tag{26}$$

For the signals with middle channel similarity, the adaptive method reserves half of total available bits *B* and allocates it to M and S channels equally. This mechanism guarantees that each channel has at least one-fourth of total available bits allocated to the frame. Moreover, for the signals with high channel similarity, the adaptive method directly allocates bits according to the ratio of allocation entropies of each channel. Such adaptive mechanism can eliminate the artifacts and preserve the advantages of our previous work in M/S coding. Figure 25 shows the difference between the previous method and the proposed method.

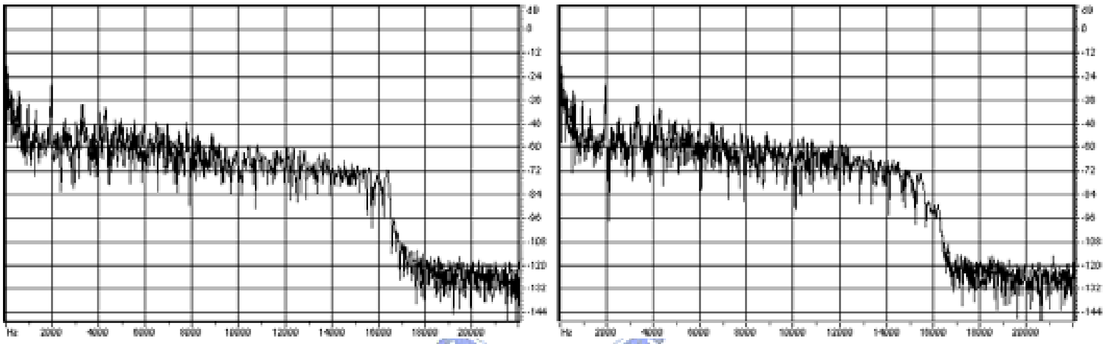


Figure 25: The difference between our previous work and the proposed adaptive method

Chapter 6 Experiments

In this chapter, extensive experiments are made to prove the enhancement of proposed methods based on the MPEG test tracks and the music database collected in our lab. After the extensive experiments, the computing complexity and bit usage of proposed methods are analyzed to reveal to cost of the method.

6.1 Experiment Environment

Computer Status:

Platform	Personal Computer
Operating System	Windows XP with SP2
CPU	Intel Pentium 4 2.4GHz
Memory	256MB DDR400 * 2
Mother Board	ASUS P4P800
Sound Card	ADI AD1985 AC' 97
Headphone	ALESSANDRO MUSIC SERIES PRO

Objective Quality Measurement Tool:

In the objective test, we choose EAQUAL [61] as the tool to assess the audio quality. EAQUAL stands for Evaluation of Audio Quality. The intention of EAQUAL is to provide an objective quality measurement for coded/decoded audio signals especially useful for audio codec development. The implementation of EAQUAL is based on the ITU-R recommendation BS.1387 [28].

Subjective Quality Measurement Tool:

In subjective quality test, we choose the tool called “MUSHRA” [62] to assist the assessment. Multi stimulus test with hidden reference and anchors (MUSHRA) has been designed to give a reliable and repeatable measure of the audio quality of intermediate-quality signals. MUSHRA has the advantage that it provides an absolute measure of the audio quality of a codec which can be compared directly with the reference. MUSHRA follows the test method and impairment scale recommended by ITU-R BS.1116 [15].

6.2 Objective Quality Measurement in MPEG Test Tracks

MPEG recommended twelve tracks to be included in the assessment of audio quality. These twelve tracks contain critical music balancing on the percussion, string, wind instruments and human vocal. The characteristics and details of these twelve tracks are shown in the table below. In this section, the quality enhancement of proposed methods at different bit rates is verified based on the MPEG test tracks.

Table 6: The twelve tracks recommended by MPEG

Tracks		Signal Description			
		Signals	Mode	Time (sec)	Remark
1	es01	Vocal (Suzan Vega)	stereo	10	(c)
2	es02	German speech	stereo	8	(c)
3	es03	English speech	stereo	7	(c)
4	sc01	Trumpet solo and orchestra	stereo	10	(b) (d)
5	sc02	Orchestral piece	stereo	12	(d)
6	sc03	Contemporary pop music	stereo	11	(d)
7	si01	Harpsichord	stereo	7	(b)
8	si02	Castanets	stereo	7	(a)
9	si03	pitch pipe	stereo	27	(b)
10	sm01	Bagpipes	stereo	11	(b)
11	sm02	Glockenspiel	stereo	10	(a) (b)
12	sm03	Plucked strings	stereo	13	(a) (b)

Remarks:

(a) Transients: pre-echo sensitive, smearing of noise in temporal domain.

(b) Tonal/Harmonic structure: noise sensitive, roughness.

(c) Natural vocal (critical combination of tonal parts and attacks): distortion sensitive, smearing of attacks.

(d) Complex sound: stresses the device under test.

Table 7: Objective measurements through the ODGs for proposed PNS approaches at 128 kbps

Codec	NCTU-AAC			
Bit Rate	128 kbps			
Tracks	M0	M1	M2	M3
es01	-1.56	-1.54	-1.54	-1.54
es02	-1.98	-1.94	-1.94	-1.93
es03	-2.24	-2.2	-2.2	-2.19
sc01	-0.7	-0.74	-0.74	-0.81
sc02	-0.98	-0.71	-0.84	-0.64
sc03	-0.6	-0.58	-0.58	-0.57
si01	-1.09	-1.08	-1.08	-1.07
si02	-3.28	-3.2	-3.22	-3.17
si03	-1.21	-1.21	-1.21	-1.21
sm01	-0.81	-0.81	-0.81	-0.81
sm02	-1.54	-1.55	-1.55	-1.54
sm03	-1.2	-1.16	-1.18	-1.14
Max	-0.6	-0.58	-0.58	-0.57
Min	-3.28	-3.2	-3.22	-3.17
Average	-1.4325	-1.3933	-1.4075	-1.385
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

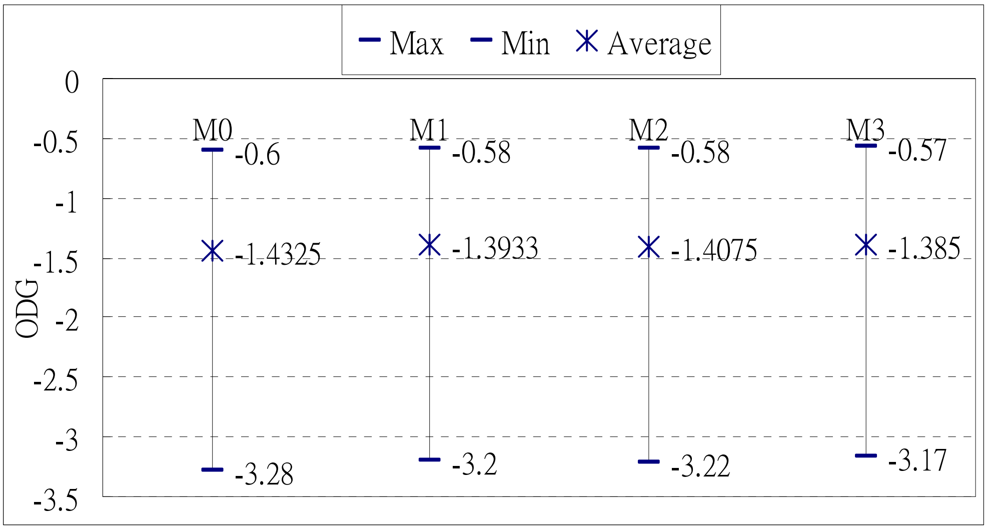


Figure 26: The variance in the ODGs of proposed PNS approaches at 128 kbps

Table 8: Objective measurements through the ODGs for proposed PNS approaches at 112 kbps

Codec	NCTU-AAC			
Bit Rate	112 kbps			
Tracks	M0	M1	M2	M3
es01	-2.11	-2.06	-2.04	-2
es02	-2.46	-2.39	-2.35	-2.3
es03	-2.82	-2.66	-2.59	-2.57
sc01	-1.01	-1	-1	-1.11
sc02	-1.5	-1.21	-1.18	-1.01
sc03	-0.98	-0.82	-0.84	-0.77
si01	-1.55	-1.54	-1.54	-1.55
si02	-3.52	-3.43	-3.42	-3.38
si03	-1.91	-1.91	-1.91	-1.91
sm01	-1.63	-1.62	-1.62	-1.62
sm02	-1.96	-1.98	-1.97	-1.95
sm03	-1.92	-1.77	-1.68	-1.58
Max	-0.98	-0.82	-0.84	-0.77
Min	-3.52	-3.43	-3.42	-3.38
Average	-1.9475	-1.8658	-1.845	-1.8125
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

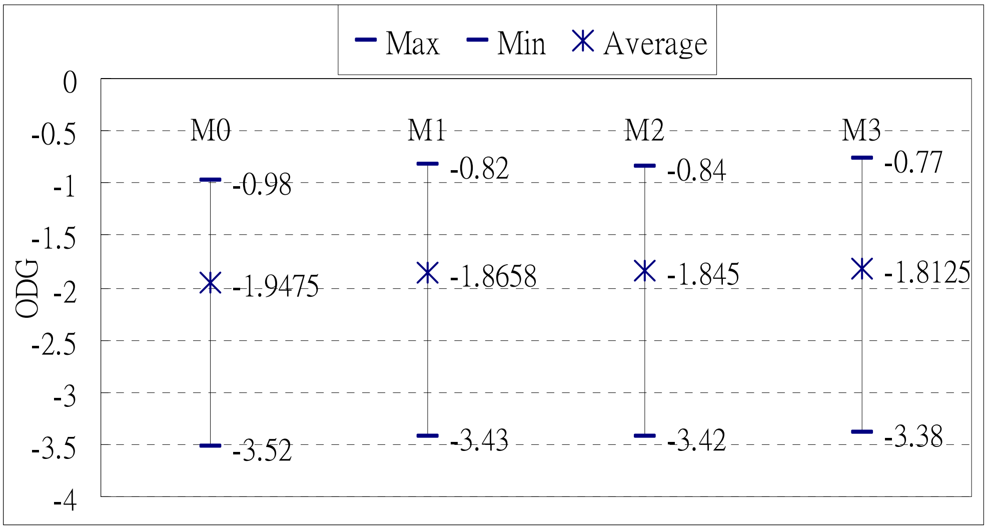


Figure 27: The variance in the ODGs of proposed PNS approaches at 112 kbps

Table 9: Objective measurements through the ODGs for proposed PNS approaches at 96 kbps

Codec	NCTU-AAC			
Bit Rate	96 kbps			
Tracks	M0	M1	M2	M3
es01	-2.98	-2.49	-2.61	-2.37
es02	-3.11	-2.71	-2.71	-2.54
es03	-3.52	-3.17	-3.12	-3.09
sc01	-1.52	-1.5	-1.43	-1.49
sc02	-2.04	-1.64	-1.65	-1.62
sc03	-2.15	-1.61	-1.43	-1.41
si01	-2.55	-2.41	-2.45	-2.41
si02	-3.73	-3.37	-3.46	-3.36
si03	-3.02	-2.37	-3.02	-2.37
sm01	-2.98	-2.51	-2.96	-2.51
sm02	-2.62	-2.52	-2.63	-3.2
sm03	-2.82	-2.39	-2.12	-2.21
Max	-1.52	-1.5	-1.43	-1.41
Min	-3.73	-3.37	-3.46	-3.36
Average	-2.7533	-2.3908	-2.4658	-2.3817
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

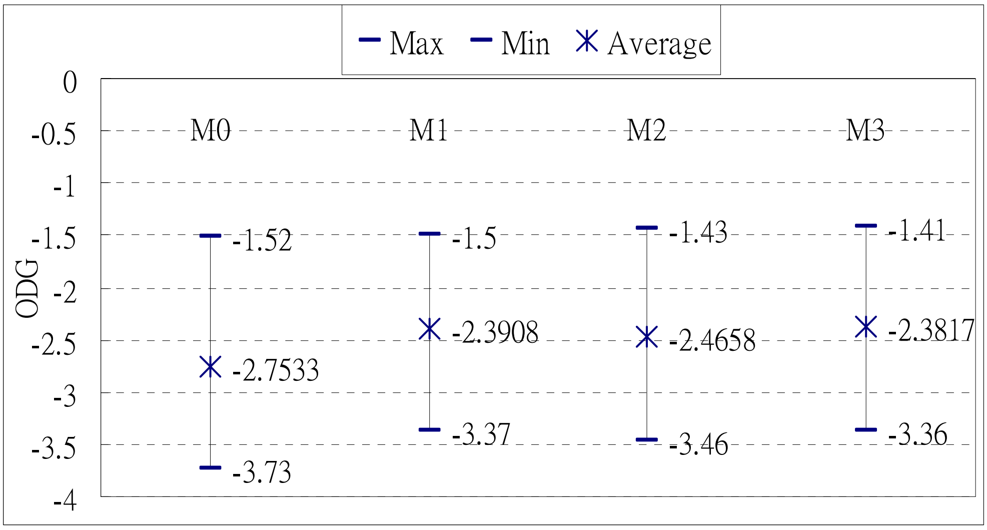


Figure 28: The variance in the ODGs of proposed PNS approaches at 96 kbps

Table 10: Objective measurements through the ODGs for proposed PNS approaches at 80 kbps

Codec	NCTU-AAC			
Bit Rate	80 kbps			
Track	M0	M1	M2	M3
es01	-3.65	-2.79	-3.14	-2.73
es02	-3.68	-2.98	-3.09	-2.94
es03	-3.85	-3.47	-3.46	-3.27
sc01	-1.94	-1.95	-1.95	-1.96
sc02	-2.66	-1.92	-2.26	-1.89
sc03	-3.26	-2.13	-2.13	-2.03
si01	-3.47	-3.33	-3.34	-3.32
si02	-3.84	-3.55	-3.62	-3.54
si03	-3.8	-2.56	-3.5	-2.65
sm01	-3.71	-3.15	-3.41	-3.13
sm02	-3.38	-3.21	-3.39	-3.18
sm03	-3.47	-2.64	-2.78	-2.62
Max	-1.94	-1.92	-1.95	-1.89
Min	-3.85	-3.55	-3.62	-3.54
Average	-3.3925	-2.8067	-3.0058	-2.7717
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

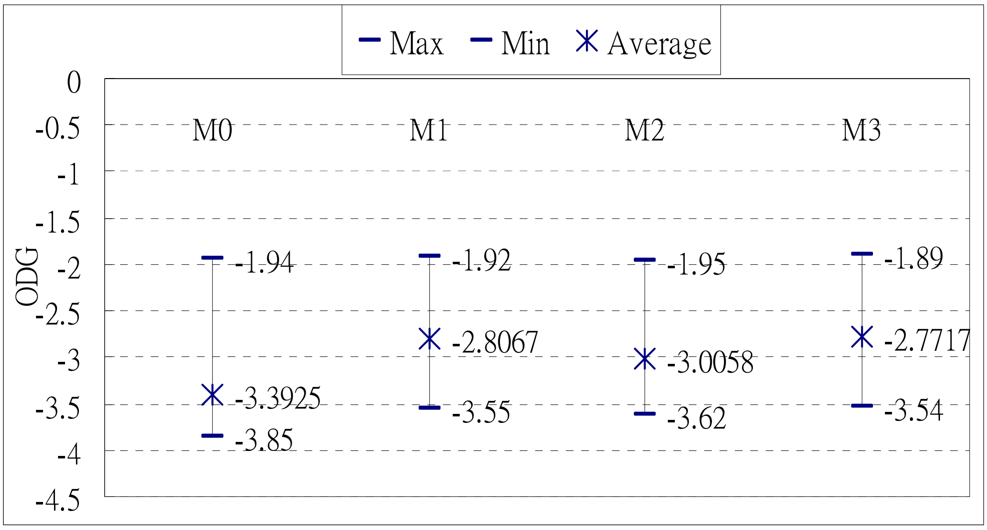


Figure 29: The variance in the ODGs of proposed PNS approaches at 80 kbps

Table 11: Objective measurements through the ODGs for proposed PNS approaches at 64 kbps

Codec	NCTU-AAC			
Bit Rate	64 kbps			
Track	M0	M1	M2	M3
es01	-3.84	-3.28	-3.46	-3.15
es02	-3.86	-3.25	-3.3	-3.11
es03	-3.9	-3.65	-3.64	-3.64
sc01	-2.01	-1.98	-1.97	-1.95
sc02	-2.93	-2.27	-2.64	-2.17
sc03	-3.59	-2.59	-2.96	-2.37
si01	-3.79	-3.43	-3.75	-3.44
si02	-3.89	-3.78	-3.82	-3.72
si03	-3.88	-2.81	-3.78	-2.81
sm01	-3.85	-3.64	-3.72	-3.64
sm02	-3.74	-3.71	-3.7	-3.7
sm03	-3.68	-2.96	-3.36	-2.95
Max	-2.01	-1.98	-1.97	-1.95
Min	-3.9	-3.78	-3.82	-3.72
Average	-3.58	-3.1125	-3.3417	-3.0542
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

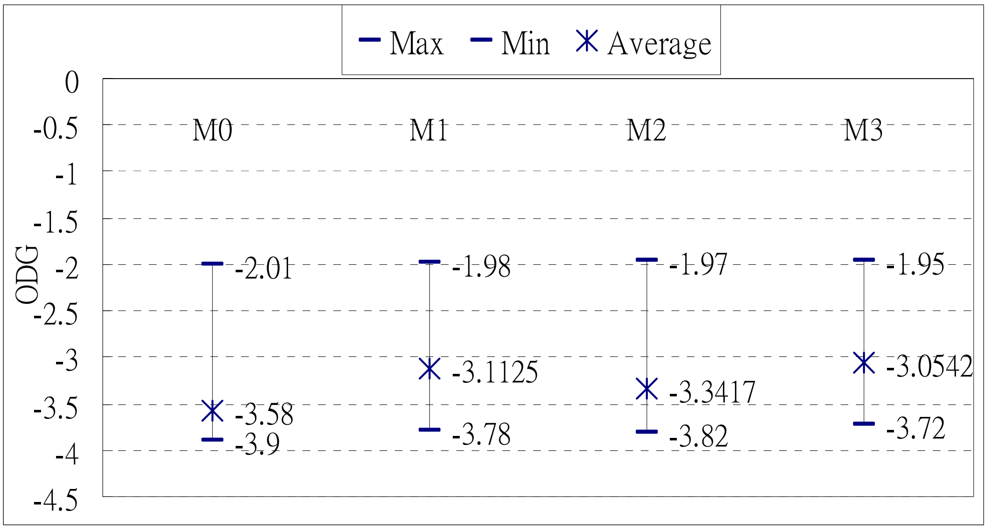


Figure 30: The variance in the ODGs of proposed PNS approaches at 64 kbps

Table 12: Objective measurements through the ODGs for adaptive masking offset in the psychoacoustic model

Codec	NCTU-AAC	
Bit Rate	128 kbps	
Tracks	Original Psycho	Enhanced Psycho
es01	-0.27	-0.26
es02	-0.07	-0.05
es03	-0.12	-0.07
sc01	-0.44	-0.42
sc02	-0.76	-0.79
sc03	-0.62	-0.61
si01	-0.62	-0.56
si02	-0.69	-0.65
si03	-0.95	-0.87
sm01	-0.6	-0.54
sm02	-0.48	-0.45
sm03	-0.8	-0.76
Max	-0.07	-0.05
Min	-0.95	-0.87
Average	-0.535	-0.5025
Enhanced psycho: the psychoacoustic model with adaptive masking offset		

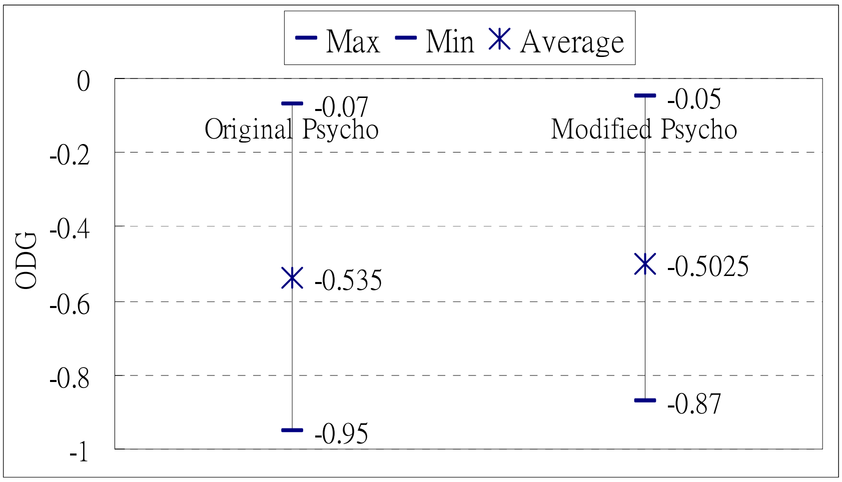


Figure 31: The variance in the ODGs of original and modified psychoacoustic models

The verification of the enhancement of enhanced M/S coding is not performed on the MPEG 12 test tracks. Because the MPEG 12 tracks contain only a few tracks which have middle channel similarity in L and R channels. The result of objective quality measurement in MPEG 12 tracks can not reflect the improvement of enhanced M/S coding. Furthermore, since the motivation of enhanced M/S coding is to eliminate the artifacts occurred in our previous M/S coding. The objective quality measurement should be performed on the tracks which have artifacts induced by previous M/S coding. Therefore, the experiment of modified M/S coding is performed on other ten tracks chosen from the music database in our lab [63]. The characteristics of these ten tracks are shown in Table 13.

Table 13: The ten tracks used to verify the enhanced M/S coding

Tracks		Signal Description			
		Signal	Mode	Time (sec)	Remark
1	41_30sec	Drum and cymbals	stereo	30	(a) (d)
2	CYMBALS	Electric guitar and cymbals	stereo	9.3	(d)
3	PinkFloydTime	Vocal and orchestra	stereo	18.8	(c) (d)
4	Scars	Flute and violin	stereo	22.8	(b) (d)
5	SmashingSample	Cymbals and vocal	stereo	12.7	(a) (c) (d)
6	taking_you_home	Piano and cymbals	stereo	10	(a) (d)
7	test_4_artifacts	Contemporary pop music	stereo	19.5	(c) (d)
8	Trust	Applause and singing	stereo	29	(c) (d)
9	Velvet	Electric music	stereo	11.9	(a)
10	youwalkaway	Drum and electric guitar	stereo	9.9	(d)

Remarks:

(a) Transients: pre-echo sensitive, smearing of noise in temporal domain.

(b) Tonal/Harmonic structure: noise sensitive, roughness.

(c) Natural vocal (critical combination of tonal parts and attacks): distortion sensitive, smearing of attacks.

(d) Complex sound: stresses the device under test.

Table 14: Objective measurements through the ODGs for enhanced M/S coding

Codec	NCTU-AAC	
Bit Rate	128 kbps	
Tracks	Original M/S	Enhanced M/S
41_30sec	-0.9	-0.89
CYMBALS	-0.81	-0.78
PinkFloydTime	-0.7	-0.67
Scars	-0.62	-0.61
SmashingSample	-0.45	-0.44
taking_you_home	-0.6	-0.58
test_4_artifacts	-0.61	-0.56
Trust	-0.66	-0.64
Velvet	-1.19	-1.18
youwalkaway	-0.62	-0.58
Max	-0.45	-0.44
Min	-1.19	-1.18
Average	-0.716	-0.693
Enhanced M/S: the M/S coding with inter channel bit allocation modified		

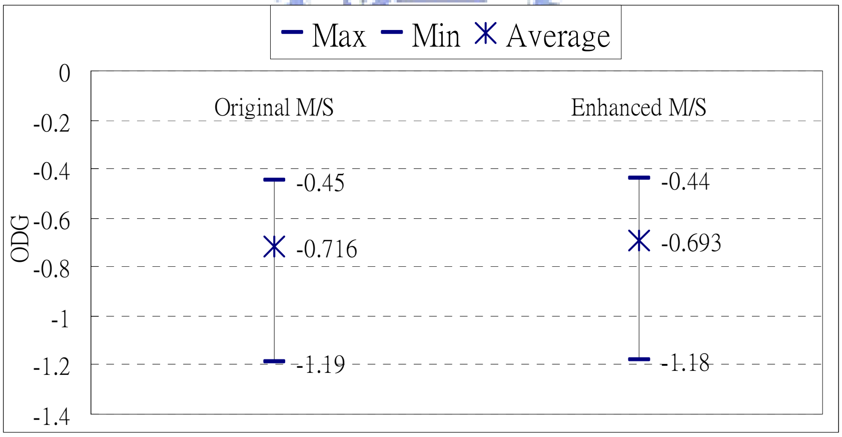


Figure 32: The variance in the ODGs of original and enhanced M/S coding

6.3 Objective Quality Measurement in Music Database

To confirm the possible risk and robustness of proposed methods, extensive tests are adapted to verify the qualities of these methods. For the past few years, an audio database [63] has been established in our laboratory [41]. The audio database includes 16 categories and 327 tracks with different signal properties. With the various audio content of the database, the quality of a codec can be assessed and guaranteed. The characteristics of each category in the audio database are shown in Table 15.

Table 15: The PSPLab audio database [63]

Bitstream Categories		# of tracks	Remark
1	Ff123	103	Killer bitstream collection from ff123 [64].
2	Gpsycho	24	LAME quality test bitstream [65].
3	HA64KTest	39	64 kbps test bitstream for multi-format in HA forum [66].
4	HA128KTestV2	12	128 kbps test bitstream for multi-format in HA forum [66].
5	Horrible_song	16	Collections of critical songs among all bitstreams in PSPLab.
6	Ingets1	5	Bitstream collection from the test of OGG Vorbis pre 1.0 listening test [67].
7	Mono	3	Mono test bitstream.
8	MPEG	12	MPEG test bitstream set for 48000Hz.
9	MPEG44100	12	MPEG test bitstream set for 44100 Hz.
10	Phong	8	Test bistream collection from Phong [68].
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 [69].
14	TestingSong14	14	Test bitstream collection from rshong, PSPLab.
15	TonalSignals	15	Artificial bitstream that contains sin wave etc.
16	VORBIS_TESTS_ Samples	8	Eight Vobis testing samples from HA [66].

M0: PNS module disabled

M1: Noise band substitution and high frequency signal substitution by PNS

M2: Zero-band dithered by PNS

M3: Combination of M1 and M2

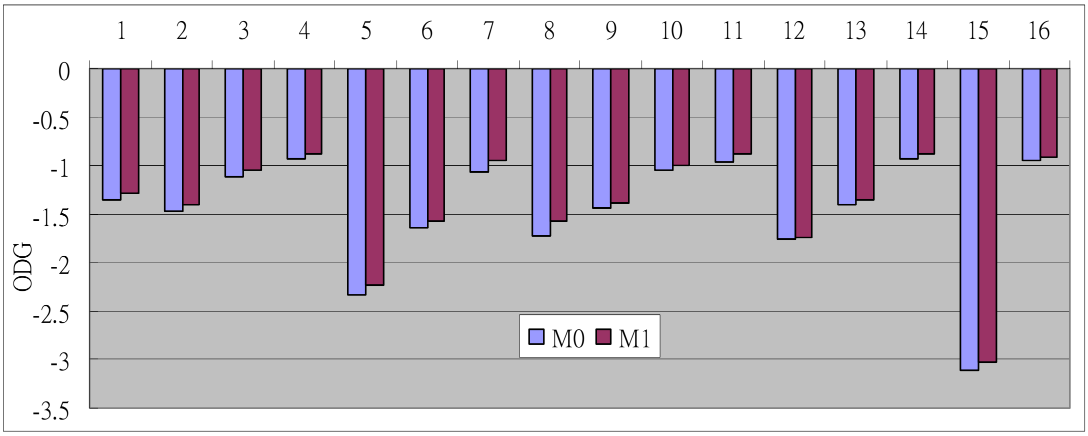


Figure 33: The average ODGs of method M0 and M1 at 128kbps in 16 categories

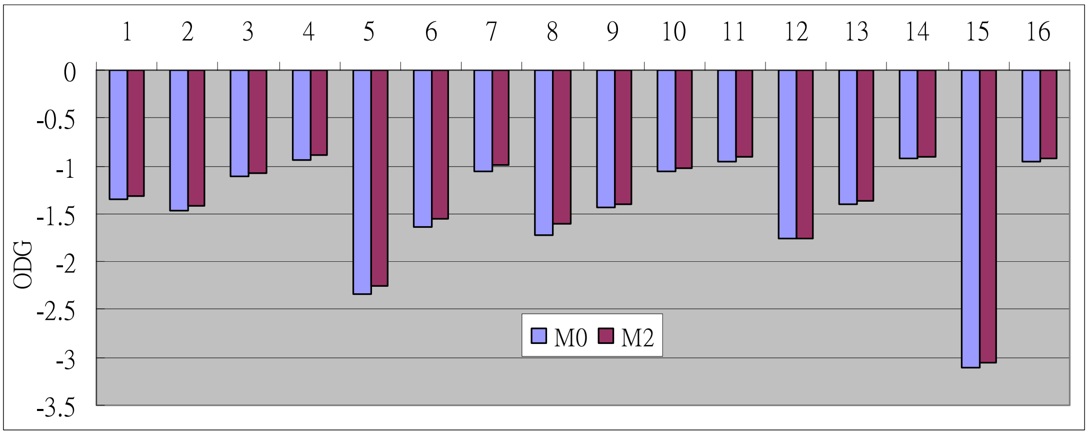


Figure 34: The average ODGs of method M0 and M2 at 128kbps in 16 categories

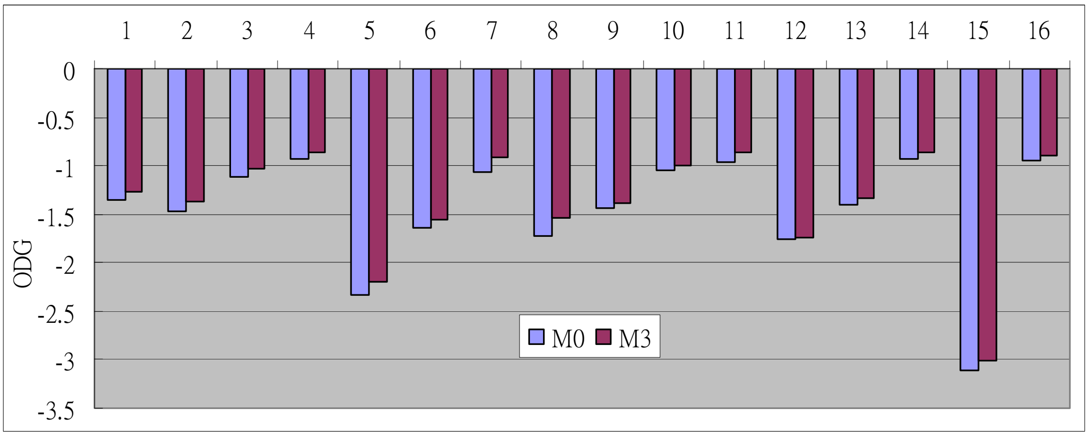


Figure 35: The average ODGs of method M0 and M3 at 128kbps in 16 categories

M0: PNS module disabled

M1: Noise band substitution and high frequency signal substitution by PNS

M2: Zero-band dithered by PNS

M3: Combination of M1 and M2

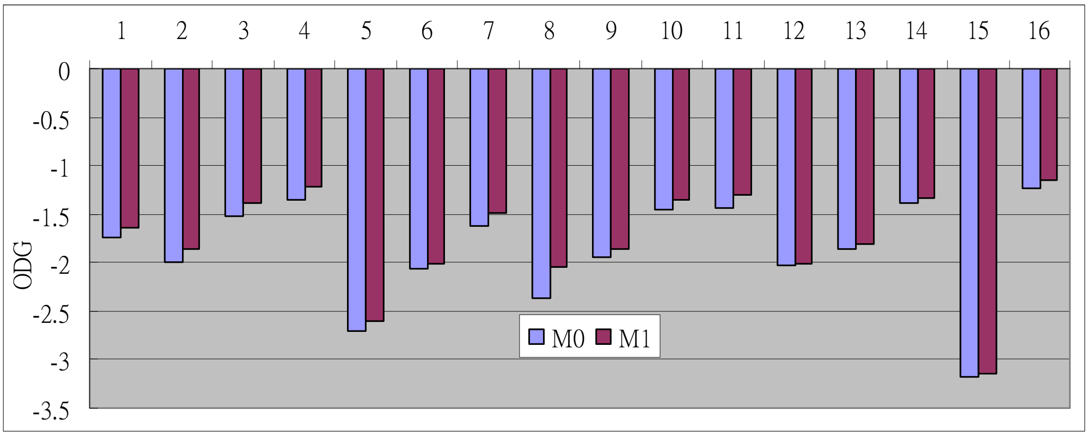


Figure 36: The average ODGs of method M0 and M1 at 112kbps in 16 categories

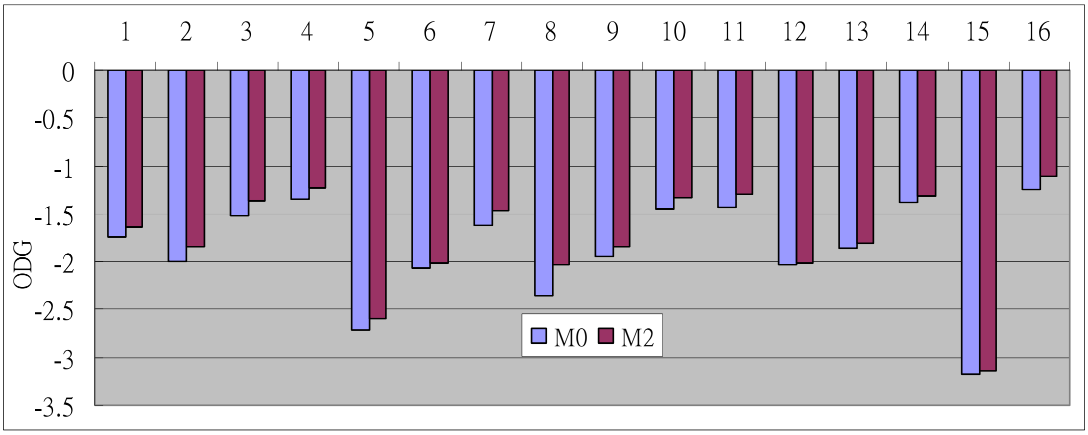


Figure 37: The average ODGs of method M0 and M2 at 112kbps in 16 categories

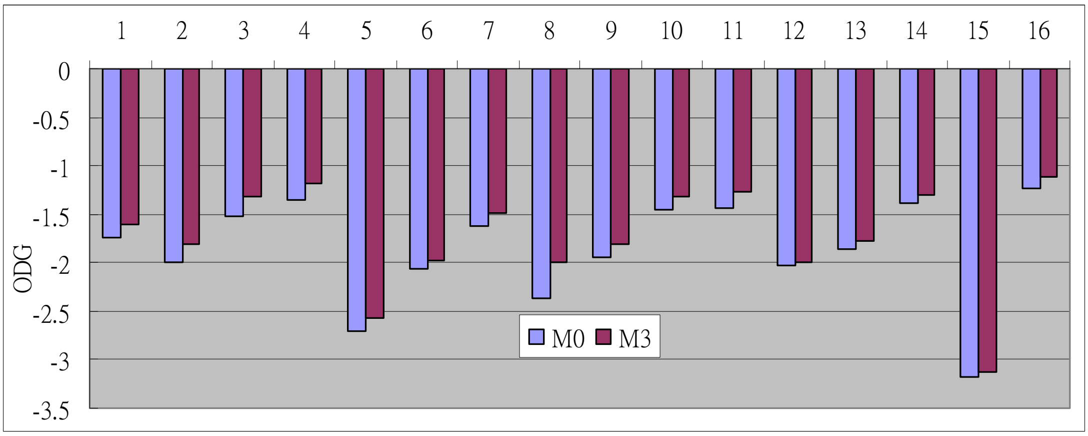


Figure 38: The average ODGs of method M0 and M3 at 112kbps in 16 categories

M0: PNS module disabled

M1: Noise band substitution and high frequency signal substitution by PNS

M2: Zero-band dithered by PNS

M3: Combination of M1 and M2

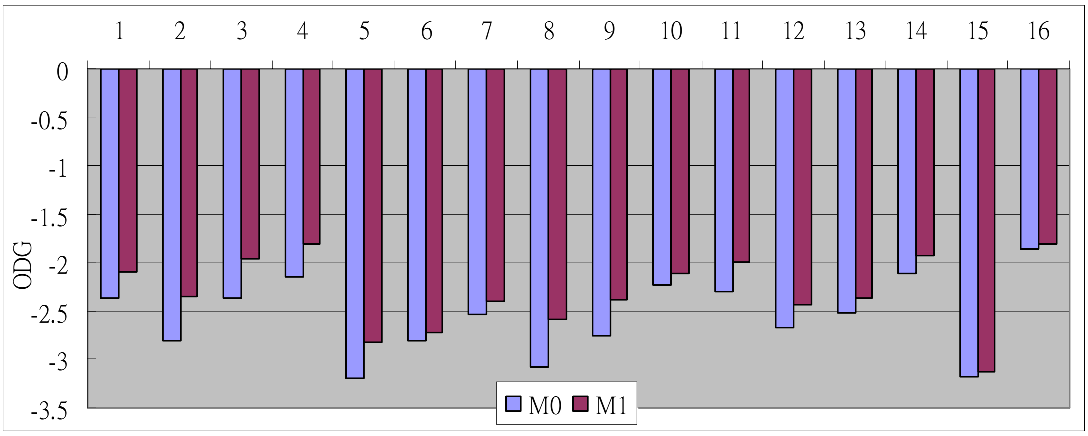


Figure 39: The average ODGs of method M0 and M1 at 96kbps in 16 categories

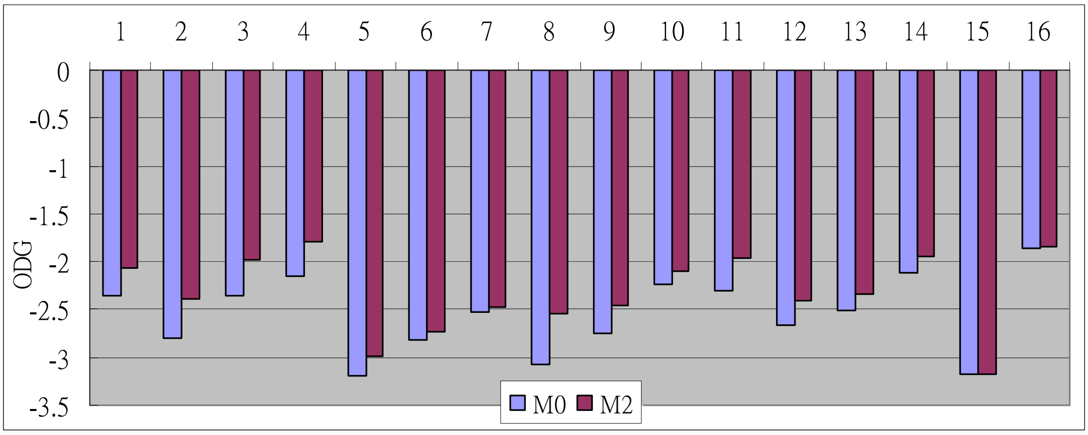


Figure 40: The average ODGs of method M0 and M2 at 96kbps in 16 categories

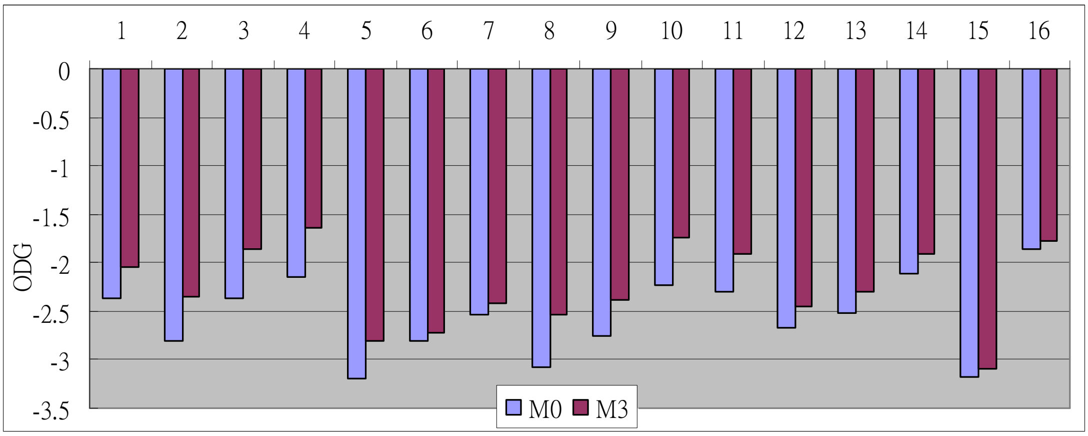


Figure 41: The average ODGs of method M0 and M3 at 96kbps in 16 categories

M0: PNS module disabled

M1: Noise band substitution and high frequency signal substitution by PNS

M2: Zero-band dithered by PNS

M3: Combination of M1 and M2

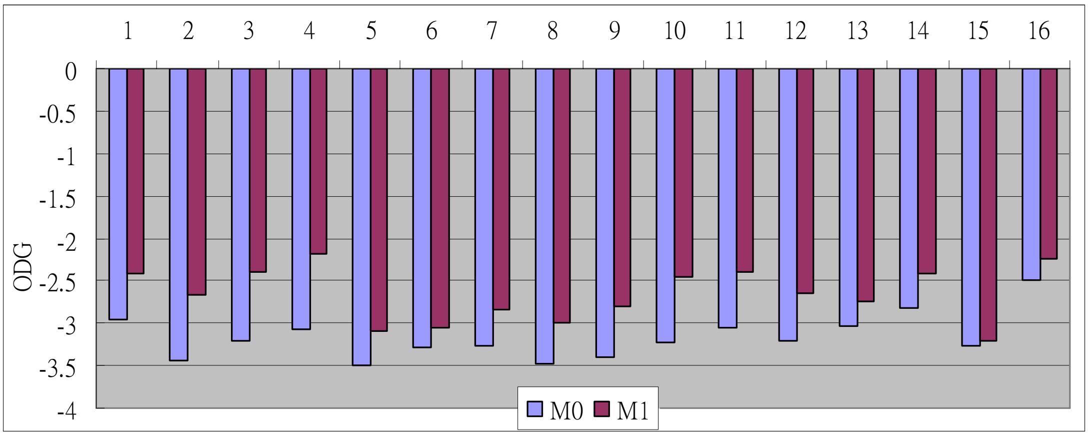


Figure 42: The average ODGs of method M0 and M1 at 80kbps in 16 categories

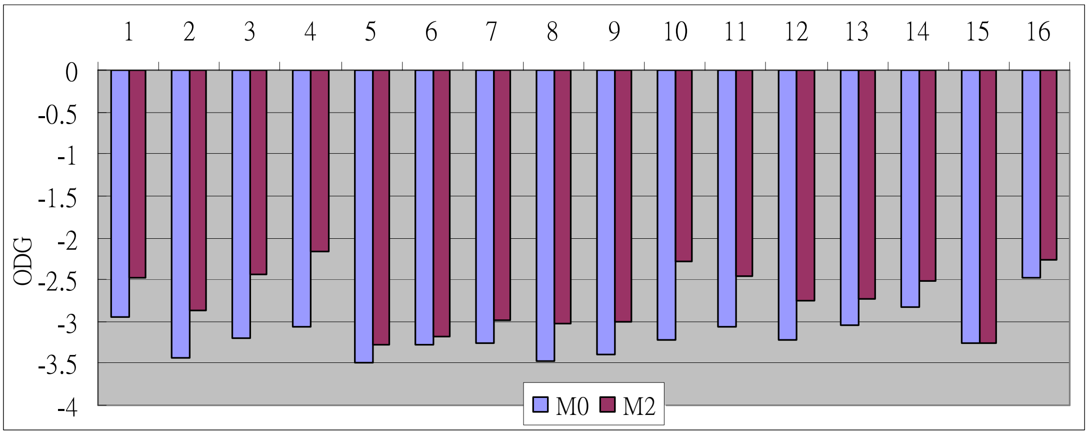


Figure 43: The average ODGs of method M0 and M2 at 80kbps in 16 categories

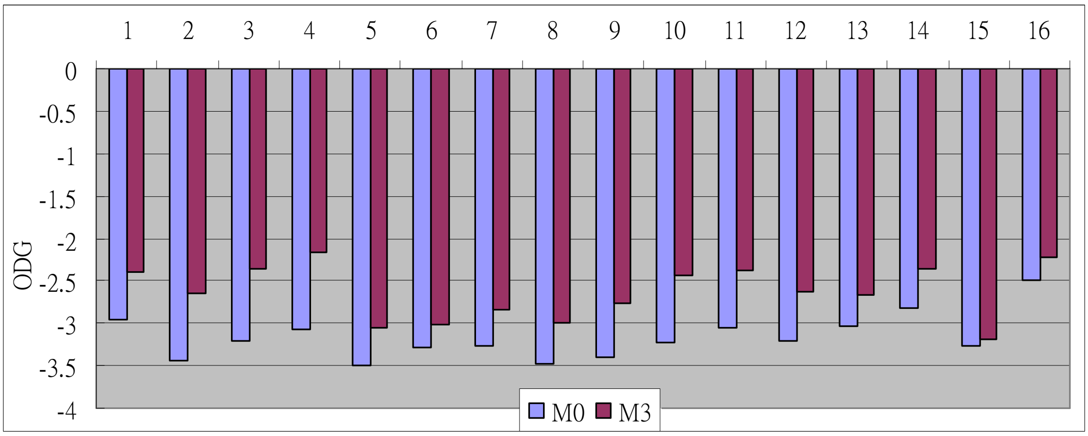


Figure 44: The average ODGs of method M0 and M3 at 80kbps in 16 categories

M0: PNS module disabled

M1: Noise band substitution and high frequency signal substitution by PNS

M2: Zero-band dithered by PNS

M3: Combination of M1 and M2

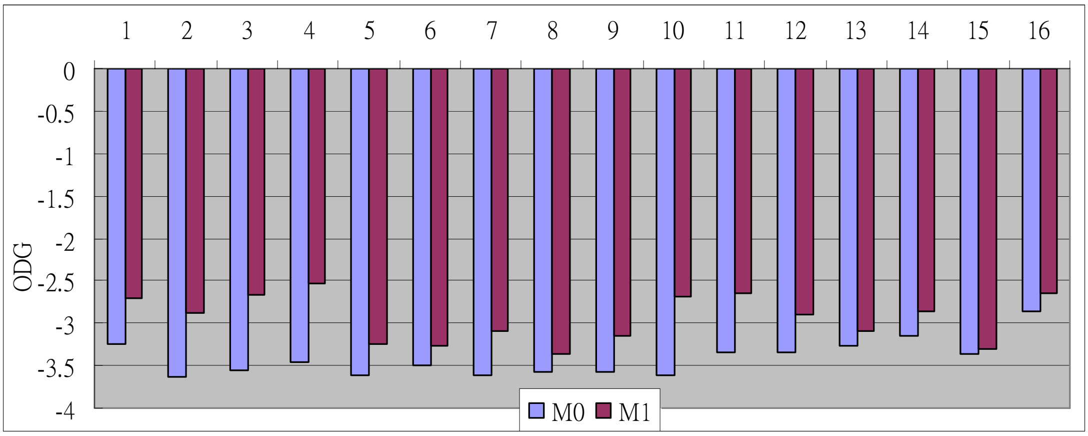


Figure 45: The average ODGs of method M0 and M1 at 64kbps in 16 categories

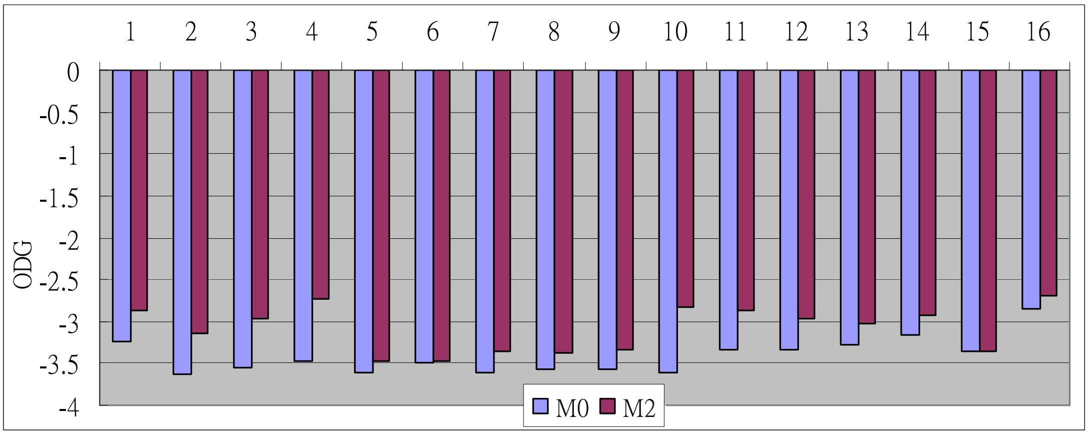


Figure 46: The average ODGs of method M0 and M2 at 64kbps in 16 categories

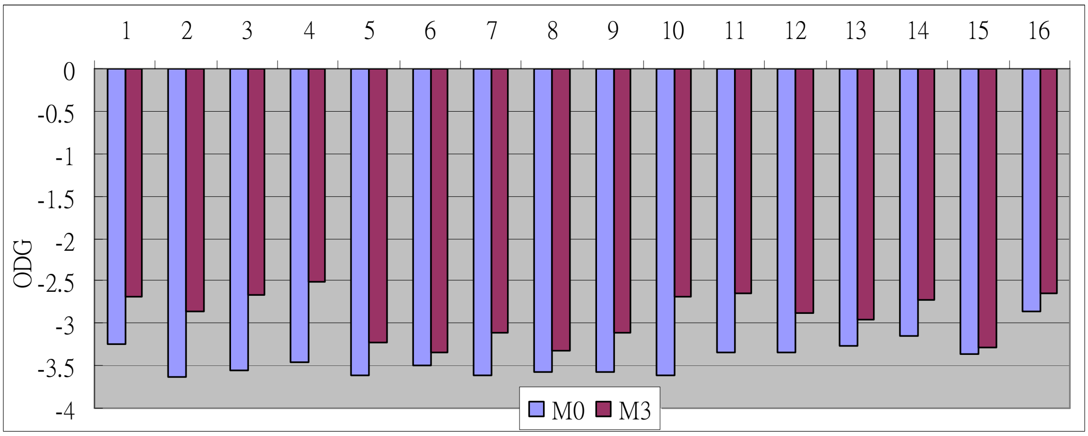


Figure 47: The average ODGs of method M0 and M3 at 64kbps in 16 categories

Table 16: The average ODGs of original psycho model and the modified psycho model with adaptive masking offset in16 categories

Codec	NCTU-AAC	
Bit Rate	128 kbps	
Categories	Original Psycho	Enhanced Psycho
FF123	-0.7586	-0.7272
Gpsycho	-0.7717	-0.7408
HA64KTest	-0.7084	-0.6792
HA128KTestV2	-0.6258	-0.5992
Horrible_song	-1.1194	-1.0794
Ingets1	-0.55	-0.522
Mono	-0.8967	-0.86
MPEG	-0.6733	-0.6342
MPEG44100	-0.535	-0.5025
Phong	-0.7613	-0.7175
PSPLab	-0.7346	-0.7041
Sjeng	-0.6167	-0.6067
SQAM	-0.3756	-0.3544
TestingSong14	-0.6221	-0.5907
TonalSignals	-2.44	-2.4187
VORBIS_TESTS_Samples	-0.5238	-0.4913
Enhanced psycho: the psychoacoustic model with adaptive masking offset		

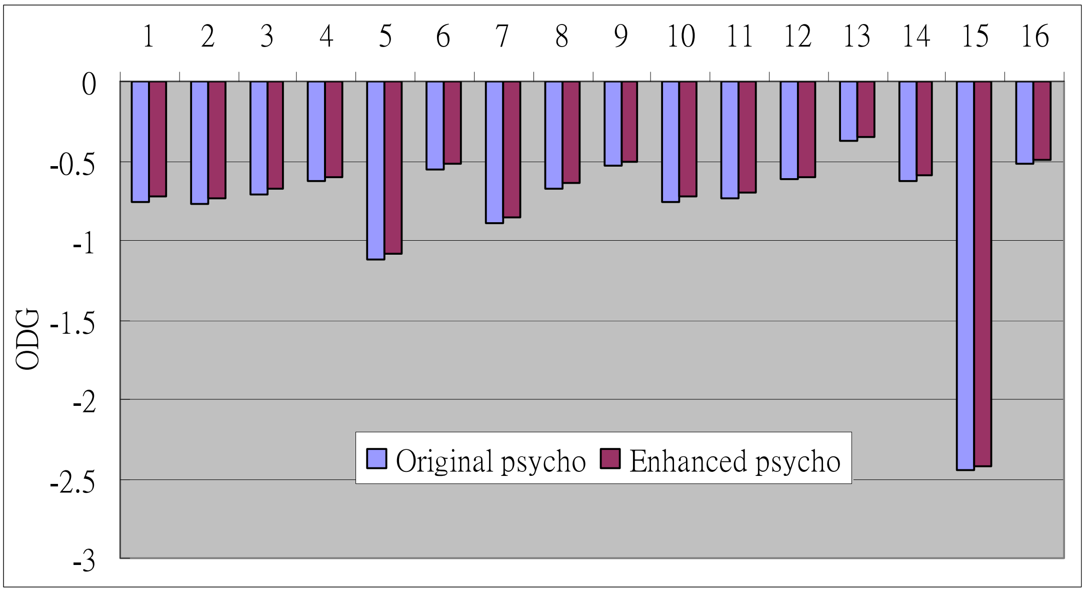


Figure 48: The ODG improvement of enhanced psycho model in 16 categories

Table 17: The average ODGs of original M/S coding and enhanced M/S coding in16 categories

Codec	NCTU-AAC	
Bit Rate	128 kbps	
Categories	Original M/S	Enhanced M/S
FF123	-0.7466	-0.7272
Gpsycho	-0.7379	-0.7283
HA64KTest	-0.6949	-0.6792
HA128KTestV2	-0.6225	-0.5992
Horrible_song	-1.0794	-1.0731
Ingets1	-0.482	-0.482
Mono	-0.86	-0.86
MPEG	-0.6317	-0.6317
MPEG44100	-0.5025	-0.5025
Phong	-0.7063	-0.6838
PSPLab	-0.7068	-0.7041
Sjeng	-0.6067	-0.6067
SQAM	-0.345	-0.3444
TestingSong14	-0.5921	-0.5907
TonalSignals	-2.406	-2.3987
VORBIS_TESTS_Samples	-0.4938	-0.4913
Enhanced M/S: the M/S coding with inter channel bit allocation modified		

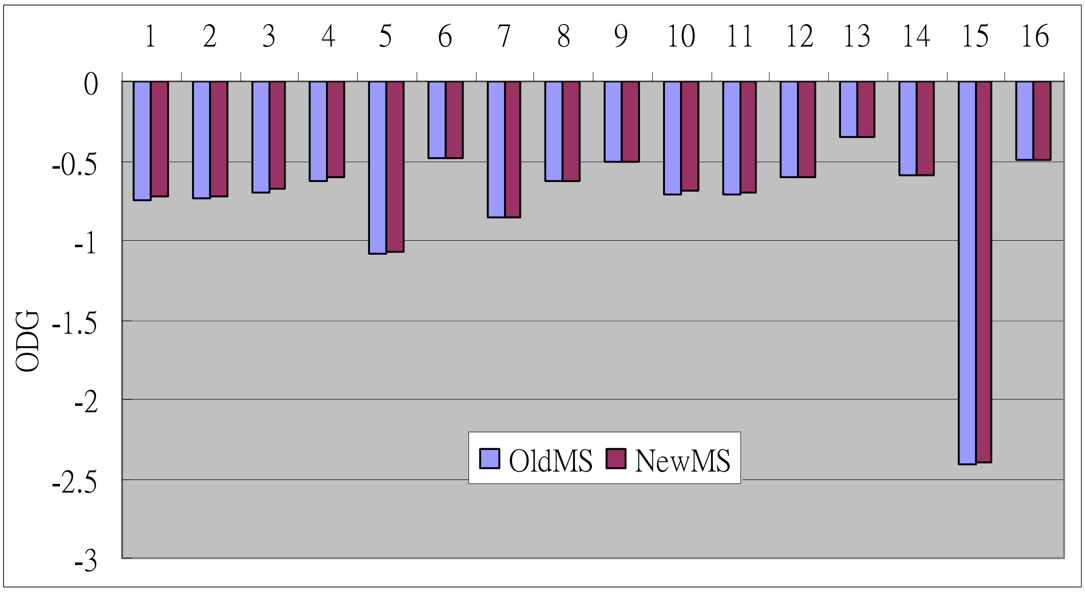


Figure 49: The ODG improvement of enhanced M/S coding in 16 categories

6.4 Subjective Quality Measurement

After passing through the objective quality measurement, we perform subjective listening test to verify the quality improvement and possible risk of proposed methods in this thesis. The subjective listening test is performed on the codec “NCTU-AAC” and use the tool called “MUSHRA” to assist the assessment. The assessment is performed on seven people in our lab. The ten tracks used for PNS are selected from the music database according ODG. As shown in the following figures, the result of subjective quality measurement is consistent with the result of objective quality measurement.

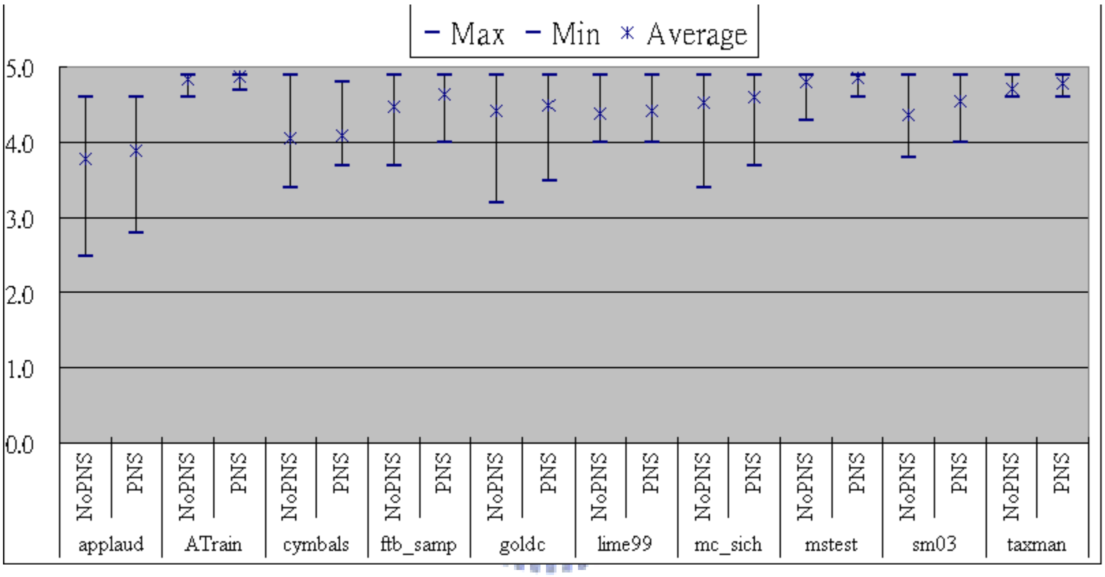


Figure 50: The result of subjective test for PNS at 96kbps.

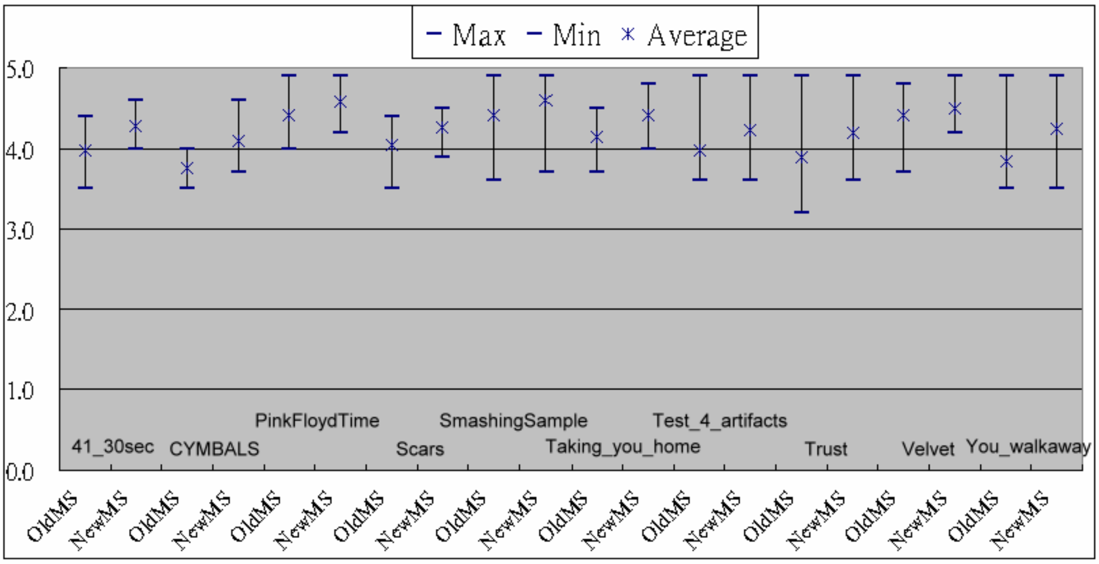


Figure 51: The result of subjective test for M/S at 128kbps.

6.5 Complexity

Computing Complexities of Proposed PNS Approaches

The computing complexities of proposed PNS approaches are measured by Intel VTune 7.0 [70]. The Intel VTune Performance Analyzer provides a performance analysis and tuning environment that helps you analyze your system and software performance. The clockticks shown in Table 18 are measured on the computer stated in section 6.1. We combine the MPEG 12 tracks into one track and duplicate it four times to form the test track. The result of the measurement is shown in Table 18. Since the signal in a PNS band is not quantized and coded, the methods M1 and M3 can reduce the computing complexity. The overhead of Method M2 in computing complexity is 0.31%, it has no obvious influence on the encoding speed. Therefore, the proposed PNS approaches can increase the encoding speed and improve the quality of encoded signal.

Table 18: Computing complexities of propose PNS approaches

Codec	NCTU-AAC			
Bit Rate	96 kbps			
	M0	M1	M2	M3
Clockticks	44549076768	41971386748	44686051168	42128532154
PNS band (%)	0.00%	12.21%	6.84%	13.26%
Overhead	0.00%	-5.79%	0.31%	-5.43%
M0: PNS module disabled				
M1: Noise band substitution and high frequency signal substitution by PNS				
M2: Zero-band dithered by PNS				
M3: Combination of M1 and M2				

Number of PNS Bands per Frame at Different Bit Rates

In this section, we measure the number of PNS bands per frame at different bit rates in MPEG 12 test tracks. We enabled the PNS to substitute the noise band and high frequency signals before quantization process and dither zero-bands after quantization. The results are shown in Table 19. For 128 to 80 kbps, the number of PNS bands is increasing when the bit rate is lower and lower. This situation is caused by the start frequency of high frequency signals substitution is lower and lower and the zero-bands appear more and more often. However, at 64kbps, the number of PNS bands begins to decrease. At 64kbps, the start frequency of high frequency substitution is lower than other bit rates and zero-bands occur more often other bit rates. However, according to our design of PNS, the number of bits used by PNS is bounded by the amount of bits saved in bit reservoir and a ratio of the total available bits of a frame. Therefore, the number of PNS bands at 64kbps is fewer than that of 80kbps.

Table 19: Number of PNS bands per frame at different bit rates

Codec	NCTU-AAC				
Tracks	# of PNS bands per frame (L and R channels)				
	128kbps	112kbps	96kbps	80kbps	64kbps
es01	7.63	13.21	19.55	26.63	19.98
es02	7.98	13.5	19.53	26.13	18.76
es03	8.52	14.03	20.56	27.56	20.73
sc01	11.04	17	23.85	31.23	24.27
sc02	15.51	19.65	24.31	29.63	24.13
sc03	9.21	12.8	17.09	22.13	16.97
si01	1.71	3.24	5.9	9.79	6.01
si02	10.87	16.25	22.15	28.97	22.43
si03	0.05	0.15	0.82	2.84	0.85
sm01	0.25	0.48	0.84	1.25	0.91
sm02	0.36	0.77	1.88	4.06	1.86
sm03	6.01	11.81	19.52	25.79	16.83
Average	5.89	9.1	13.09	17.68	12.86

Bit Usage of PNS

In this section, we measure the bits used by PNS bands. The data in the following table was measured in MPEG 12 tracks at 96kbps and PNS was enabled to substitute noise-like component, to retain high frequency signal and to dither zero-bands.

Table 20: The bits usage of PNS

Track	# of PNS band	Bits used	Bits per PNS band
es01	9050	45283	5
es02	7245	34563	4.77
es03	6743	31881	4.73
sc01	11281	43571	3.86
sc02	13348	54901	4.11
sc03	8509	40337	4.74
si01	2037	9692	4.76
si02	7375	37516	5.09
si03	988	5390	5.46
sm01	402	1841	4.58
sm02	817	3738	4.58
sm03	11769	51117	4.34
Average	6630.33	29985.83	4.67

Chapter 7 Concluding Remarks and Future Works

This thesis has presented the design of perceptual noise substitution in MPEG-4 Advanced Audio Coding through various approaches. The proposed PNS approaches can be classified into two sets. One is signal substitution before quantization process while another is zero-band dithering after quantization process. Signal substitution before quantization process consists of two parts: noise-like component substitution and high frequency signal substitution. The noise-like component substitution uses PNS parameters to replace the noisy band in the encoder end and reproduce random noise in the decoder end without any obvious distortions for human perception. The high frequency signal substitution can substitute the signal in high frequency to maintain a better precision in low frequency range. Another set, zero-band dithering after quantization process, can transform the zero-bands into PNS bands to eliminate the birdie artifacts [39] and improve the sound quality. Furthermore, this thesis also proposed two enhancing mechanisms for our previous work in psychoacoustic model and M/S coding. The proposed adaptive masking offset in psychoacoustic model can reflect the influence of frequency when determining the masking ability of a signal to gain a more precise psychoacoustic model in audio coding. The enhanced inter channel bit allocation in M/S coding can eliminate the artifacts appeared in previous M/S coding and improve the sound quality. Finally, extensive experiments have been conducted to verify the quality enhancement, possible risk and robustness of proposed methods. Through both subjective listening test and objective quality measurement, the proposed methods are verified to be able to improve the perceived quality of coded audio signals.

Although the experiments have showed the improvement of proposed methods, there are several aspects which can be enhanced in the future. First, the proposed PNS approaches do not consider the relation between M/S and PNS. In current design, M/S module is precedent to PNS module. The bands that PNS can take effect are the bands whose M/S flags are not turned on. There should be a method to switch bands between M/S and PNS, not let M/S precedent PNS always. Second, the enhanced M/S coding can be modified to be smoother, not so harsh. In the proposed method, the bits allocated to S channel will change rapidly near the threshold adopted in (26). This is caused by the discontinuity of proposed method. The function to allocate bits between M and S channels can be replaced with another continuous function (e.g. sin, cos, and etc.) to distribute bits smoothly.

References

[1] International Electrotechnical Commission/ American National Standard Institute (IEC/ANSI) CEI-IEC-908, “Compact Disc Digital Audio System” (“red book”), 1987.

[2] J. Johnston and K. Brandenburg, “Wideband Coding – Perceptual Considerations for Speech and Music,” in Advances in Speech Signal Processing (S. Furui and M. Sondhi, eds.), Ch. 4, pp. 109-140, New York: Dekker, 1992.

[3] ISO/IEC 11172-3, “Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5 Mbit/s – Part 3: Audio,” ISO/IEC JTC 1/SC 29, May 1993.

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

[5] B. Scharf, “Critical Bands,” in Foundations of Modern Auditory Theory (J. Tobias, ed.), vol. 1, Ch. 5, pp. 159-202, New York: Academic Press, 1970.

[6] E. Zwicker and H. Fastl, Psychoacoustics: Facts and Models, Ch. 6, New York: Springer-Verlag, 1990.

[7] Chris A. Lanciani, “Auditory Perception and the MPEG Audio Standard,” Georgia Institute of Technology School of Electrical and Computer Engineering, August 11, 1995.

[8] R. Hellman, “Asymmetry of Masking Between Noise and Tone,” Perception and Psychophysics, vol. 11, pp. 241-246, 1972.

[9] T. Sporer, U. Gbur, J. Herre, and R. Kapust, “Evaluating a Measurement System,” J. Audio Eng. Soc., vol 43, pp. 353-363, May 1995.

[10] K. Brandenburg and G. Stoll, “ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio,” J. Audio Eng. Soc., pp. 780-792, Oct. 1994.

[11] P. Paramichalis, “MPEG audio compression: Algorithms and implementation,” in Proc. DSP 95 Int. Conf. DSP, pp. 72-77, June 1995.

[12] ISO/IEC, JTC1/SC29/WG11 MPEG, “Generic coding of moving pictures and associated audio – Audio (non backward compatible coding, NBC),” JTC1/SC29/WG11 MPEG, Committee Draft 13818-7 1996 (“MPEG-2 NBC/AAC”).

[13] L. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, “AC-2 and AC-3: Low-complexity transform-based audio coding,” Collected Papers Digital Audio Bit-Rate Reduction, pp. 54-72, 1996.

[14] D. Shiha, J. Johnson, S. Dorward, and S. Quackcnbush, “The perceptual audio coder (PAC),” in The Digital Signal Processing Handbook, V. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC Press, pp. 42.1-42.18, 1998.

[15] International Telecommunications Union, Radiocommunication Sector BS.1116 (rev. 1), “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems,” Geneva, 1997.

[16] International Telecommunications Union, Radiocommunication Sector BS.562-3, “Subjective Assessment of Sound Quality,” Geneva 1990.

[17] ISO/IEC JTC 1/SC 29/WG 11 MPEG 94/063, “Report on the MPEG/Audio Multichannel Formal Subjective Listening Tests,” 1994.

[18] International Telecommunications Union, Radiocommunication Sector 10/51-E, “Low Bit Rate Multichannel Audio Coder Test Results,” Geneva 1995.

[19] International Telecommunications Union, Radiocommunication Sector 10/2-23-E, “Chairman Report of the Second Meeting of the Task Group 10/2,” Geneva 1992.

[20] ISO/IEC JTC 1/SC 29/WG 11 MPEG 91/010, “The MPEG/AUDIO Multichannel Subjective Listening Test,” Stockholm, April/May 1991.

[21] M. R. Schroeder, B. S. Atal and J. L. Hall, “Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear,” JASA, vol. 66 no. 6, pp. 1647-1652, December 1979.

[22] M. Karjalainen, “A New Auditory Model for the Evaluation of Sound Quality of Audio Systems,” Proc. of ICASSP, pp. 608-611, March 1985.

[23] K. Brandenburg and T. Sporer, “NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria, “Proc. of AES 11th Intl. Conf., Portland, May 1992.

[24] J. G. Beerends and J. A. Stemerdink, “A Perceptual Audio Quality Measure Based in a Psychoacoustic Sound Representation,” J. Audio Eng. Soc., vol. 40, no. 12, pp. 963-978, December 1992.

[25] B. Paillard, P. Mabilleu, S. Morissette and J. Soumagne, “PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals,” J. Audio Eng. Soc., pp. 21-31, vol. 40, January/February 1992.

[26] C. Colomes, M. Lever, J.B. Rault and Y.F. Dehéry, “A Perceptual Model Applied to Audio Bit Rate Redduction,” presented at the 95th AES Convention, Preprint 3741, New York, October 1993.

[27] ISO/IEC JTC 1/SC 29/WG 11 MPEG 95/201, “Chairman’s Report on the Work of the Audio ad Hoc Group on Objective Measurements,” Tokyo, July 1995.

[28] International Telecommunications Union, Radiocommunication Sector Bs.1387, “Method for the ObjectiveMeasurement of Perceived Audio Quality,” Geneva 1998.

[29] W. C. Treurniet and G. A. Soulodre, “Evaluation of the ITU-R Objective Audio Quality Measurement Method,” J. Audio Eng. Soc., vol. 48, no. 3, pp. 167-173, March 2000.

[30] T. Tremain, “The Government Standard Linear Predictive Coding: LPC-10,” Speech Technology,

pp. 40-49, April 1982.

[31] L. Supplee, R. Cohn, J. Collura, “MELP: The New Federal Standard at 2400 BPS,” IEEE ICASSP 1997, pp. 1591-1594.

[32] J. Herre and D. Schulz, “Extending the MPEG-4 AAC Codec by Perceptual Noise Substitution,” presented at 104th AES Convention, Amsterdam, May 1998.

[33] ISO/IEC 14996-3, “Information Technology – Coding of Audio-Visual Objects – Part 3: Audio,” JTC1/SC29/WG11, 2003.

[34] I. Varga, “Adaptive Filtering for Noise Reduction in Audio Signals,” J. Audio Eng. Soc., vol. 40, p. 436, May 1992.

[35] J. Makhoul, “Linear Prediction: A Tutorial Review,” Proc. IEEE, vol. 63, pp.567-580, April 1975.

[36] C.S. Yao, “On Improvement of Modules in MPEG-4 T/F coding,” A Master Thesis from Department of Computer Science and Information Engineering, National Chiao Tung University, June 2000.

[37] X. Serra, “A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic plus Stochastic Decomposition,” Ph.D. dissertation, Stanford University, Stanford CA, 1990

[38] J.D. Chen, “Improvement of PNS Tools for MPEG-4 General Audio Coding,” A Master Thesis from Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, June, 2002.

[39] V. Prakash, A. Kumar, P. Konda and S. C. Vadapalli, “Removal of Birdie Artifact in Perceptual Audio Coders,” presented at the 116th AES Convention, Berlin, May 8-11, 2004.

[40] NCTU-AAC, website: http://psplab.csie.nctu.edu.tw/projects/index.pl/nctu-aac.html, Date: June 03, 2005.

[41] Perceptual Signal Processing Lab, website: http://psplab.csie.nctu.edu.tw, Date: June 03, 2005.

[42] C.M. Liu and W.C. Lee, “A unified Fast Algorithm for Cosine Modulated Filter Banks in Current Audio Coding Standards,” J. Audio Eng. Soc., vol. 47, no. 12, December 1999.

[43] C.M. Liu and W.C. Lee, “Unified Recursive Decomposition Architecture for Cosine Modulated Filter Banks,” U.S. patent 6119080.

[44] W.C. Lee, “Design of the Audio Coding Standards for MPEG and AC-3,” Ph.D. dissertation, National Chiao Tung University.

[45] C.M. Liu, W.J. Lee, and R.S. Hong, “A New Criterion and Associated Bit Allocation Method for Current Audio Coding Standards,” Proc. of the 5th Int. Conference on Digital Audio Effects (DAFX-02), Hamburg, Germany, September 26-28, 2002.

[46] C.M. Liu, W.C. Lee, and Y.H. Hsiao, “M/S Coding based on Allocation Entropy,” Proc. of the 6th

Int. Conference on Digital Audio Effects (DAFX-03), London, UK, September 8-11, 2003.

[47] C.M. Liu, W.C. Lee, and H.W. Hsu, "High Frequency Reconstruction for Band-linited Audio Signals," Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03), London, UK, September 8-11, 2003.

[48] C.M. Liu, W.C. Lee, and H.W. Hsu, "High Frequency Reconstruction by Linear Extrapolation," presented at the 115th AES Convention, New York, October 10-13, 2003.

[49] C.M. Liu, W.C. Lee, and C.T. Chien, "Bit Allocation for Advanced Audio Coding Using Bandwidth Proportional Noise-shaping Criterion," Proc. of the 6th Int. Conference on Digital Audio Effects (DAFX-03), London, UK, September 8-11, 2003.

[50] H.W. Hsu, C.M. Liu, and W.C. Lee, "Audio Patch Method in Audio Decoders - MP3 and AAC," presented at the 116th AES Convention, Berlin, May 8-11, 2004.

[51] C.M. Liu, W.C. Lee, and T.W. Chang, "The Efficient Temporal Noise Shaping Method," presented at the 116th AES Convention, Berlin, May 8-11, 2004.

[52] C.M. Liu, L.W. Chen, and et al., "Efficient Bit Reservoir Design for MP3 and AAC," presented at the 117th AES Convention, San Francisco, October 28-31, 2004.

[53] C.M. Liu, and C.H. Yang, and et al., "Design of MPEG-4 AAC Encoder," presented at the 117th AES Convention, San Francisco, October 28-31, 2004.

[54] H.W. Hsu, C.M. Liu, W.C. Lee, and Z.W. Li, "Audio Patch Method in MPEG-4 HE-AAC Decoder," presented at the 117th AES Convention, San Francisco, October 28-31, 2004.

[55] T. Chiou, "Design of Psychoacoustic Model for MPEG-4 Advanced Audio Coding and MPEG Layer III," A Master Thesis from Department of Computer Science and Information Engineering, National Chiao Tung University, June, 2004.

[56] M. B. and R. E. Goldberg, "Introduction to Digital Audio Coding and Standards," published by Kluwer Academic Publishers, 2003.

[57] J.D. Johnston, "Perceptual Transform Coding of Wideband Stereo Signals," Proc. of ICASSP, pp. 1993-1996, 1989.

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

[59] J.D. Johnston, "Estimation of Perceptual Entropy Using Noise Masking Criteria," Proc. of ICASSP, pp. 2524-2527, February 1988.

[60]J.D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE J. Select. Areas Commun., vol. 6, pp. 314-323, 1988.

[61] EAQUAL, mirror site: http://public.planetmirror.com/pub/sourceforge/e/ea/eaqual/, Date: June 03,

2005.

[62] MUSHRA, website: http://ff123.net/abchr/abchr.html, Date: June 03, 2005.

[63] The Audio Database Collected in Perceptual Signal Processing Lab, website: http://psplab.csie.nctu.edu.tw/projects/index.pl/testbitstreams.html, Date: June 03, 2005.

[64] Samples for Testing Audio Codecs from ff123, website: http://ff123.net/samples.html, Date: June 03, 2005.

[65] Quality and Listening Test Information for LAME, website: http://lame.sourceforge.net/gpsycho/quality.html, Date: June 03, 2005.

[66] Hydrogen Audio, website: http://www.hydrogenaudio.org, Date: June 03, 2005.

[67] OGG Vorbis Pre 1.0 Listening Test, website: http://hem.passagen.se/ingets1/vorbis.htm, Date: June 03, 2005.

[68] Phong’s Audio Samples, website: http://www.phong.org/audio/samples.xhtml, Date: June 03, 2005.

[69] Sound Quality Assessment Material, website: http://sound.media.mit.edu/mpeg4/audio/sqam/, Date: June 03, 2005.

[70] Intel VTune Performance Analyzers, website: http://www.intel.com/software/products/vtune/, Date: June 6, 2005.