

國立交通大學

資訊科學與工程研究所

碩士論文

Parametric Stereo Coding 中參數抽取與聲道
合成之設計



Design of T/F Stereo Parameter Extraction and
Downmix Method in Parametric Stereo Coding

研究生：李侃峻

指導教授：劉啟民 教授

李文傑 博士

中華民國九十五年六月

Parametric Stereo Coding 中參數抽取與聲道合成之設計
Design of T/F Stereo Parameter Extraction and Downmix Method in
Parametric Stereo Coding

研究生：李侃峻

Student : Kan-Chun Lee

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

資訊科學與工程研究所



Submitted to Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
Master
in
Computer Science

June 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年六月

Parametric Stereo Coding 中參數抽取與聲道合成 之設計

學生：李侃峻

指導教授：劉啓民 博士
李文傑 博士

國立交通大學資訊科學與工程研究所

中文論文摘要

在低位元音訊壓縮的需求之下，Parametric Stereo Coding (PS)會是一個有幫助的工具。PS 的壓縮概念來自於將雙聲道訊號合成為單一聲道訊號，再加上由左右聲道所抽取出的雙聲道特徵來進行壓縮。而在解碼端即可利用單聲道訊號以及抽取的參數來重建回雙聲道訊號。由於所抽取的參數只需要少量的位元需求，所以整體在位元的使用量就可以大幅的降低。對於 PS 壓縮的設計，本篇論文會著重在參數抽取和聲道合成上。參數抽取控制參數在時間與頻率上的解析度，以達到合適的特徵更新，而聲道合成部份將討論單一聲道訊號的產生辦法。最後在實作方面，由於 PS 可以架構在 HE-AAC 之上，讓 HE-AAC 處理單聲道的壓縮，所以實驗平台將架設在 NCTU-HEAAC 上。實驗方式會包含主觀測試以及客觀測試兩方面，由這兩種測試來驗證本論文所提出方法的壓縮品質。

Design of T/F Stereo Parameter Extraction and Downmix Method in Parametric Stereo Coding

Student: Kan-Chun Lee

Advisor: Dr. Chi-Min Liu

Dr. Wen-Chieh Lee

Institute of Computer Science and Information Engineering
National ChiaoTung University

ABSTRACT

Parametric Stereo Coding (PS) tool is an efficient tool for low bit-rate audio compression. It combines the stereo signal into a monaural downmix signal and extracts the parameters of stereo signal as the stereo image. By above information, the PS decoder can reconstruct the stereo signal. Moreover, the stereo parameters requires only a small overhead, the bit usage is substantially decreased. For the design of PS tool, the thesis will focus on the T/F stereo parameter extraction which controls the delivery of stereo parameters both in time domain and frequency domain, and downmix method for generating the monaural downmix signal. Because the PS tool can be jointly used with the HE-AAC for downmix signal encoding, these methods are integrated in the NCTU-HEAAC for implementation and to verify the coding quality. The extensive experiments are executed for both subjective and objective quality measurement in the end of the thesis.

致謝

感謝劉啓民老師兩年來的栽培及李文傑博士給予的指導，實驗室的楊宗翰學長、許瀚文學長、唐守宏同學、張家銘同學和楊詠成同學,以及學弟胡正倫、曾信耀和曾詩忠的協助，在研究上提供我寶貴的意見，讓我在專業知識及研究方法獲得非常多的啟發。

最後，感謝我的父母與家人及系上同學，在我研究所兩年的生活中，給予我無論在精神上以及物質上的種種協助，使我能全心全意地在這個專業的領域中研究探索在此一併表達個人的感謝。



Contents

Contents	iv
Figure List.....	vi
Table List.....	viii
Chapter 1 Introduction	1
Chapter 2 Basic Components.....	3
2.1 Parametric Stereo Coding Concept.....	3
2.1.1 Hybrid Filterbank.....	4
2.1.2 Framing.....	6
2.1.3 De-correlation Signal.....	6
2.1.4 Stereo Processing.....	7
Chapter 3 T/F Stereo Parameter Extraction	9
3.1 T/F Stereo Parameter Extraction Concept	9
3.2 Existing Approaches in Other Codecs	9
3.3 Adaptive T/F Stereo Parameter Extraction	10
3.3.1 Region Decision.....	11
3.3.1.1 Stereo Parameter Estimation and Error Calculation	11
3.3.1.2 Dynamic Programming.....	12
3.3.1.3 Region Decision based on Dynamic Programming.....	13
3.3.2 Stereo Band Decision.....	16
3.4 T/F Stereo Parameter Extraction Summary	17
Chapter 4 Downmix Method	18
4.1 Downmix Method Concept.....	18
4.2 Averaging Approach.....	18
4.3 Downmix Method based on Karhunen-Loève Transform	19
4.3.1 KLT-based Approach.....	19
4.3.2 Artifacts under KLT-based Approach.....	21
4.3.2.1 Tone Leakage Effect	21
4.3.2.2 Tone Modulation Effect	23
4.3.2.3 Pre- and Post-processing for KLT-based Approach.....	25
4.4 Downmix Method Summary.....	27
Chapter 5 Experiments.....	29
5.1 Experiment Environment	29
5.2 Objective Quality Measurement in MPEG Test Tracks.....	31

5.3 Objective Quality Measurement in Music Database	36
5.4 Subjective Quality Measurement	41
Chapter 6 Conclusion Remark and Future Works.....	42
References.....	43



Figure List

Figure 1: The quality comparison of AAC, HE-AAC (SBR), and Parametric Stereo Coding based on HE-AAC	1
Figure 2: An example for parametric stereo coding.....	2
Figure 3: Block diagram of the PS decoder based on QMF [10].....	3
Figure 4: Block diagram of the PS encoder with HE-AAC.....	4
Figure 5: Hybrid QMF analysis filterbank for the 10 and 20 stereo bands configuration [10]	5
Figure 6: Hybrid QMF analysis filterbank for the 34 stereo bands configuration [10]	5
Figure 7: Framing of the PS Coding.....	6
Figure 8: Block diagram of de-correlation procedure.....	7
Figure 9: Difference between two mixing procedures.....	8
Figure 10: Block diagram of the PS encoder including RD and SBD modules	10
Figure 11: Dynamic programming in Region Decision.....	14
Figure 12: $E_{i,j}^{(0)}$, the time region from time slot i to time slot j	14
Figure 13: Flow chart of DP in Region Decision.....	15
Figure 14: Complete flow chart of Region Decision.....	16
Figure 15: Stereo Band Decision in stereo band k	16
Figure 16: Flow chart of Stereo Band Decision.....	17
Figure 17: Energy cancellation problem in averaging approach	19
Figure 18: Flow chart of Karhunen-Loève Transform.....	20
Figure 19: Block diagram of the PS encoder including KLT module.....	21
Figure 20: Simple example for tone leakage	22
Figure 21: Reconstructed signal under averaging method for tone leakage	22
Figure 22: Reconstructed signal under KLT method for tone leakage	23
Figure 23: Example of tone modulation	24
Figure 24: The smooth methods of KLT coefficient vector.....	25
Figure 25: Linear smoothness of coefficient vectors.....	26
Figure 26: Cosine smoothness of coefficient vector.....	27
Figure 27: Block diagram of the PS encoder with KLT-based downmix method.....	28
Figure 28: The variance in the ODGs of proposed methods at 48 kbps	32

Figure 29: The variance in the ODGs of proposed methods at 36 kbps	33
Figure 30: The variance in the ODGs of proposed methods at 24 kbps	34
Figure 31: The average ODGs of method M0 and M1 at 48 kbps in 15 categories	37
Figure 32: The average ODGs of method M0 and M2 at 48 kbps in 15 categories	37
Figure 33: The average ODGs of method M0 and M3 at 48 kbps in 15 categories	37
Figure 34: The average ODGs of method M0 and M1 at 36 kbps in 15 categories	38
Figure 35: The average ODGs of method M0 and M2 at 36 kbps in 15 categories	38
Figure 36: The average ODGs of method M0 and M3 at 36 kbps in 15 categories	38
Figure 37: The average ODGs of method M0 and M1 at 24 kbps in 15 categories	39
Figure 38: The average ODGs of method M0 and M2 at 24 kbps in 15 categories	39
Figure 39: The average ODGs of method M0 and M3 at 24 kbps in 15 categories	39
Figure 40: The subjective test result for PS coding at 36 kbps.....	41

Table List

Table 1: Two region types for error calculation	12
Table 2: The relative threshold for DP under three target bit-rates.....	15
Table 3: The twelve tracks recommended by MPEG	31
Table 4: Objective measurements through the ODGs for proposed methods at 48 kbps	32
Table 5: Objective measurements through the ODGs for proposed methods at 36 kbps	33
Table 6: Objective measurements through the ODGs for proposed methods at 24 kbps	34
Table 7: The PSPLab audio database [25]	36



Chapter 1

Introduction

During the last decade, there are many psycho-based perceptual audio coding. MPEG Layer-1 3 (MP3) [1] is one of the world's well-known codec. Furthermore, MPEG-4 Advance Audio Coding (AAC) [2] is popularized currently and will become the next-generation audio codec. However, to string along the vogue of mobile usage, there is congenital deficiency in psycho-based perceptual audio coding. As we know, the mobile usage demands the low bit-rate to increase its capacity no matter in transmission or in disk. Further, it does not need the high quality because of its using environment. For the purposes of resolving the above demands, some approaches for low bit-rate perceptual audio coding are created. Spectral Band Replication (SBR) [3][4][5][6][7] is one of these new audio coding enhancement tools. The idea of SBR is to replace the high frequency spectrum by the low frequency spectrum which is encoded by AAC. Under this idea, it only uses about half the bits of AAC with some extra information for the high frequency reconstruction. Thus, the bit-rate is decrease but the quality is kept. The research for SBR indicates the coding quality under 80 kbps for the stereo signal is better than AAC, extremely equals to 96 kbps of AAC.

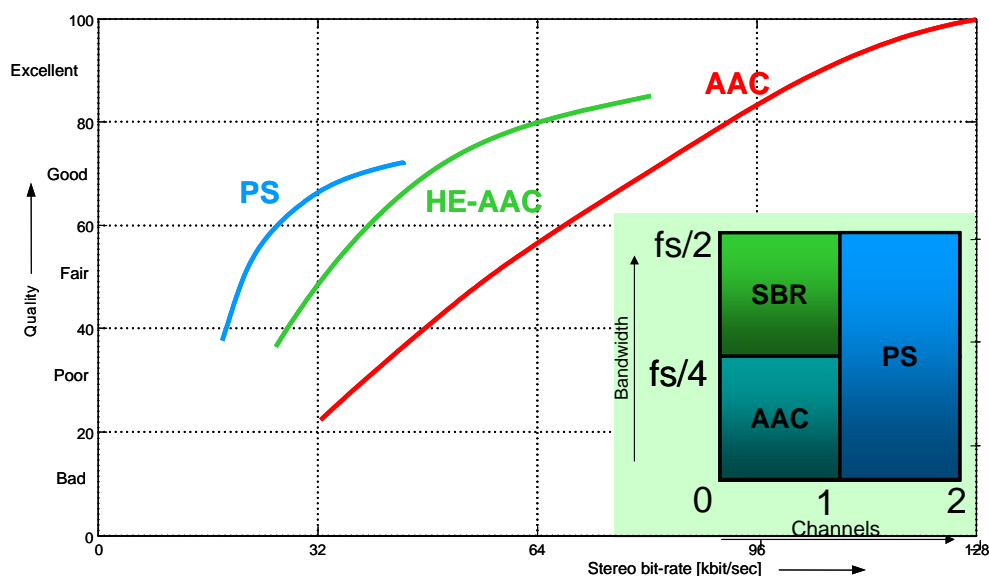


Figure 1: The quality comparison of AAC, HE-AAC (SBR), and Parametric Stereo Coding based on HE-AAC

Parametric Stereo Coding (PS) [8][9] goes a step further to achieve the demand of low bit-rate usage. It is also an audio coding enhancement tool. The originality of parametric stereo coding is to combine the stereo signal into a monaural signal with some stereo images. That is, encoding a stereo signal by PS tool uses almost half of the bits used in other codec like HE-AAC with small extra bits for stereo images. Also, according to the human hearing, the stereo images can be recorded as the intensity level difference between left and right band-limited signals as inter-channel intensity difference (IID), the similarity as inter-channel coherence (ICC), and the phase behavior as inter-channel and overall phase differences (IPD and OPD). The PS decoder then uses the monaural signal and these stereo images to reconstruct the stereo signal. Figure 1 shows the coding efficiency of AAC, HE-AAC (SBR), and parametric stereo coding based on HE-AAC. The parametric stereo coding extends the high quality audio coding of AAC and HE-AAC to bit rates 24-48 kbps. Figure 2 illustrates an example for encoding and decoding in parametric stereo coding.

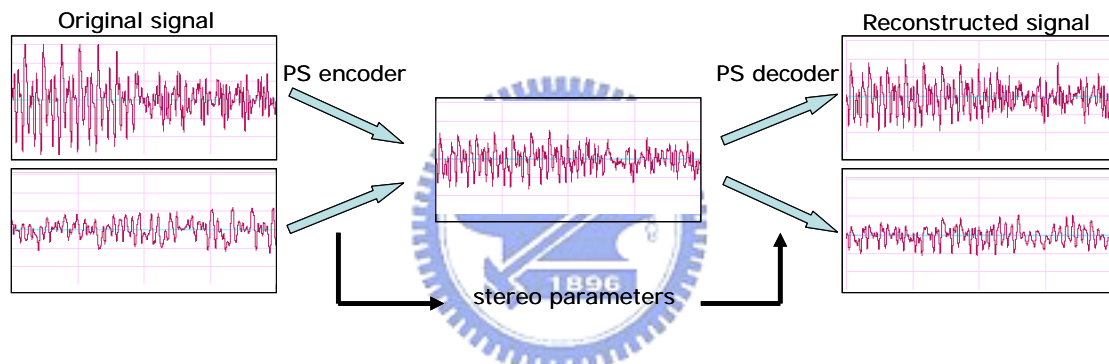


Figure 2: An example for parametric stereo coding

The target of this thesis focuses on T/F stereo parameter extraction and downmix method of the parametric stereo encoder. Chapter 2 will firstly introduce the basic idea and some components of parametric stereo coding. Chapter 3 discusses the approach of T/F stereo parameter extraction module which is designed to control the stereo parameter sets. Chapter 4 presents the method to generate the monaural signal which is done in downmix method. After all, the experiments are conducted in Chapter 5 to verify the audio quality and efficiency of our approaches. In Chapter 6, the thesis gives the conclusion for our design.

Chapter 2

Basic Components

2.1 Parametric Stereo Coding Concept

The Parametric Stereo coding (PS) is a tool in the MPEG-4 audio parametric coding scheme for compressing high quality stereo audio at bit rates around 24 kbps. From the original stereo input signal and the monaural downmix of the stereo input signal generated by the parametric stereo coding tool, the PS module extracts the stereo parameter sets. The parametric stereo decoding can reconstruct the stereo signal by using the monaural downmix signal with the delivered stereo parameters. Figure 3 illustrates the PS decoder based on QMF. The detail of reconstruction idea in PS decoder will be presented in the later sections.

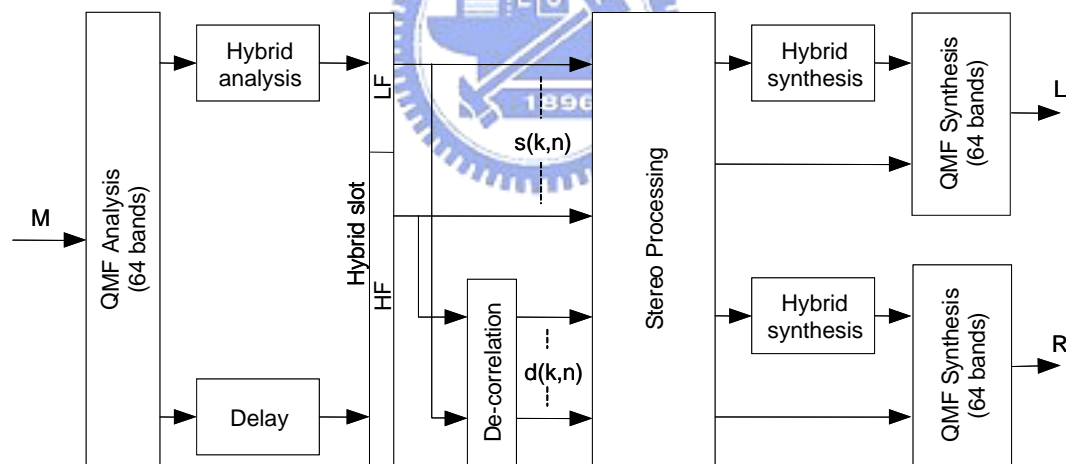


Figure 3: Block diagram of the PS decoder based on QMF [10]

For the encoding of the monaural downmix signal, it can operate in combination with any monaural coder such as SSC [10] or MPEG-4 HE-AAC (SBR). In this thesis, the design of PS encoder is built on HE-AAC. The block diagram of the PS encoder with HE-AAC is illustrated in Figure 4. The downmix method and T/F stereo parameter extraction, which controls the hybrid filter and stereo parameters calculation module, are the focal point of this thesis.

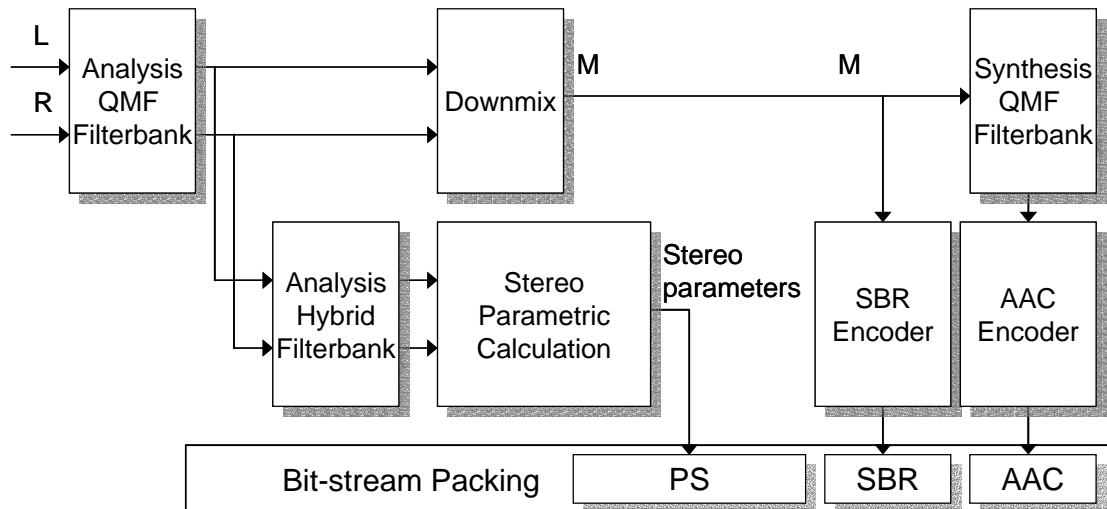


Figure 4: Block diagram of the PS encoder with HE-AAC

Before discussing our design in PS encoder, this chapter gives the basic idea and components for stereo reconstruction in PS decoder.

2.1.1 Hybrid Filterbank

In Figure 3, the reconstructed monaural signal is firstly passed into QMF analysis filterbank, which is defined in ISO/IEC 14496-3/AMD1:2003, subclause 4.B.18.2. As we know, each subband of QMF analysis filterbank has the same bandwidth. This does not conform to the psychoacoustic model which the lower bands have narrow bandwidth, it means the lower bands have high frequency resolution, and the higher bands have wide bandwidth. QMF exactly has the opposite property. Neither the lower subbands nor the high subbands of QMF analysis filterbank, the frequency resolution can not response the sense of hearing. Therefore, the PS Coding uses a hybrid filterbank to procure this property. Figure 5 and Figure 6 are two type of hybrid configuration depending on the number of stereo bands. After the filter, the lower QMF subbands are split into many sub-subbands to obtain a higher frequency resolution. The PS tool may split the 64 QMF subbands to 71 or 91 sub-subbands and rearrange these sub-subbands into 10, 20, or 34 stereo bands. The stereo band is the basic unit of frequency domain in PS Coding.

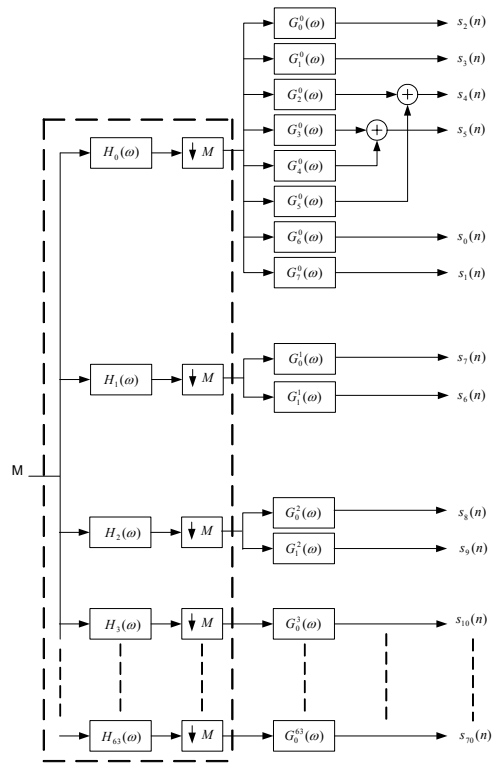


Figure 5: Hybrid QMF analysis filterbank for the 10 and 20 stereo bands configuration [10]

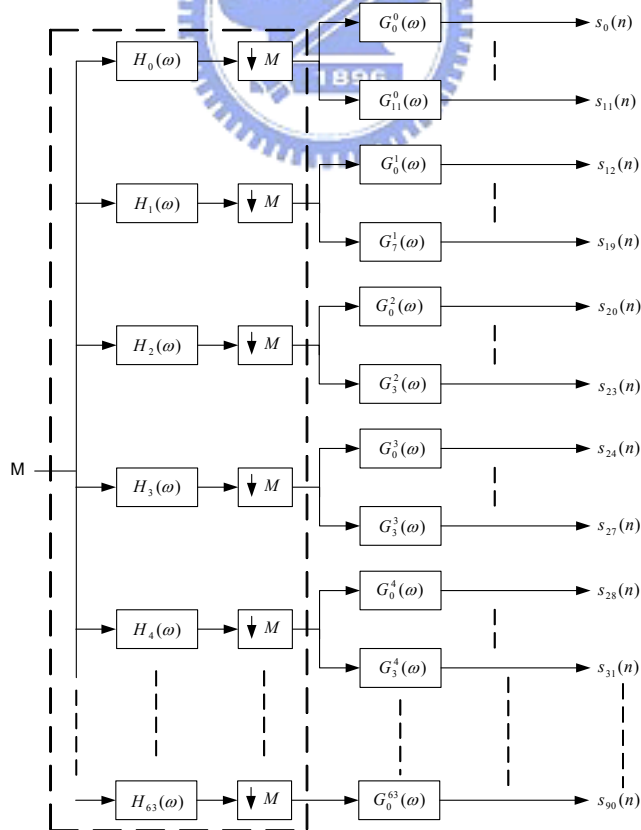


Figure 6: Hybrid QMF analysis filterbank for the 34 stereo bands configuration [10]

2.1.2 Framing

In the previous section, we see the definition of frequency resolution using in the PS Coding. Moreover, there are some restrictions for the resolution of time domain. In a PS frame, there are 32 time slots in time domain and 10, 20, or 34 stereo bands in frequency domain. Stereo parameters in one frame can be assigned to at most four time slots, called time borders. Each time border can be the location for updating the stereo parameters. “Region” is defined as the part between two time borders or the frame boundaries. As Figure 7 illustrates, there are four time borders for updating of stereo parameters and five regions in the frame F_i . Because the stereo parameters only can be assigned to some time slots, the non-assigned time slots use interpolation to obtain their stereo parameters. The bold solid lines in the figure illustrate the interpolation of stereo parameter. For the very first region, it respects to the default stereo parameters. Specially, for the case like region 5 of the frame F_i , in which there is no updating of stereo parameters, each time slot of the region obtains its stereo parameters from the latest time border. Thus, in this kind of region, all un-assigned time slots will get the same stereo parameters according to the latest time border.

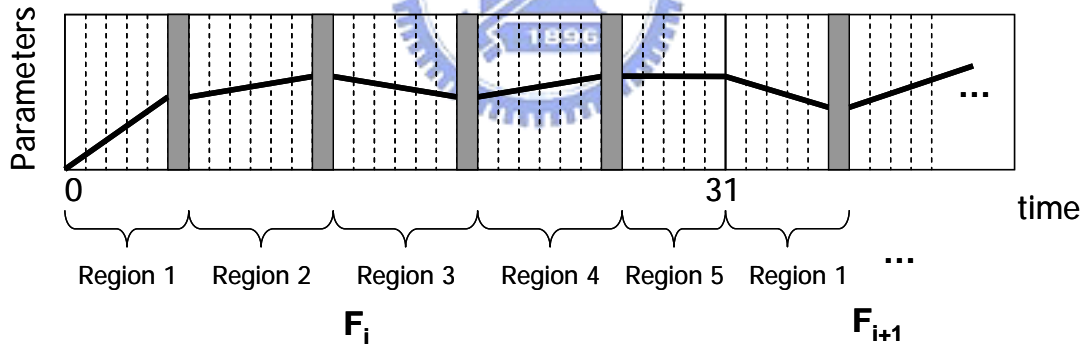


Figure 7: Framing of the PS Coding

2.1.3 De-correlation Signal

The PS decoder reconstructs the stereo signal based on the reconstructed monaural signal and the de-correlation signal. In the related researches [11][12], the signal which is mixed with the monaural signal is often a white noise signal. The noise signal is expected to have the zero cross-correlation property to the monaural signal. This property is useful to get the correlation between the original stereo channels by mixing different percentage of noise signal in two reconstructed

channels. However, one disadvantage for using the noise signal as the mixed signal. Is that the noise signal does not have the same time-envelop in monaural signal. If mixing these two signals, the time-envelop of the reconstructed signal will be destroyed.

To avoid this problem, PS Coding uses the de-correlation procedure to get the mixed signal. The mixed signal is based on monaural signal. Because the zero cross-correlation property is the basic idea for reconstructing signal and time-envelop must be kept, the de-correlation process is using the all-pass filter to randomize the phase in monaural signal. Besides, transients and other fast time-envelopes must be pre-processed. The transient detection and smooth is used before the filter in PS Coding. Thus, the de-correlation procedure in PS Coding can hold the similar time-envelop to monaural signal and also gives the zero cross-correlation property. Figure 8 illustrates the block diagram in which m , d separately indicates the monaural signal and de-correlation signal, k is sub-subband index, and n is time domain index.

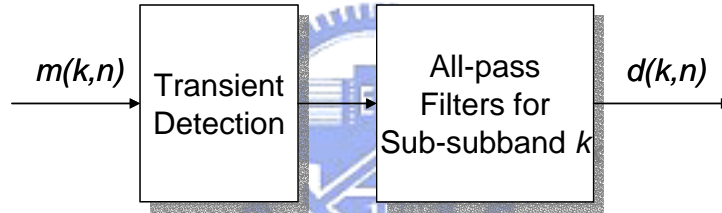


Figure 8: Block diagram of de-correlation procedure

2.1.4 Stereo Processing

The stereo processing module, which is also called mixing procedure in the PS Coding, is to reconstruct the stereo signal by the downmix monaural signal, the de-correlation signal, and the delivered stereo parameters. For each sub-subband, they are reconstructed as:

$$\begin{aligned} l_k(n) &= H_{11}(k,n)m(k,n) + H_{21}(k,n)d(k,n) \\ r_k(n) &= H_{12}(k,n)m(k,n) + H_{22}(k,n)d(k,n) \end{aligned} \quad (1)$$

In (1), $m(k,n)$ is the n -th sample of the downmix monaural signal in sub-subband k , and d is indicated the de-correlation signal. Consequently, H is the mixing matrix and the mixing procedure calculates this matrix by the delivered stereo parameters. From the de-correlation procedure, the correlation between m and d is nearly zero and their energies are the same because of the all-pass filter. Thus, to keep the energy and correlation in reconstructed signal, H is demanded the restriction as:

$$\left\{ \begin{array}{l} H_{11}^2 + H_{21}^2 = \frac{2l^2}{l^2 + r^2}, \\ H_{12}^2 + H_{22}^2 = \frac{2r^2}{l^2 + r^2}, \\ \frac{H_{11}H_{12} + H_{21}H_{22}}{\sqrt{H_{11}^2 + H_{21}^2}\sqrt{H_{12}^2 + H_{22}^2}} = \rho \end{array} \right. \quad (2)$$

where l^2 and r^2 are the energy of left and right channel, ρ is the correlation of stereo signal. If the energy of the monaural signal is the average energy of two channels, the first two restrictions of H can guarantee the conservation of energy for two channels.

In PS draft [10], there are two mixing procedures, which are named R_a and R_b , are introduced. Both two approaches comply with the restriction of mixing matrix and reconstruct the signal to keep the original energy ratio and correlation between the stereo channels. The difference between the two approaches is the strategy for using the de-correlation signal. To compare H_{21} and H_{22} under the two approaches, there is a negative sign difference. Thus, the de-correlation signal effects on two channels under two approaches are reversed. Figure 9 illustrates this difference. The x axis denotes the energy ratio, y axis is correlation of signal, and z axis indicates the value of mixing matrix.

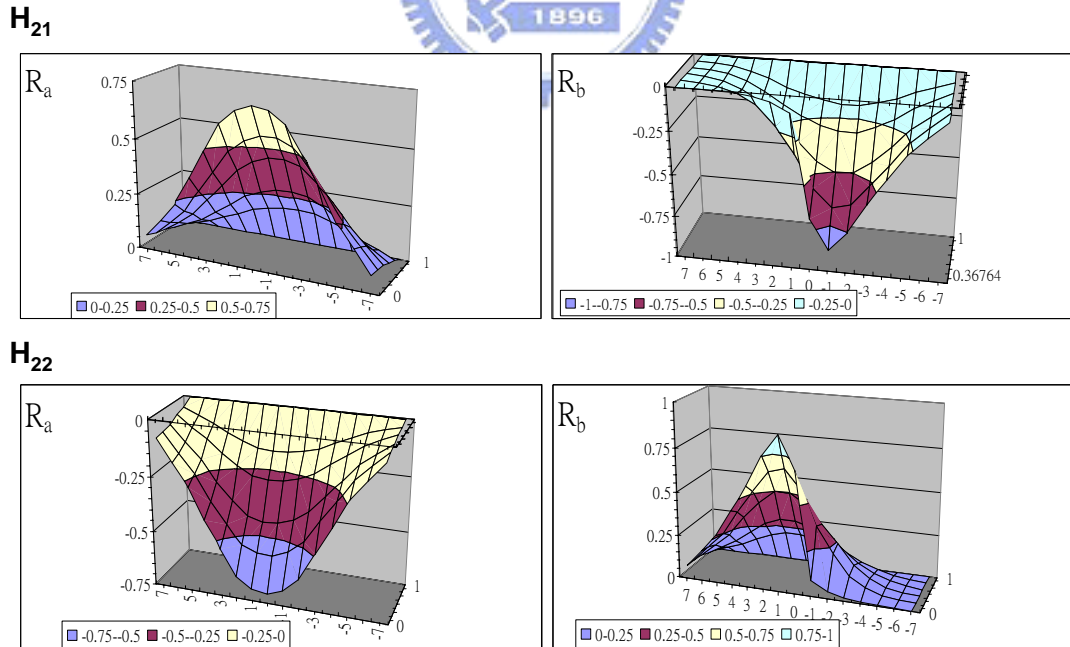


Figure 9: Difference between two mixing procedures

Chapter 3

T/F Stereo Parameter Extraction

3.1 T/F Stereo Parameter Extraction Concept

As mentioned above, in accordance with the information in delivered stereo parameters the PS tool reconstructs the stereo signal from monaural signal. In each frame, there are 32 time slots in time domain and many stereo bands in frequency domain. This T/F resolution is too high to deliver the stereo parameters in each frame. Consequently, the PS draft [10] limits the number of stereo parameter delivery as discussion in chapter 2. In each frame, there can be at most four time borders for time domain and three types of stereo band for frequency domain can be selected. In other words, the number of stereo parameter sets send in one frame is from zero, when there is no time border in the frame, to 136, which is under four time borders and 34 stereo bands. The design issue of T/F stereo parameter extraction module is the way to decide the setting of time borders and stereo band.

3.2 Existing Approaches in Other Codecs

In the literature, there have been limited publications on the stereo parametric extraction. The code in 3GPP [13] uses basically a fixed stereo parameter set, which always uses fixed stereo band resolution with two time regions, each region is half of frame, in each frame. Similarly, to dump the encoded tracks which are encoded by NERO 7.0.8.2 [14] or Coding Technologies 7.0.5 [15] can be found out these tracks always use fixed stereo band resolution with only one time regions in each frame. All these codecs do not make adaptive decision on the T/F stereo parameter extraction. That is, they only update the stereo parameter sets regularly without any decision. Because this approach does not reflect the content of signals, it might have some disadvantages. For stationary signal, because the signal is steady, the stereo parameters in same time slot should be similar to the previous ones. Under this situation, there could be no time border assigned to these kind frames. Thus, each frame follows the stereo setting in the previous frame. Nevertheless, the fixed stereo

parameter extraction for stationary signal updates the stereo parameters too often and so the bits used here are wasted. On the contrary, the transient signal or varied signal might need many stereo parameter sets to record the variation of the signal. The merit of the fixed approach is the low complexity but the method has not reflects the signal contents.

3.3 Adaptive T/F Stereo Parameter Extraction

As illustrated in Figure 10, the thesis segments “T/F Stereo Parameter Extraction” into two modules: “Region Decision” (RD) for time domain and “Stereo Band Decision” (SBD) for frequency domain which adapt to the signal content. Both region decision and stereo band decision are the control modules. In the block diagram, their outputs are dotted lines to original PS encoder for controlling the relative modules. Region decision is designed to find the time borders for updating the stereo parameter sets. The other module, Stereo band decision, decides the frequency resolution and informs “Hybrid Filterbank Analysis” module to split lower QMF subbands to achieve a higher frequency resolution. These two modules decide the T/F resolution in a frame.

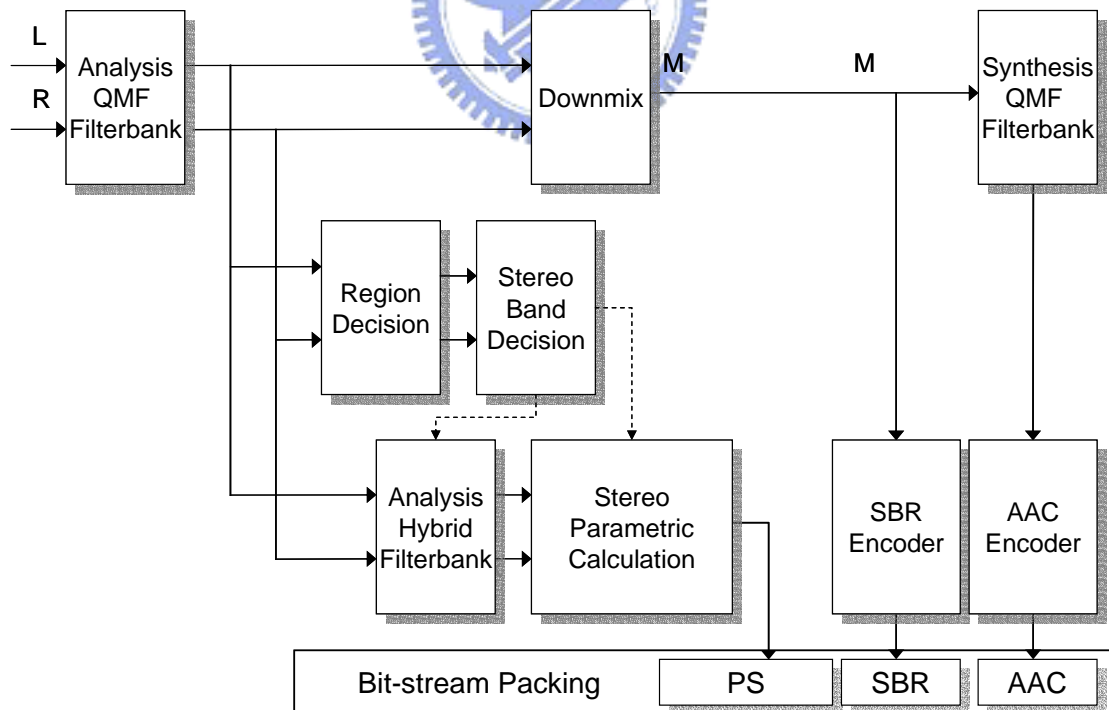


Figure 10: Block diagram of the PS encoder including RD and SBD modules

3.3.1 Region Decision

First, we consider the way the PS decoder reconstructs the stereo signal by using the monaural signal and the delivered parameter sets. In each stereo band of a time region, there is only one stereo parameter set assigned to the last time slot in the region and the parameter sets of other time slots in the region are assigned by means of interpolation. Our algorithm is to find the time regions and its border positions by the content of stereo signal with a well-controlled reconstruction error.

3.3.1.1 Stereo Parameter Estimation and Error Calculation

Before searching time regions in a frame, the stereo parameter for each data sample, which is in a subband of a time slot, should be built. Because the stereo parameter is a statistical characteristic, only one sample can not be extracted with the stereo parameters. The estimation method here is to window the nearly data samples for obtaining the approximate stereo parameters. For example, to calculate an approximate inter-channel intensity difference (IID), the formula is

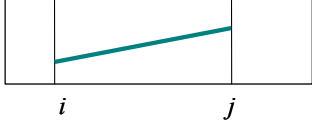

$$IID(t, b) = 10 \log_{10} \frac{\sum_{n=0}^{L-1} \left| w(n) \cdot l\left(t + n - \left\lfloor \frac{L}{2} \right\rfloor + 1, b\right) \right|^2}{\sum_{n=0}^{L-1} \left| w(n) \cdot r\left(t + n - \left\lfloor \frac{L}{2} \right\rfloor + 1, b\right) \right|^2}. \quad (3)$$

t indicates time index and b is frequency index. w is the window sequence with length L . l and r are the left and right channels. After parameter estimation, because there are different types of stereo parameters, the same kind of estimated stereo parameters should be normalized firstly. Thus, each type of stereo parameter has zero mean and equal standard derivation and so different type of parameter can be averaged together. The normalized formula is

$$\tilde{x}(t, b) = \frac{x_{t,b} - \text{mean}(x(t, b))}{\sigma(x(t, b))}, \quad (4)$$

where x means a type of stereo parameter, and $\sigma(x_{t,b})$ is the standard derivation of $x_{t,b}$. By this estimation, the reconstruction error can be calculated in each possible region. There are two types of parameter generating for unassigned time slot in a region. If a time region has time borders as its boundaries, the parameter generating is by means of linear interpolation. However, a region also can be defined as the space between time border and frame boundary, the parameters for unassigned time slots are same as the previous time border until the end of frame. These two region types are arranged in Table 1, where $\Delta_{i,j}(b)$ indicates the linear interpolation slope for type-I.

Table 1: Two region types for error calculation

	Type-I: Linear interpolation	Type-II: constant interpolation
Description	Region between two time borders	Region between time border and frame boundary
Figure		
Error Calculation	$e_{i,j} = \sum_b \sum_{l=0}^{L-2} x(i+l,b) - x(l-1,b) - \Delta_{i,j}(b) \times (l+1) $	$e_{i,j} = \sum_b \sum_{l=0}^{L-2} x(i+l,b) - x(l-1,b) $

By the reconstruction error of each possible region, any given region set can be used to calculate the reconstruction error. First, the search space can be defined as

$$\Gamma = \{RS^{i_1, i_2, i_3, i_4} \mid 0 \leq i_1 \leq i_2 \leq i_3 \leq i_4 \leq 31\}, \quad (5)$$

where RS indicates a region set which contains time borders at i_1 , i_2 , i_3 , and i_4 . Any region set conforms to this space is a possible solution for region decision. Also, the objective function for a region set is given as

$$\Phi(RS^{i_1, i_2, i_3, i_4}) = \sum_{k=0}^4 e_{i_k+1, i_{k+1}}, \quad (6)$$

which means the error summation of all regions in this region set. By above formulas, the optimal solution is

$$RS^{i_1^*, i_2^*, i_3^*, i_4^*} = Arg \left\{ \min_{0 \leq i_1 \leq i_2 \leq i_3 \leq i_4 \leq 31} [\Phi(RS^{i_1, i_2, i_3, i_4})] \right\}. \quad (7)$$

Unfortunately, using the ‘‘Bruce-Force Method’’ to search the optimal solution is too complexity. The number of total possible region sets is

$$1 + C_1^{32} + C_2^{32} + C_3^{32} + C_4^{32} = 41449. \quad (8)$$

Therefore, an efficient search algorithm should be adopted for solving this problem. The thesis will introduce a method based on dynamic programming to substantially decrease the coding complexity.

3.3.1.2 Dynamic Programming

Dynamic programming (DP) is a subset of the general theory concerned with discrete sequential decisions. The varied history and an extensive bibliography can be found in the article by Silverman and Morgan [16]. The prolific application of the DP to various fields has been credited to Professor Richard Bellman [17]. The

basic principle of DP is to break an optimization problem down into stages of decisions that follow a criterion leading to a recurrence relations. For audio compression, we have already applied the DP algorithm to efficiently design the MS coding, and Huffman code book search [18]. This thesis will consider the applying of the DP algorithm to efficiently search the time borders.

3.3.1.3 Region Decision based on Dynamic Programming

Applying dynamic programming to region decision, we need to define the problem into a decision problem and break the problem into stages of the recursive decision of sub-problems. Above all, we define the symbols as follow

- s_i i -th inner time borders.
- $e_{i,j}^{s_1, s_2, \dots, s_k}$ the reconstruction error of parameters in the range from time slot i to time slot j under the time region set with k inner time borders and there is a border at time slot $i-1$.
- $E_{i,j}^{(k)}$ the minimum construction error of parameters in the range from time slot i to time slot j among all possible time region sets with k inner borders and there is a border at time slot $i-1$.

By above symbol definitions, the minimum construction error can be written as

$$E_{i,j}^{(k)} = \min \{ e_{i,j}^{s_1, s_2, \dots, s_k} \mid i \leq s_1 < s_2 < \dots < s_k < j \}, \forall k > 0. \quad (9)$$

Furthermore, let $E_{i,j}^{(0)}$ for the case whose no inner border. The optimum sub-structure of $E_{i,j}^{(k)}$ can be explored as follow. Assume the optimum k borders are s_1', s_2', \dots, s_k' , we have

$$E_{i,j}^{(k)} = E_{i,s_1'}^{(0)} + e_{s_1'+1,j}^{s_2', \dots, s_k'} \quad (10)$$

By the definition of $E_{s_1'+1,j}^{(k-1)}$, it gives

$$E_{i,j}^{(k)} = E_{i,s_1'}^{(0)} + e_{s_1'+1,j}^{s_2', \dots, s_k'} \geq E_{i,s_1'}^{(0)} + E_{s_1'+1,j}^{(k-1)} \quad (11)$$

Since $E_{i,j}^{(k)}$ is the optimum solution, the equality must hold,

$$E_{i,j}^{(k)} = E_{i,s_1'}^{(0)} + E_{s_1'+1,j}^{(k-1)} \quad (12)$$

Hence, to inspect all the possible s_1' , $E_{i,j}^{(k)}$ is determined in (13) and Figure 11 illustrates this condition.

$$E_{i,j}^{(k)} = \min_{t \in \{i, i+1, \dots, j\}} \{ E_{i,t}^{(0)} + E_{t+1,j}^{(k-1)} \} \quad (13)$$

Number of border = k

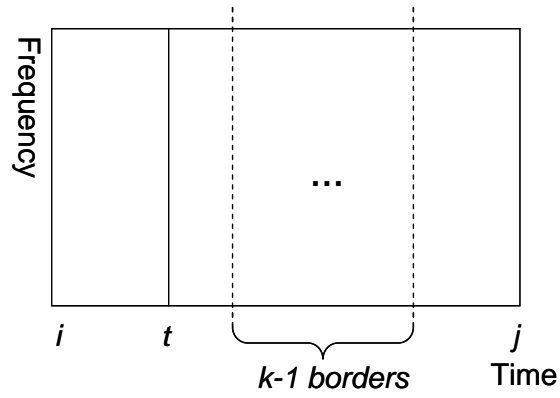


Figure 11: Dynamic programming in Region Decision

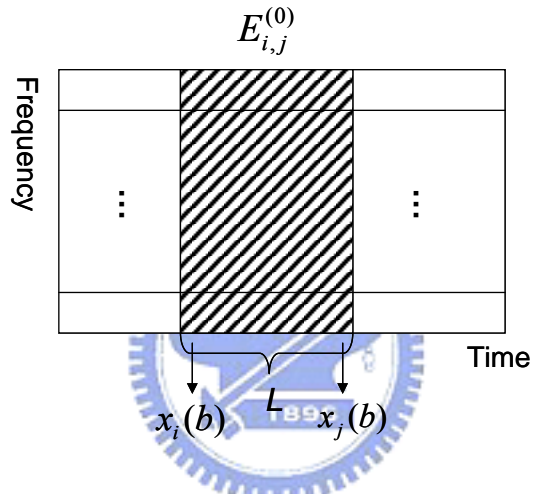


Figure 12: $E_{i,j}^{(0)}$, the time region from time slot i to time slot j

Each $E_{i,j}^{(k)}$ can be recursively constructed for all $k > 0$. Therefore, we need to calculate $E_{i,j}^{(0)}$ for all i and $j > i$ at initialization of dynamic programming. Figure 12 illustrates $E_{i,j}^{(0)}$, where b is the QMF subbands index, x is the variable which indicates the stereo parameters value, and L is the length of $E_{i,j}^{(0)}$.

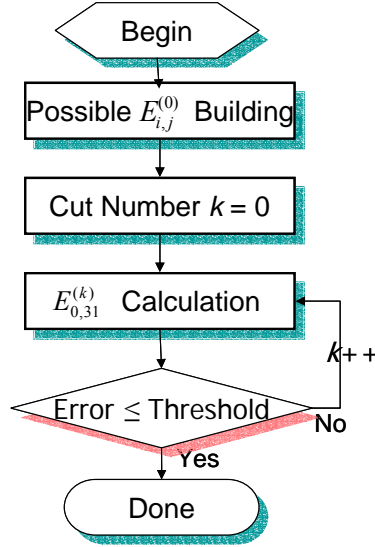


Figure 13: Flow chart of DP in Region Decision

The flow chart of dynamic programming is illustrated in Figure 13. The dynamic programming algorithm calculates the minimum errors from the case without inner border, up to the case with four inner borders in the frame. Also there is a termination threshold shown in Figure 13. The number of regions increases with the decrease of reconstruction error with overhead from the bits usage in a frame. By this reason, the optimal solution might lead to too much bit overhead. That is to say, although searching the minimum reconstruction error gives the optimal solution for region set, the coding quality might be influenced. Under the requirement of quality and the limited available bits, the threshold is useful to provide this condition. Therefore, instead of finding the regions with minimum reconstruction error, the algorithm uses a threshold to stop the DP search and finds the wanted region set with the minimum reconstruction error among all region sets under the smallest region number in which the error is firstly less than the threshold. Moreover, this threshold must relate with bit-rates. By our tuning result, the thesis suggests three thresholds under target bit-rates: 48, 36, and 24 kbps. For other bit-rates, it uses interpolation method to get relative threshold.

Table 2: The relative threshold for DP under three target bit-rates

Bit-rate (kbps)	48	36	24
Relative Threshold	1.4	1.5	1.6

However, this threshold needs to be conservative, and hence there could be risky in some tracks. This thesis suggests an aggressive threshold for avoid the risk. If there are successive frames which have high reconstruction errors but can not achieve the

threshold, our algorithm will not update the stereo parameters in these frames. The aggressive threshold is used here to coercively update the stereo parameters to avoid this situation. Another method is to update parameters regularly. This method also can solve the seek problem for playback usage. Figure 14 illustrates the complete flow chart of region decision based on the dynamic programming.

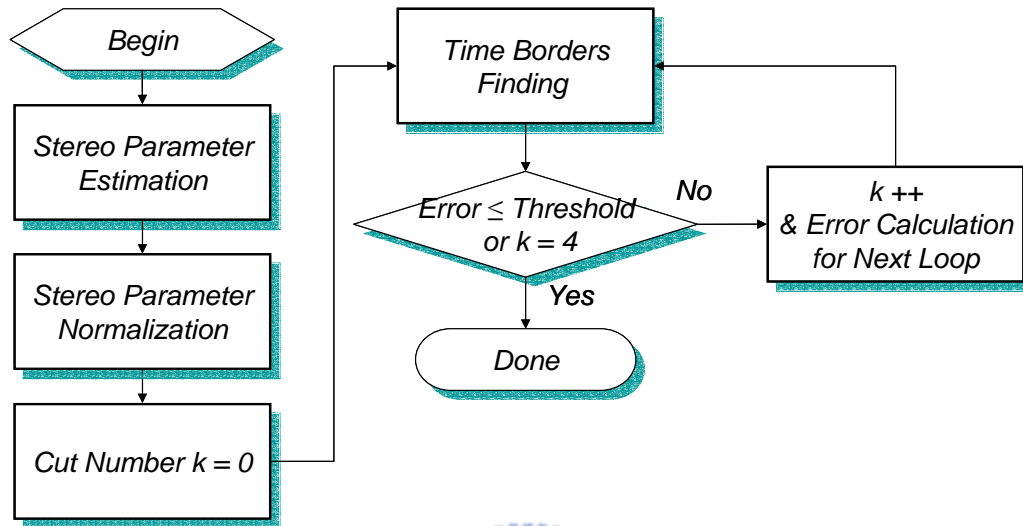


Figure 14: Complete flow chart of Region Decision

3.3.2 Stereo Band Decision

In stereo band decision module, it controls the hybrid analysis filterbank to decide the resolution for frequency domain. The 64 QMF subbands might be divided to 71 or 91 sub-subbands and combined into 10, 20, or 34 stereo bands. Since only one stereo parameter set can be updated for a stereo band in a time region, the QMF subbands in the same stereo band are supposed to share same stereo parameter set. That is, these QMF subbands should have similar stereo characteristic. By this viewpoint, the similarity of subbands should be measured.

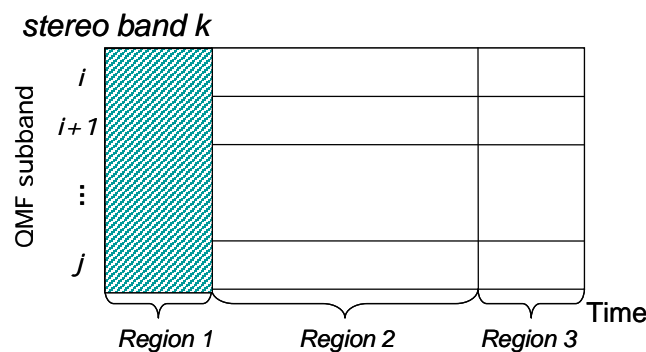


Figure 15: Stereo Band Decision in stereo band k

Figure 15 illustrates a stereo band k , which contains QMF subbands which are indexed as $i, i+1, \dots$ to k . In this example, there are three time regions in stereo band k . For each time region and QMF subband, it firstly estimates a stereo parameter set for region T and QMF subband b . Like the discussion in region decision, the stereo parameters here also need to be normalized. The normalized method is same as (4). Using these normalized parameters, SBD module calculates their variance in each stereo band containing more than one QMF subband. If the variance in this stereo band is small enough, this QMF subbands in this stereo band have the similar stereo parameters and can be combined. Therefore, these variances are the measurement for similarity. Finally, to sum up these variances under different types of stereo band resolution, the smallest one indicates that this frequency resolution is the most suitable for content of signal. Figure 16 illustrates the flow chart of SBD module.

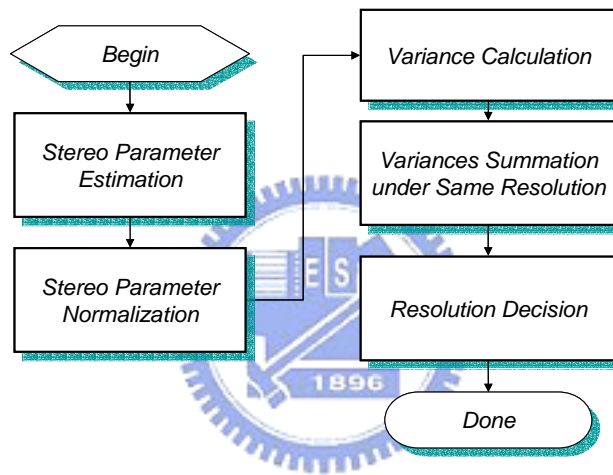


Figure 16: Flow chart of Stereo Band Decision

3.4 T/F Stereo Parameter Extraction Summary

From above sections, the thesis introduces the existing method, which uses the fixed stereo parameter sets, and suggests an adaptive T/F stereo parameter extraction to avoid the lack of existing method. This adaptive method has been already published in [19]. The quality measurement will be shown in Chapter 5 to assess the improvement of this method.

Chapter 4

Downmix Method

4.1 Downmix Method Concept

The purpose of PS coding is to combine the stereo signal into a monaural signal with some stereo parameters. Because of that, bit-rate can be greatly reduced. Moreover, in such low bit-rate, only few bits provide PS usage, the most bits suppose the encoding of monaural signal to keep the certain coding quality. Therefore, the monaural downmix signal is the essential quality source for the reconstructed stereo signal. The design issue of downmix method seems to be the way to preserve the information of the original stereo signal. Furthermore, in Chapter 2 the mixing procedure suggests some restrictions of the downmix signal. In downmix method, it should also consider these restrictions. Following sections will firstly discuss the averaging approach and then suggest an approach to avoid the problems in the averaging approach.

4.2 Averaging Approach

In PS draft [10], the monaural downmix signal is generated according to

$$M = \frac{L + R}{2}. \quad (14)$$

This approach only calculates the average signal of the stereo input signal. Therefore, there might be energy cancellation problem in the downmix signal illustrated in Figure 17. If two channels are anti-phase and have same magnitude, the monaural downmix signal will be totally cancelled. Under this energy cancellation, the PS tool can not reconstruct the original signal. In other words, the stereo information between two channels might be lost. Also this problem violates the restriction of mixing procedure which demands the energy of downmix signal is the average energy of the stereo signal to keep the conservation of energy.

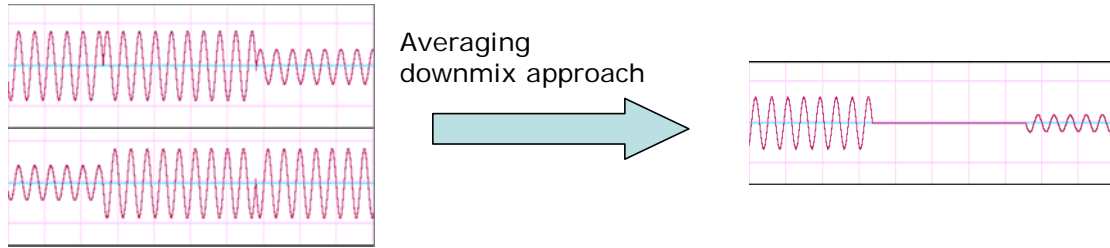


Figure 17: Energy cancellation problem in averaging approach

The method adopted in 3GPP [13] is based on the averaging approach. It uses a post-processing to cater for the demand of mixing procedure. An energy adjusting scale is used in the post-processing. This method seems to procure the mixing restriction, but it also can not save the stereo information which is disappeared in the average signal. Besides, the destroyed spectrum structure which faces cancellation problem is still same as before when using scaling method. Therefore, it is easy to implement the average approach but this approach which does not conform to the signal content would wreck the signal structure.

4.3 Downmix Method based on Karhunen-Loève Transform



As mentioned above, the design issue of the downmix method is to preserve the most information of original stereo signal. In the following sections, the thesis will introduce an approach based on the Karhunen-Loève Transform.

4.3.1 KLT-based Approach

Our goal here is to generate the features via linear transforms of the stereo signal. The basic concept is to transform a given set of samples to a new set of features. Karhunen-Loève Transform (KLT) [20] is such an approach. Its transform domain features can exhibit high information packing properties. This means that the most of the classification-related information is squeezed in a relatively small number of features, leading to a reduction of the necessary feature space dimension.

To adopt KL transform as the downmix coefficient vector, it firstly defines some symbols as follow:

$U_{2 \times 32}$	sample matrix which contains samples in two relative subbands
$V_{2 \times 32}$	transformed matrix

R_U, R_V auto-correlation matrixes of sample matrix and transformed matrix
 $\Phi_{2 \times 2}$ KL transform matrix

By above symbol definitions, the transform equation can be written as:

$$V = \Phi^T U \quad (15)$$

The KL transform matrix Φ is given by orthonormal eigenvectors of R_U . After KL transform, R_V will be an uncorrelated matrix. Because the transform is used for downmix usage, the resultant sample set is the row in V with the eigenvector which corresponds to the large eigenvalue. In other words, the eigenvector with the large eigenvalue is the downmix coefficient vector to generate the monaural signal. Therefore, it is the optimal transform in terms of energy compaction because it makes basis vectors uncorrelated and orthogonal. Figure 18 shows a simple view for KLT flow path. It firstly build auto-correlation matrix from two relative subbands and then calculate the wanted eigenvector. The eigenvector is delivered to downmix method for generating the monaural signal. The block diagram which illustrated the PS encoder with KLT module is shown in Figure 19.

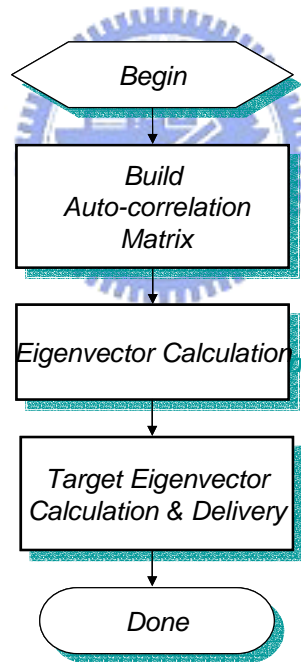


Figure 18: Flow chart of Karhunen-Loève Transform

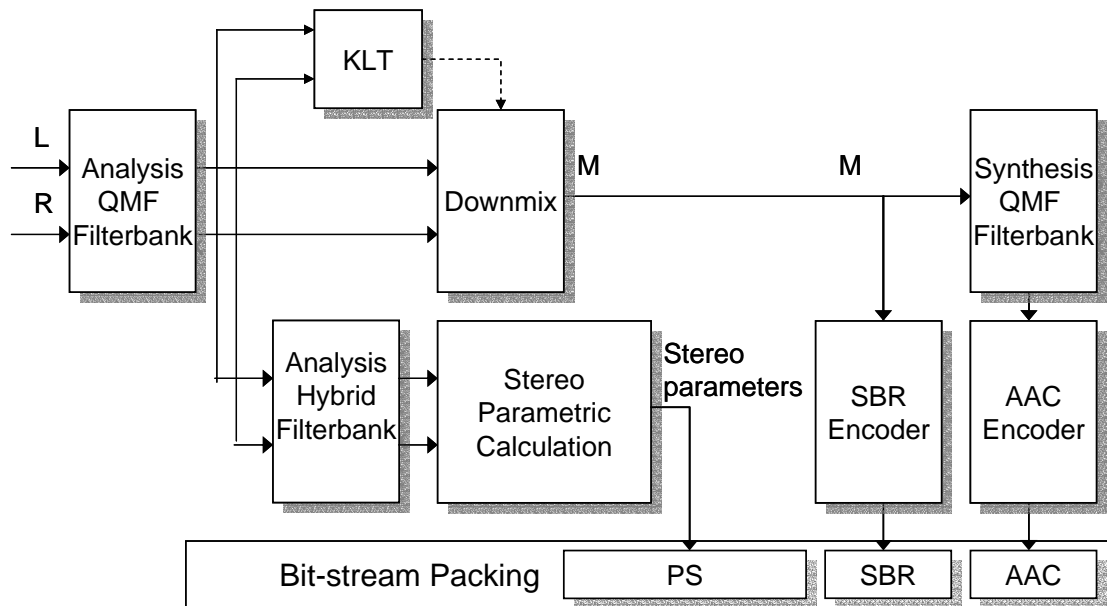


Figure 19: Block diagram of the PS encoder including KLT module

4.3.2 Artifacts under KLT-based Approach

To achieve the advantage of energy compactness, the KLT method suffers more risks than the simple average method. The feature that the weaker signal component is usually discarded and the adaptive property of signal combination coefficients are the main causes of the unwanted artifacts from the KLT method. The following subsections focus on the two critical artifact phenomena which named “Tone Leakage” effect and “Tone Modulation” effect.

4.3.2.1 Tone Leakage Effect

Here the two types of the “tone leakage” effect are defined to describe the two difference phenomena caused from the downmixing process. At first, the type-I tone leakage effect is defined to specify the phenomenon that one tone in some channel leaks to another channel after the upmixing procedure. This is an inherent artifact of downmixing coding, and hence both the average and the KLT methods suffer unavoidably. On the other hand, another situation is that one tone entirely or almost disappears in the both channels. The situation is named as the type-II tone leakage effect. Both the two method have the kind of artifact due to the different causing reason. However, the type-II tone leakage effect is usually inseparable from the KLT method. To show the tone leakage effect, a simple example is introduced and we compare the resultant severity between the two methods.

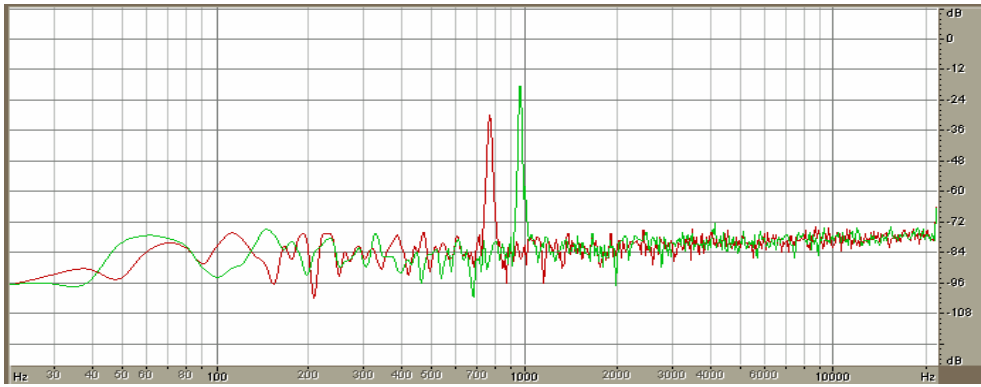


Figure 20: Simple example for tone leakage

In Figure 20, the left channel and right channel has a tone of different frequency respectively and there is a magnitude difference of 12 dB between two tones. By the average method, the monaural signal retains both the two tones from the different channel and only decreases their magnitude. As shown in Figure 21, the type-I tone leakage effect occurs on the two tones. Although each channel maintains itself tone component, the imposed external tones are also introduced into the opposite channel after the upmixing procedure.

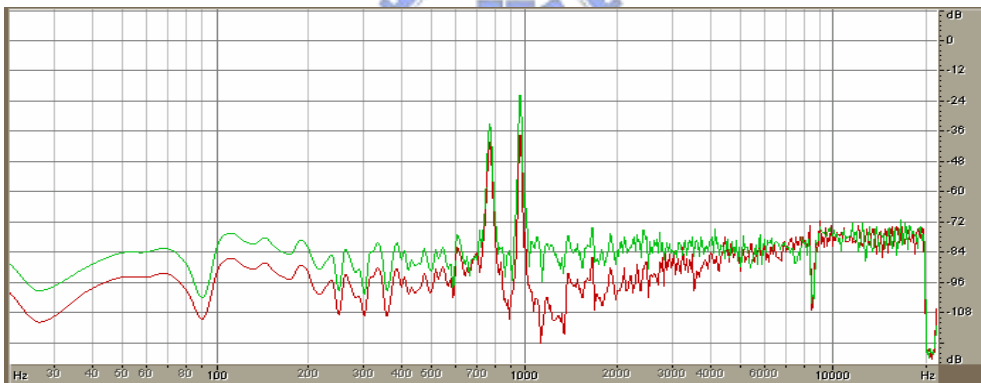


Figure 21: Reconstructed signal under averaging method for tone leakage

On the contrary, to ensure the maximum variance of data information by the KLT method, the transform process of data dimension reduction may cause the biased preference between the two channels. Especially, when there is a great difference of energy between the two channels, the coefficient vector tends to save the more dominate channel for energy compactness. In other word, the weaker channel is sacrificed inevitably and hence loses its spectral structure in the extracted downmixing signal. Therefore as shown in Figure 22, the reconstructed stereo signal only keeps the more dominate tone, and the weaker tone is suppressed to nearly disappear. This presents an example of the type-II tone leakage effect for the KLT method.

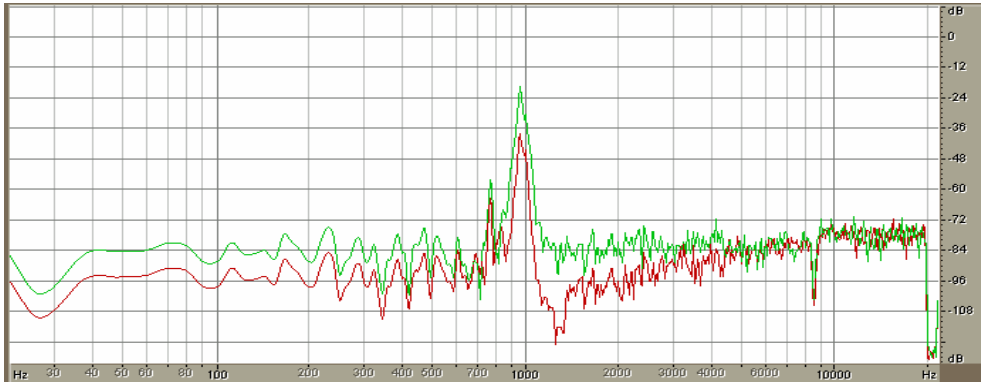


Figure 22: Reconstructed signal under KLT method for tone leakage

In conclusion, either the average method or the KLT method has the type-I and type-II tone leakage effects. Because of the inherent property that the component in the monaural signal is certainly reproduced into the reconstructed binaural signal, any downmixing method without other auxiliary information will suffer the type-I tone leakage effect. On the other hand, although the dominate tone can be held better by the KLT method than by the average method, the weaker channel is always sacrificed due to the biased ratio of combination.

Unlike for the KLT method, the type-II tone leakage effect occurs infrequently in the average method unless the two tones are exactly cancelled to each other nearly. However, energy weakness degrades greatly the quality of the average method. The contradictory between the stereo image conservation and the energy compactness is obviously the main design issue and compromise for the downmixing policy. Therefore, some amendments need to be done in KLT method to avoid the type-II tone leakage effect.

4.3.2.2 Tone Modulation Effect

Unlike the fixed combination coefficient for the average method, the coefficient vectors of the KLT must be adaptive frame by frame to achieve the optimal energy conservation. However, the adaptation results in the connection discontinuity of adjacent spectrums of the monaural, and brings an annoying effect that sounds like “click”.

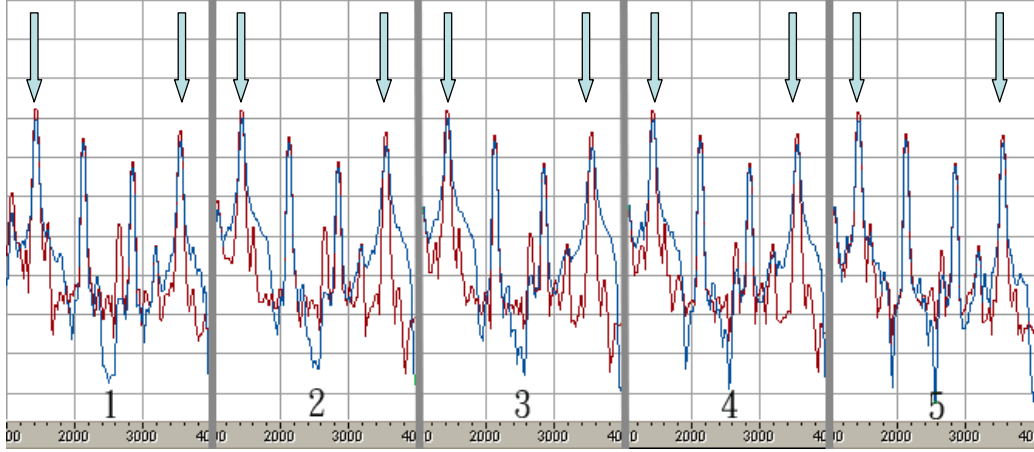


Figure 23: Example of tone modulation

Figure 23 illustrates a series of reconstructed spectrum under the KLT method. It contains spectrums of successive five frames. The red line indicates original spectrum and the blue line is reconstructed spectrum. There is an unusual phenomenon, named as “tone modulation” effect, shows the tone shape expands and contracts as time goes on. To analytically understand the cause, a downmixing subband signal should be represent by the linear combination of the left and right subband signals as

$$d[n] = \lambda_1[n] \exp(i\theta_1(n))l[n] + \lambda_2[n] \exp(i\theta_2(n))r[n], \quad (16)$$

where $\lambda_k[n] \exp(i\theta_k[n])$ $k=1,2$ means the polar form of the combination coefficient, and $l[n], r[n]$ are the left and right subband signals respectively. The influence of multiplier $\lambda_k[n] \exp(i\theta_k[n])$ will cause the modulation in both amplitude and phase. For example, consider a sinusoid signal

$$s[n] = A \exp(i(\omega n + \Theta)), \quad (17)$$

and the modulated signal

$$\hat{s}[n] = (A \cdot \lambda[n]) \exp(i(\omega n + \Theta + \theta[n])). \quad (18)$$

The multiplier $\lambda[n] \exp(i\theta[n])$ can be viewed as a step function of time index n that is constant in each frame, but may jumps hugely at the frame boundaries. Assume $s[n]$ is a tone component coupled into the monaural channel, its amplitude and frequency should be changed largely frame by frame, and hence its spectral structure will have the modulation effect. In other word, the downmixing procedure of the KLT method is equivalent to combine the two signals with mixed modulation in both amplitude and frequency, and results in the annoying “tone modulation” effect.

4.3.2.3 Pre- and Post-processing for KLT-based Approach

In the previous subsections, it introduces the risk under KLT method. However, the energy cancellation in the simple average method greatly degrades the quality. Consequently, to keep the advantage from KLT method but improve the unwanted artifacts is main issue to be solved.

Energy Normalize during pre-processing:

Energy normalization is a pre-processing for KLT method. From above discussion of tone leakage effect, the weaker signal component is usually entirely or nearly discarded in KLT method. Thus, KLT method should be revised to avoid these imbalance situations. To eschew KLT method snubs the weaker signal components, a method which lets two channels adjust their samples by the energy is used here. This method can give two channels equal priority for transform calculation. Therefore, the coefficient weight of two channels will be decided by theirs signal variation.

Coefficient Vector Smooth during post-processing:

Any adaptive mechanism of channel coupling, like the KLT method, will result in the spectrum discontinuity problems such as the tone modulation effect. An enhancement is to smooth the coupling coefficients to avoid the transient spectral discontinuity in the monaural signal. Similar to the PSOLA method commonly used to waveform synthesis in speech processing, the coefficients of adjacent frames can be smoothed by the connection of the smooth function. Thereinafter, the thesis introduces the smooth methods for KLT method as a post-processing.

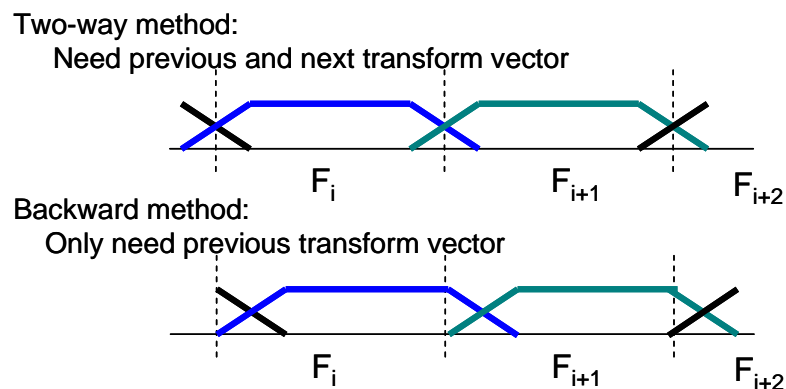


Figure 24: The smooth methods of KLT coefficient vector

Figure 24 shows the two different smooth methods: “Two-way method” and “Backward method”. Two-way method refers to the previous and next coefficient

vectors and backward method references only previous one. Here the thesis suggests using the backward method. Because two-way method must look ahead the component in next frame which let the implementation inconveniently. Also, the smooth length in both two methods is the same as each other. Thus, the backward method can handle the smooth procedure and is easily implemented.

Also the smooth curve is an issue. The simple idea is the linear smoothness. But the linear smoothness has the discontinuity at the beginning and end points. Figure 25 shows this discontinuity problem, where λ_i and λ_{i+1} indicate the previous and current coefficient vectors and the smooth area is from time index 0 to k .

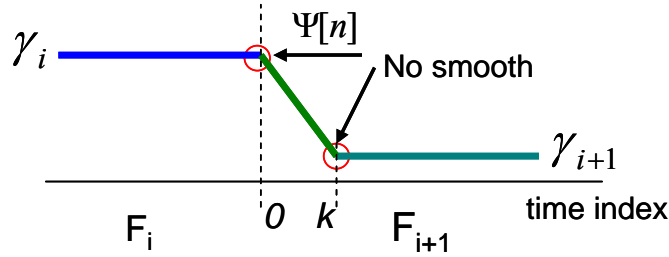


Figure 25: Linear smoothness of coefficient vectors

Because this disadvantage, the cosine curve seems to give smoother. Let the continuous smooth function $\Psi(x)$ as:

$$\Psi(x) = A \cos \pi \frac{x}{k} + B \quad (19)$$

A and B will be defined the following equations. $\Psi[n]$ must satisfy the following restriction,

$$\begin{cases} \Psi(0) = a, \Psi'(0) = 0 \\ \Psi(k) = b, \Psi'(k) = 0 \end{cases} \quad (20)$$

which exhibit the continuous connection at both beginning and end points. By substitution, the equations can be

$$\begin{cases} \Psi(0) = A \cos \pi \frac{0}{k} + B = a, \Psi'(0) = -\frac{A\pi}{k} \sin \pi \frac{0}{k} = 0 \\ \Psi(k) = A \cos \pi \frac{k}{k} + B = b, \Psi'(k) = -\frac{A\pi}{k} \sin \pi \frac{k}{k} = 0 \end{cases} \quad (21)$$

$$\begin{cases} \Psi(0) = A + B = a \\ \Psi(k) = -A + B = b \end{cases} \Rightarrow \begin{cases} A = \frac{a-b}{2} \\ B = \frac{a+b}{2} \end{cases} \quad (22)$$

Finally, $\Psi[n]$ is defined as:

$$\Psi[n] = \frac{\gamma_i - \gamma_{i+1}}{2} \cos \pi \frac{n}{k} + \frac{\gamma_i + \gamma_{i+1}}{2} . \quad (23)$$

The cosine smoothness may support smoother curve. The diagram is shown in Figure 26.

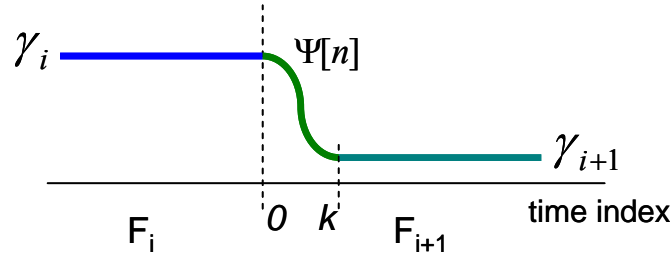


Figure 26: Cosine smoothness of coefficient vector

After all, the resultant coefficient for the frame F_{i+1} becomes a continuous function as

$$\hat{\gamma}[n] = \begin{cases} \Psi[n], & \forall 0 \leq n \leq k \\ \gamma_{i+1}, & \forall k < n \end{cases} . \quad (24)$$

According to the subjective test, the “click” noise suffered from discontinuity is enhanced largely by the smooth process and the quality is improved.

4.4 Downmix Method Summary

This chapter has discussed the downmix procedure in PS coding. Also, we have proposed the “Karhunen-Loève Transform” to naturally avoid the weakness of averaging approach. Furthermore, the chapter has considered the perceptual artifacts generated in PS coding referred to as the tone leakage effect and tone modulation effect. The thesis has suggested the pre- and post-processing of KLT for reducing these artifacts. Figure 27 illustrates the block diagram for the KLT-based downmix approach.

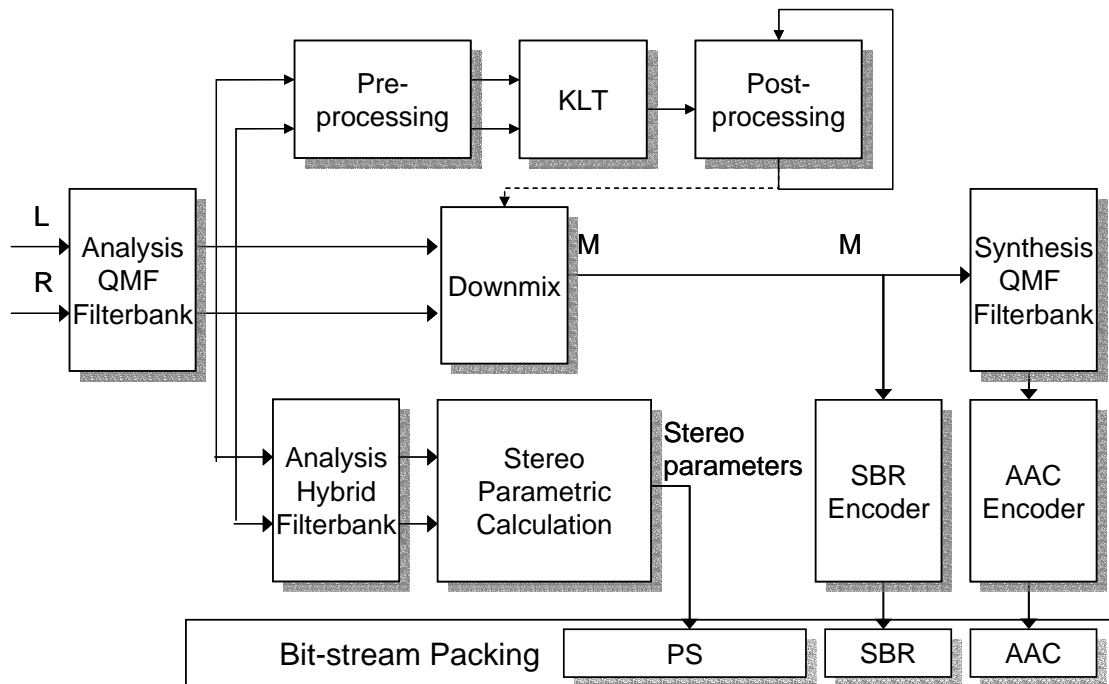


Figure 27: Block diagram of the PS encoder with KLT-based downmix method



Chapter 5

Experiments

In this chapter, a lot of tracks are conducted for verifying the proposed approaches. The tracks are based on the MPEG test tracks and the music database collected in our lab. The experiments include both objective quality measurement and subjective measurement.

5.1 Experiment Environment

Computer Status:

Platform	Personal Computer
Operating System	Windows XP
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:

For objective quality evaluation, the thesis mainly adopts the PEAQ system (perceptual evaluation of audio quality) [21] which is the recommendation system by ITU-R Task Group 10/4. The system includes a subtle perceptual model to measure the difference between two tracks. The objective difference grade (ODG) is the output variable from the objective measurement method. The ODG values should range from 0 to -4 , where 0 corresponds to an imperceptible impairment and -4 to impairment judged as very annoying. The improvement up to 0.1 is usually perceptually audible. The PEAQ has been widely used to measure the compression technique due to the capability to detect perceptual difference sensible by human hearing systems.

Subjective Quality Measurement Tool:

For subjective quality evaluation, the thesis mainly adopts the MUSHRA system [22]. The system allows the blind comparison of multiple audio files. Multi stimulus test with hidden reference and anchors 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 [23].



5.2 Objective Quality Measurement in MPEG Test Tracks

The twelve test tracks recommended by MPEG are shown in Table 3. These tracks include the critical music balancing on the percussion, string, wind instruments, and human vocal. In this section, the quality enhancement of proposed methods at different bit rates is verified based on these MPEG test tracks and NCTU-HEAAC [24] is adopted as the platform.

Table 3: 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 4: Objective measurements through the ODGs for proposed methods at 48 kbps

Codec	NCTU-HEAAC			
Bit Rate	48 kbps			
Tracks	M0	M1	M2	M3
es01	-1.54	-1.34	-1.52	-1.34
es02	-1.44	-1.43	-1.42	-1.41
es03	-1.63	-1.63	-1.63	-1.62
sc01	-3.37	-3.24	-3.29	-3.10
sc02	-3.12	-2.92	-3.06	-2.90
sc03	-2.58	-2.35	-2.59	-2.42
si01	-2.74	-2.56	-2.75	-2.58
si02	-2.51	-2.33	-2.45	-2.31
si03	-1.74	-1.66	-1.92	-1.69
sm01	-2.87	-2.66	-2.92	-2.70
sm02	-3.06	-3.06	-2.86	-2.76
sm03	-2.68	-2.46	-2.64	-2.43
Max	-1.44	-1.34	-1.42	-1.34
Min	-3.37	-3.24	-3.29	-3.10
Average	-2.4400	-2.3033	-2.4208	-2.2717

M0: Fixed stereo parameter sets with averaging downmix approach
M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
M2: Fixed stereo parameter sets with KLT-based downmix approach
M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

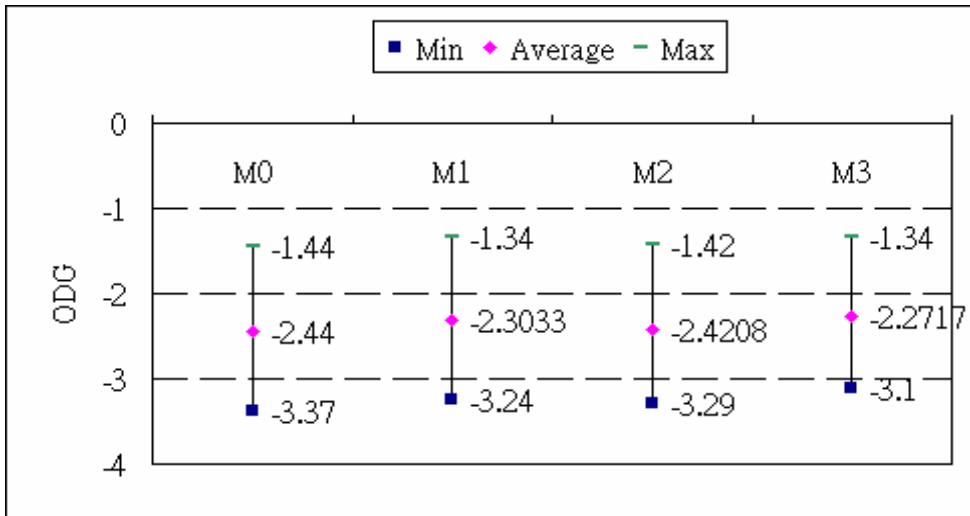


Figure 28: The variance in the ODGs of proposed methods at 48 kbps

Table 5: Objective measurements through the ODGs for proposed methods at 36 kbps

Codec	NCTU-HEAAC			
Bit Rate	36 kbps			
Tracks	M0	M1	M2	M3
es01	-2.37	-2.18	-2.32	-2.16
es02	-2.71	-2.71	-2.72	-2.71
es03	-2.81	-2.82	-2.82	-2.83
sc01	-3.41	-3.29	-3.34	-3.22
sc02	-3.27	-3.04	-3.28	-3.09
sc03	-2.86	-2.67	-2.88	-2.70
si01	-2.90	-2.73	-3.14	-3.00
si02	-2.70	-2.58	-2.65	-2.57
si03	-2.31	-2.11	-2.60	-2.36
sm01	-3.14	-2.90	-3.32	-3.04
sm02	-3.29	-3.29	-3.20	-3.14
sm03	-2.77	-2.55	-2.77	-2.56
Max	-2.31	-2.11	-2.32	-2.16
Min	-3.41	-3.29	-3.34	-3.22
Average	-2.8783	-2.7392	-2.9200	-2.7817

M0: Fixed stereo parameter sets with averaging downmix approach
M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
M2: Fixed stereo parameter sets with KLT-based downmix approach
M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

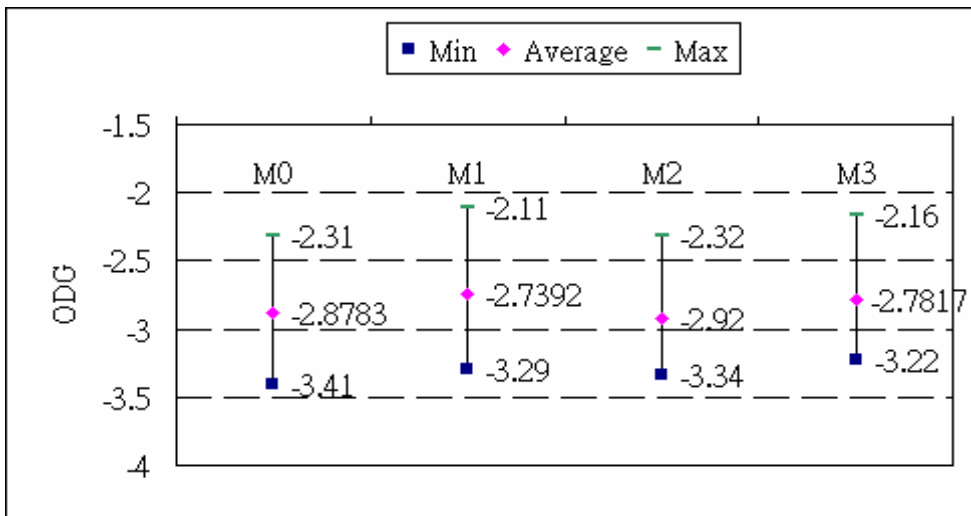


Figure 29: The variance in the ODGs of proposed methods at 36 kbps

Table 6: Objective measurements through the ODGs for proposed methods at 24 kbps

Codec	NCTU-HEAAC			
Bit Rate	24 kbps			
Tracks	M0	M1	M2	M3
es01	-3.43	-3.19	-3.45	-3.21
es02	-3.56	-3.47	-3.59	-3.47
es03	-3.74	-3.68	-3.73	-3.70
sc01	-3.48	-3.40	-3.43	-3.32
sc02	-3.30	-3.26	-3.28	-3.26
sc03	-3.17	-3.02	-3.28	-3.08
si01	-3.36	-3.00	-3.56	-3.26
si02	-3.35	-3.26	-3.30	-3.24
si03	-3.64	-3.03	-3.79	-3.38
sm01	-3.79	-3.46	-3.84	-3.62
sm02	-3.63	-3.59	-3.66	-3.65
sm03	-3.10	-2.86	-3.12	-2.89
Max	-3.10	-2.86	-3.12	-2.89
Min	-3.79	-3.68	-3.84	-3.70
Average	-3.4625	-3.2683	-3.5025	-3.3400

M0: Fixed stereo parameter sets with averaging downmix approach
M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
M2: Fixed stereo parameter sets with KLT-based downmix approach
M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

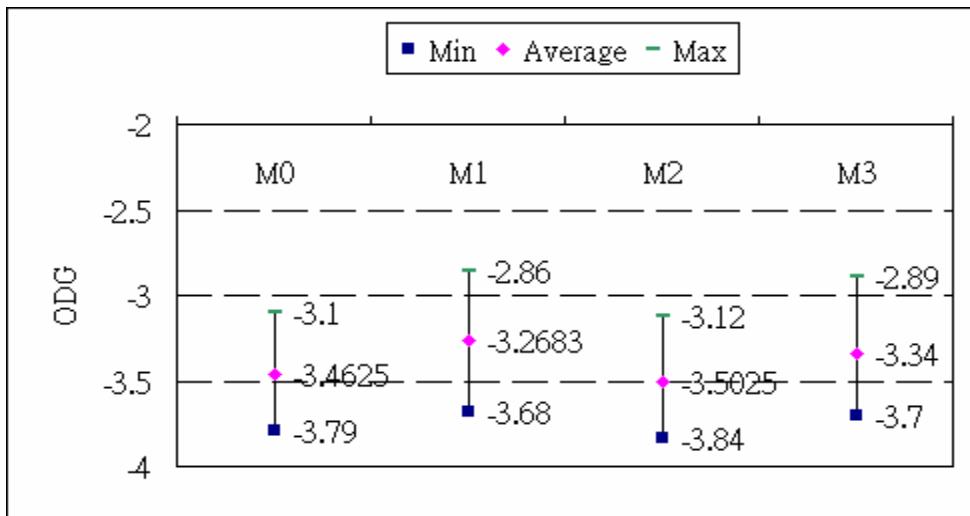


Figure 30: The variance in the ODGs of proposed methods at 24 kbps

The proposed methods are verified under the target bit-rate for PS coding. Under each bit-rate, the thesis compares the quality of four conditions: fixed stereo parameter sets with averaging downmix approach as M0, adaptive T/F stereo parameter extraction with averaging downmix approach as M1, fixed stereo parameter sets with KLT-based downmix approach as M2, and adaptive T/F stereo parameter extraction with KLT-based downmix approach as M3.

As the result shows, adaptive T/F stereo parameter extraction is verified to increase the audio quality. However, KLT-based downmix approach only improves the quality at 48 kbps. That is, under the lower bit-rate, KLT-based downmix approach isn't verified. From discussion of PS coding, the monaural downmix signal is the reconstruction basis. However, the innate character of signal in M0 and M2 is much different to each other. Thus, objective measurement here can not give sufficiently correct assessment. The thesis will adopt subjective measurement at later section to confirm the quality of this part.



5.3 Objective Quality Measurement in Music Database

From previous section, the proposed methods are verified on MPEG test set. However, there are only twelve tracks in MPEG set, but the proposed methods need great quantity of test tracks to prove its possible risk and robustness. Table 7 shows the audio database which is collected by our laboratory [25]. There are 15 categories and 320 test tracks in this audio database. Each category has its signal properties. With these large numbers of experiments, the quality of proposed methods can be assessed.

Table 7: The PSPLab audio database [25]

Bitstream Categories		# of tracks	Remark
1	ff123	101	Killer bitstream collection from ff123 [26].
2	Gpsycho	24	LAME quality test bitstream [27].
3	HA64KTest	37	64 kbps test bitstream for multi-format in HA forum [28].
4	HA128KTestV2	12	128 kbps test bitstream for multi-format in HA forum [28].
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 [29].
7	MPEG	12	MPEG test bitstream set for 48000Hz.
8	MPEG44100	12	MPEG test bitstream set for 44100 Hz.
9	Phong	8	Test bitstream collection from Phong [30].
10	PSPLab	37	Collections of bitstream from early age of PSPLab. Some are good as killer.
11	Sjeng	3	Small bitstream collection by sjeng.
12	SQAM	16	Sound quality assessment material recordings for subjective tests [31].
13	TestingSong14	14	Test bitstream collection from rshong, PSPLab.
14	TonalSignals	15	Artificial bitstream that contains sin wave etc.
15	VORBIS_TESTS_Samples	8	Eight Vorbis testing samples from HA [28].

- M0: Fixed stereo parameter sets with averaging downmix approach
- M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
- M2: Fixed stereo parameter sets with KLT-based downmix approach
- M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

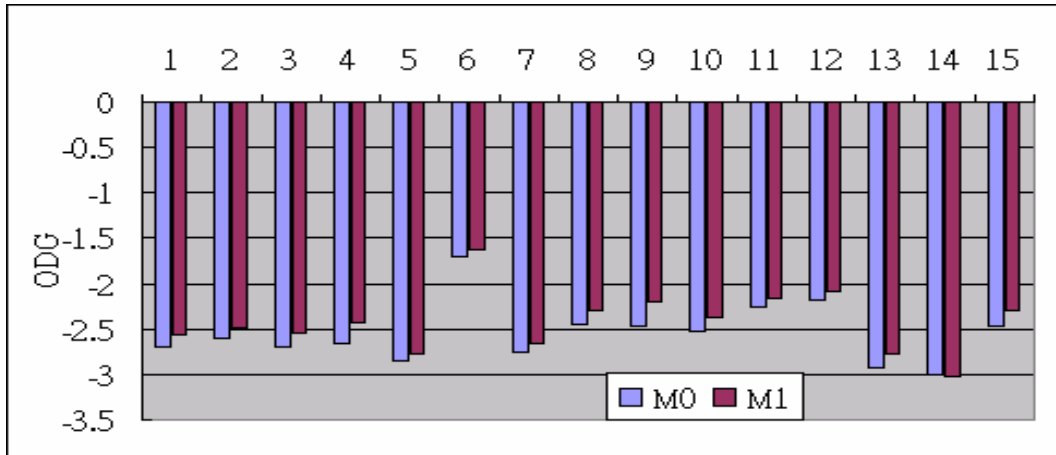


Figure 31: The average ODGs of method M0 and M1 at 48 kbps in 15 categories

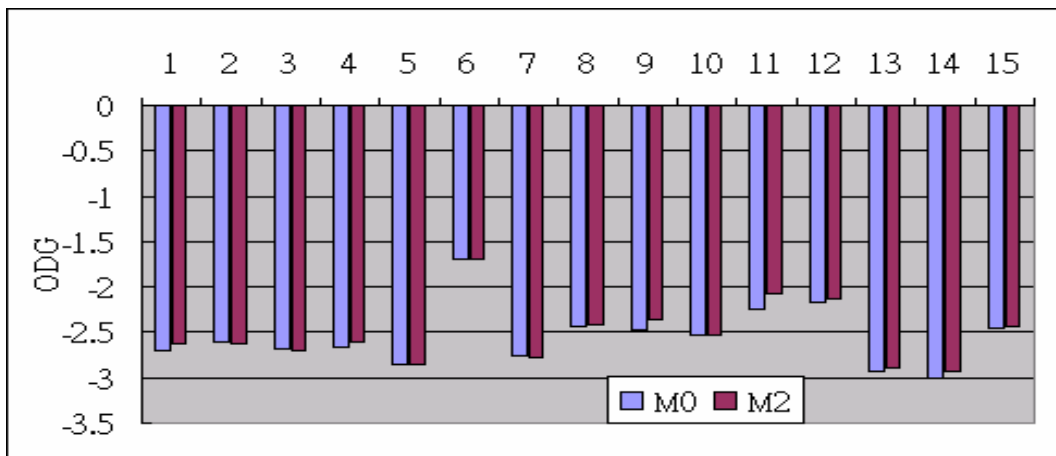


Figure 32: The average ODGs of method M0 and M2 at 48 kbps in 15 categories

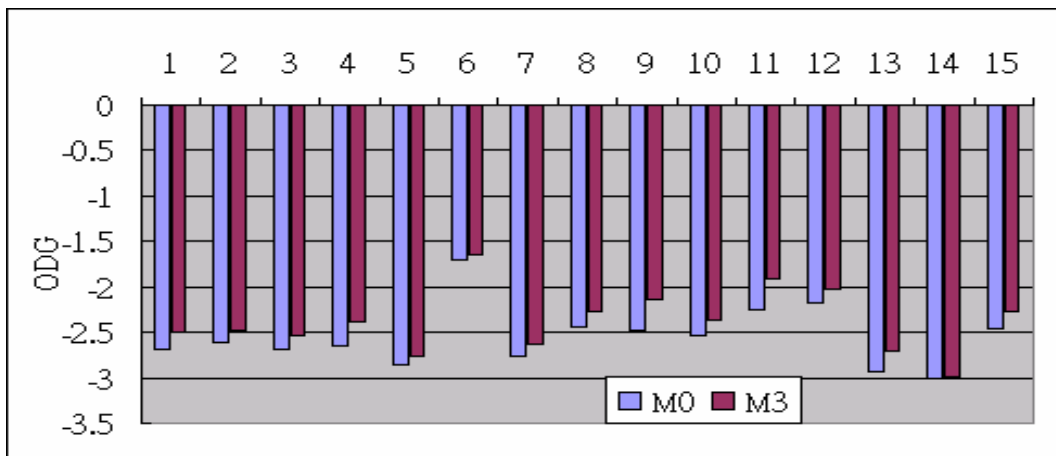


Figure 33: The average ODGs of method M0 and M3 at 48 kbps in 15 categories

- M0: Fixed stereo parameter sets with averaging downmix approach
- M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
- M2: Fixed stereo parameter sets with KLT-based downmix approach
- M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

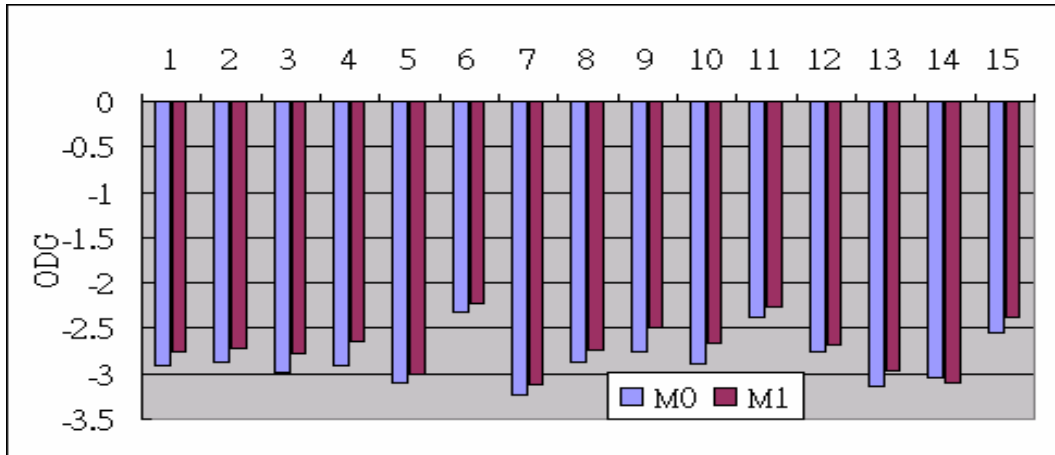


Figure 34: The average ODGs of method M0 and M1 at 36 kbps in 15 categories

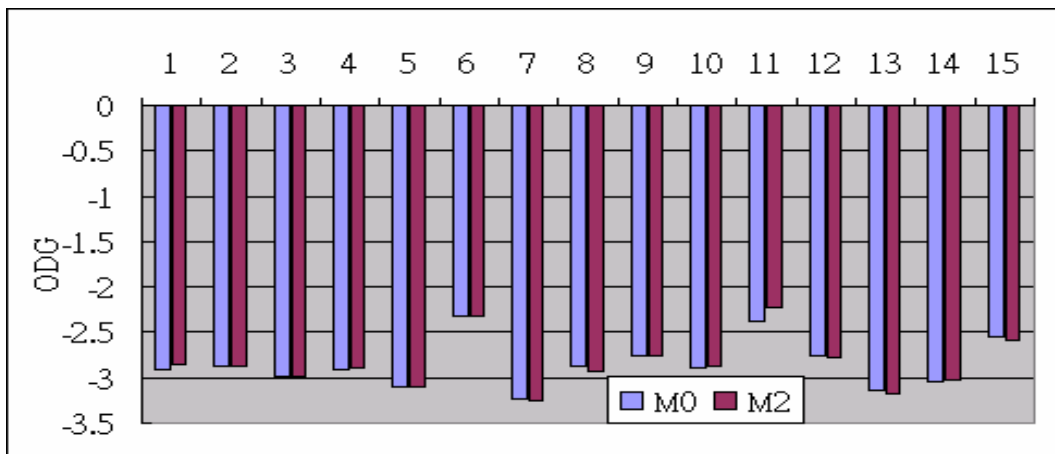


Figure 35: The average ODGs of method M0 and M2 at 36 kbps in 15 categories

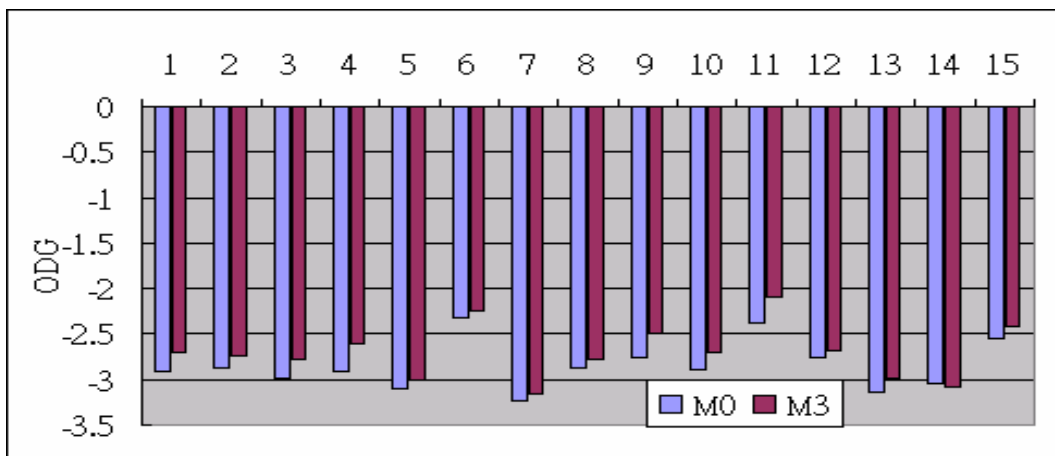


Figure 36: The average ODGs of method M0 and M3 at 36 kbps in 15 categories

- M0: Fixed stereo parameter sets with averaging downmix approach
- M1: Adaptive T/F stereo parameter extraction with averaging downmix approach
- M2: Fixed stereo parameter sets with KLT-based downmix approach
- M3: Adaptive T/F stereo parameter extraction with KLT-based downmix approach

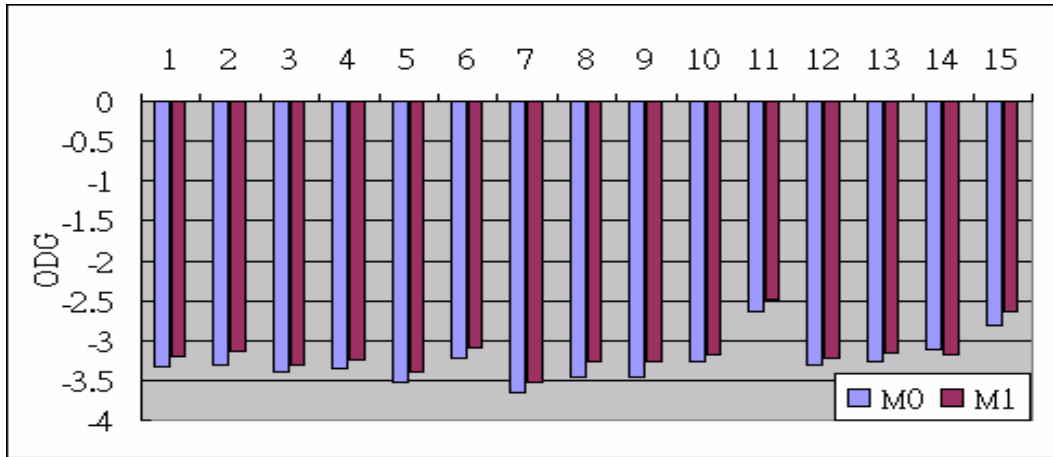


Figure 37: The average ODGs of method M0 and M1 at 24 kbps in 15 categories

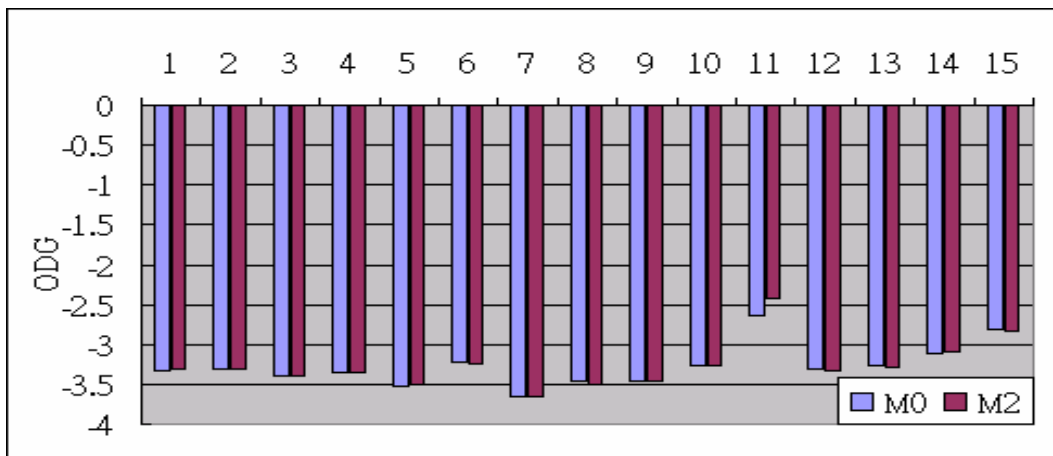


Figure 38: The average ODGs of method M0 and M2 at 24 kbps in 15 categories

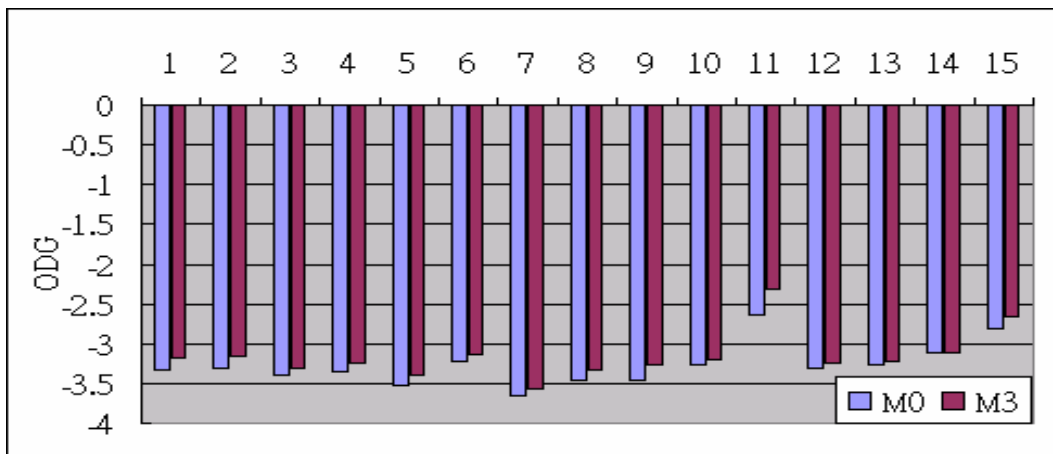


Figure 39: The average ODGs of method M0 and M3 at 24 kbps in 15 categories

Like the discussion in previous section, the adaptive T/F stereo parameter extraction is also proven in the huge music database except a test track named “square1.wav” in “TonalSignals” category. The ODG of this track is from -1.99 under M0 to -2.94 under M1. However, to check its content and compare with both encoded signals in M0 and M1, the time envelope under adaptive T/F stereo parameter extraction is more similar to the original signal. The ODG resultant of this track might be the erroneous judgment in PEAQ system.

Besides, KLT-based downmix approach gives instable improvement in the resultant. The same reason is introduced in previous section. The thesis will demonstrate the efficiency of KLT-based downmix approach by subjective quality measurement which is shown in later section.



5.4 Subjective Quality Measurement

In objective quality measurement, it proves the adaptive T/F stereo parameter extraction. Furthermore, a subjective listening test is used to ensure the quality improvement and possible risk for this proposed method. On the other hand, KLT-based downmix approach which has the different innate character against averaging approach needs the subjective measurement to verify its quality improvement. The subjective listening test is performed on the codec “NCTU-HEAAC” [24] and use the tool called “MUSHRA” [22] to assist the assessment. The twelve tracks in “MPEG44100” category are selected as the testing tracks which contain many kinds of audio environment.

Figure 40 shows the comparison of coding quality for MPEG44100 category at 36 kbps. There are three coding methods: M0, M2, and M3 which are defined in previous sections. The resultant assesses the quality improvement of KLT-based downmix approach. Besides, the improvement of KLT-based downmix approach is more sensitive in subjective test than adaptive T/F stereo parameter extraction.

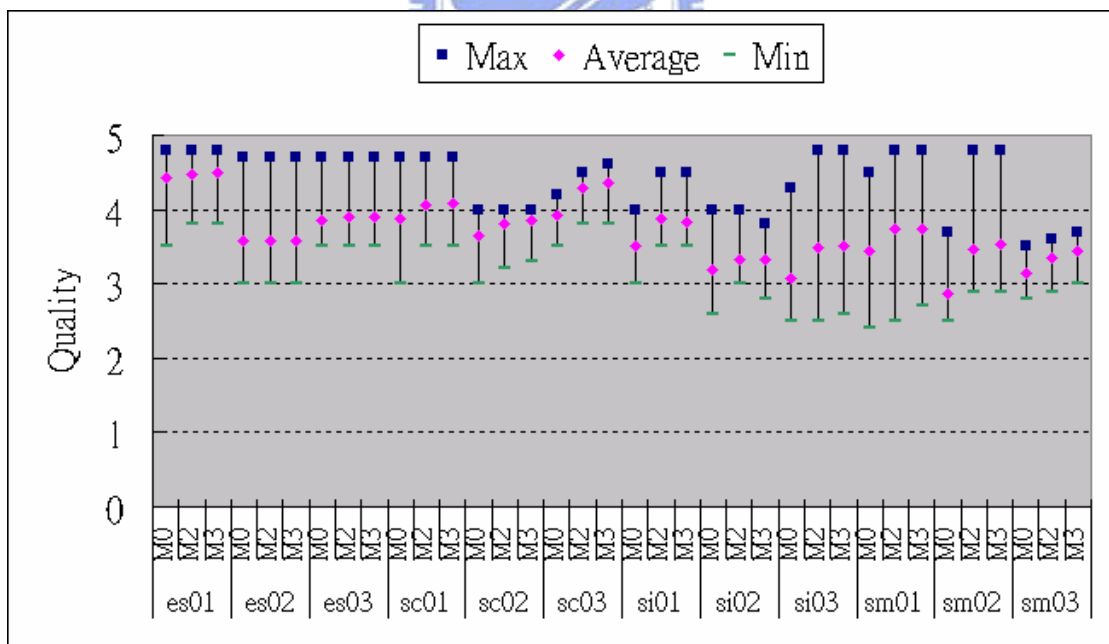


Figure 40: The subjective test result for PS coding at 36 kbps

Chapter 6

Conclusion Remark and Future Works

The thesis has brought out a new design of parametric stereo coding which includes T/F stereo parameter extraction and downmix approach. The adaptive T/F stereo parameter extraction controls the updating frequency of stereo parameters and downmix approach generates the basis monaural signal in PS coding. Because the existing methods adopt fixed approach in all signals, to avoid this content neglect, the new design controls the PS tool to be suitable for the encoded signals. In T/F stereo parameter extraction module, the adaptive method decides time borders by dynamic programming search of reconstruction error and finds out frequency resolution from the similarity between QMF subbands. Thus, these methods can naturally response to the variation of signals. Furthermore, in downmix approach, we adopt Karhunen-Loève Transform for generating the monaural downmix signal. KLT can transform the data set from high dimension into low dimension preserving the maximum possible information. Therefore, the thesis introduces the KLT-based downmix approach. However, the transform may produce some unexpected perceptual artifacts. The thesis has suggested the pre- and post-processing for KLT-based downmix approach to reduce these problems. Finally, the new methods are implemented on NCTU-HEAAC [24] and the extensive experiments are conducted for both subjective and objective measurements. The results can verify the quality of our new design for the target bit-rates for PS coding.

These new methods preliminary give the quality improvement for PS coding. Therefore, there are several aspects which can be enhanced in the future. First, in adaptive T/F stereo parameter extraction, the algorithm does not consider the influence between different types of stereo parameters. Hence, there might be a new approach to explore the relationship between different types of parameters. Second, for extension of the PS coding, although the PS is used for stereo signal compression, the basic idea also can help for multi-channel signal compression. The new codec, which called Spatial Audio Coding (SAC) [32][33][34], is such an audio enhancement tool. Under the same idea, the PS coding method also can be extended for SAC usage.

References

- [1] ISO/IEC JTC1/SC2/WGII MPEG, International Standard ISO 11172-3 “Coding of moving pictures and associated audio for digital storage media at up to about 1.5Mbit/s.”
- [2] ISO/IEC 14496-3:1999, “Information Technology–Coding of Audiovisual objects, Part3: Audio.”
- [3] ISO/IEC, “Text of ISO/IEC 14496-3:2001/FPDAM 1, Bandwidth extensions,” ISO/IEC JTC1/SC29/WG11/N5203, October 2002, Shanghai, China.
- [4] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz, “Spectral Band Replication, a novel approach in audio coding,” 112nd AES Convention, Munich, Germany, May 2002, Preprint 5553.
- [5] M. Wolters, K. Kjörling, D. Homm, H. Purnhagen, “A closer look into MPEG-4 High Efficiency AAC,” 115th AES Convention, New York, USA, October 2003, Preprint 5871.
- [6] H.W. Hsu, C.M. Liu, and W.C. Lee, “Audio Patch Method in MPEG-4 HE AAC Decoder,” 117th AES Convention, San Francisco, USA, October 2004, Preprint 6221.
- [7] C.M. Liu, L.W. Chen, H.W. Hsu, and W.C. Lee, “Bit Reservoir Design for HE-AAC,” at the 118th AES Convention, Barcelona, Spain, May 28~31, 2005.
- [8] H. Purnhagen: “Low Complexity Parametric Stereo Coding in MPEG-4”, 7th International Conference on Audio Effects (DAFX-04), Naples, Italy, October 2004.
- [9] E. Schuijers, J. Breebaart, H. Purnhagen, J. Engdegård: “Low complexity parametric stereo coding”, Proc. 116th AES convention, Berlin, Germany, 2004, Preprint 6073.
- [10] Draft ISO/IEC 14496-3 (Audio 3rd Edition), “Coding of Moving Pictures and Audio, Subpart 8: Technical description of parametric coding for high quality audio.”

- [11] C. Faller, F. Baumgarte: “Efficient Representation of Spatial Audio Using Perceptual Parametrization”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York 2001.
- [12] C. Faller and F. Baumgarte, “Binaural Cue Coding - Part II: Schemes and applications,” IEEE Trans. on Speech and Audio Proc., vol. 11, no. 6, Nov. 2003.
- [13] 3GPP TS 26.410 V6.0.0 (2004-09), website: <http://www.3gpp.org>.
- [14] Nero, website: <http://www.nero.com>.
- [15] Coding Technologies, aacPlusEval v2 Evaluation Package, Version 7.0.5, website: <http://portal.codingtechnologies.de/eval/aacPlusEval/>.
- [16] H.F. Silverman and D.P. Morgan, “The application of dynamic programming to connected speech recognition,” IEEE Acoustics, Speech, and Signal Processing Magazine, vol. 7, pp. 6-25, July, 1990.
- [17] R. Bellman, “On the theory of dynamic programming,” Proceedings of the National Academy of Sciences, vol. 38, pp. 716-719.
- [18] C.M. Liu, W.C. Lee, C.H. Yang, K.Y. Pang, T. Chiou, T.W. Chang, Y.H. Hsiao, H.W. Hsu, C.T. Chien, “Design of MPEG-4 AAC Encoder,” 117th AES Convention, San Francisco, USA, October 2004, Preprint 6201.
- [19] K.C. Lee, C.H. Yang, H.W. Hsu, C.M. Liu, W.C. Lee, and T.W. Chang, “Efficient Design of Time-Frequency Stereo Parameter Sets for Parametric HE-AAC,” 119th AES Convention, New York, USA, October 2005, Preprint 6600.
- [20] A. Rezayee and S. Gazor. “An Adaptive KLT Approach for Speech Enhancement,” IEEE Speech, and Audio Processing Magazine, vol. 9, pp. 87~95, Feb, 2001.
- [21] ITU Radiocommunication Study Group 6, “Draft Revision to Recommendation ITU-R BS.1387- Method for objective measurements of perceived audio quality.”
- [22] MUSHRA, website: <http://ff123.net/abrhr/abchr.html>.
- [23] ITU Radiocommunication Sector BS.1116 (rev.1). “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems,” Geneva, 1997.

- [24] NCTU-HEAAC, website:
<http://psplab.csie.nctu.edu.tw/projects/index.pl/nctu-mp3.html>.
- [25] The Audio Database Collected in Perceptual Signal Processing Lab, website:
<http://psplab.csie.nctu.edu.tw/projects/index.pl/testbitstreams.html>.
- [26] Samples for Testing Audio Codecs from ff123, website:
<http://ff123.net/samples.html>.
- [27] Quality and Listening Test Information for LAME, website:
<http://lame.sourceforge.net/gpsycho/quality.html>.
- [28] Hydrogen Audio, website: <http://www.hydrogenaudio.org>.
- [29] OGG Vorbis Pre 1.0 Listening Test, website:
<http://hem.passagen.se/ingets1/vorbis.htm>.
- [30] Phong's Audio Samples, website: <http://www.phong.org/audio/samples.xhtml>.
- [31] Sound Quality Assessment Material, website:
<http://sound.media.mit.edu/mpeg4/audio/sqam/>.
- [32] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, P. Kroon, "Spatial Audio Coding: Next-Generation Efficient and Compatible Coding of Multi-Channel Audio," 117th AES Convention, San Francisco, USA, October 2004, Preprint 6186.
- [33] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjör ling, E. Schuijers, J. Hilpert, F. Myburg, "The Reference Model Architecture for MPEG Spatial Audio Coding," 118th AES Convention, Barcelona, Spain, May 2005, Preprint 6477
- [34] J. Breebaart, J. Herre, C. Faller, J. Rödén, F. Myburg, S. Disch, H. Purnhagen, G. Hotho, M. Neusinger, K. Kjör ling, W. Oomen, "MPEG Spatial Audio Coding/MPEG Surround: Overview and Current Status", 119th AES Convention, New York, USA, October 2005, Preprint 6599.