國立交通大學

多媒體工程研究所

碩士論文

基於 Gram-Schmidt 正交化的 MPEG Surround 降混音器與去相關器之設計

Gram-Schmidt-based Downmixer and Decorrelator in the MPEG Surround Coding

研究生：陳德沛

指導教授：蕭旭峯 教授

劉啟民 教授

中華民國 九十九 年 八月

基於 Gram-Schmidt 正交化的 MPEG Surround 降混音器與去相關器之設計

Gram-Schmidt-based Downmixer and Decorrelator in the MPEG Surround Coding

研究生：陳德沛　　Student: Der-Pei Chen

指導教授：蕭旭峯　　Advisor: Dr. Hsu-Feng Hsiao

劉啟民　　Dr. Chi-Min Liu

國立交通大學
多媒體工程研究所
碩士論文

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

Computer Science

August 2010

Hsinchu, Taiwan, Republic of China

中華民國 九十九 年 八月

基於 Gram-Schmidt 正交化的 MPEG Surround 降混音器與去相關器之設計

學生：陳德沛　　　　　　　　　　　　　　指導教授：蕭旭峯　博士
劉啟民　博士
國立交通大學多媒體工程研究所碩士班

中文論文摘要

MPEG Surround 乃一項低位元率多聲道音訊壓縮標準。其壓縮的原理是透過降混音(down-mix)處理將多聲道訊號耦合成雙聲或單聲道訊號，並計算出聲源定位的空間參數(spatial parameter)，來達到減少聲道數與紀錄聲場之目的。解碼端透過去相關器產生的去相關訊號(decorrelated signal)，並根據空間參數進行升混音(up-mix)處理來重建聲源定位與空間寬廣度的環繞效果。因此，編碼端如何將多聲道耦合成雙聲或單聲道訊號，與去相關訊號的生成，將影響重建聲音的品質。

此篇論文，使用 Gram-Schmidt 正交化的概念，改進去相關器與二轉一降混音模組的運作方式。方法的改進效果將透過升混音後訊號與原始訊號的聲道間能量差與相關性差值變化，及 ODG(Objective Difference Grade)客觀量測與 MUSHRA 主觀測試來驗證。

Gram-Schmidt-based Downmixer and Decorrelator in the MPEG Surround Coding

Student：Der-Pei Chen　　　　Advisor：Dr. Hsu-Feng Hsiao

Dr. Chi-Min Liu

Institute of Multimedia and Engineering

College of Computer Science

National Chiao Tung University

Abstract

MPEG Surround (MPS) coding is an efficient method for multichannel audio coding. In an MPS encoder, downmixing from multichannel signals into a less number of channels is an efficient way to achieve high compression rate. In decoder, an upmixing module combining with the decorrelator is the key module to reconstruct the multichannel signals. This thesis considers the design of the downmixer and the decorrelator with the assistance of the Gram-Schmidt orthogonal process. The performance of the proposed downmixer and decorrelator is verified through the differences of Channel Level Difference (*CLD*) and Inter-Channel Coherence (*ICC*) between the upmixed signals and the original signals. Also, the objective and subjective quality measurements are conducted.

Acknowledgement

I would like to express my sincere gratitude to my advisor, Prof. Hsu-Feng Hsiao. His wide knowledge has broadened my horizon on the field of Computer Science and his logical thinking shows me the way to do better research. Also, his understanding, encouraging, and personal guidance have provided a solid foundation for the present thesis.

I appreciate Prof. Chi-Min Liu for his valuable advice. His suggestions and discussions during the weekly meeting are always been helpful for this study.

I am grateful to Mr. Han-Wen Hsu because he gave me a lot of inspirations when I bottlenecked on my research. If I did not have his advice or help, I might have much more hard time than I have already experienced.

I am thankful to the labmates or schoolmates I met during my graduate school life at National Chiao Tung University, especially Mr. Yun-Hsiu Tung. They make my life full of laughter and give me a hand when I am in need of help.

Finally, I owe my loving thanks to my parents, Mr. Yow-Shin Chen and Ms. Hsia-Hua Yu. I couldn't have become what I am right now without their love and care. Therefore, I would like to dedicate my thesis to them.

Part of this work has been published at 128th Convention of Audio Engineering Society. This work is supported by the National Science Council under the contract numbers NSC98-2221-E-009-093 and NSC98-2221-E-009-120, which are gratefully acknowledged.

Jeffrey Chen

August 2010

Contents

Figure List

Table List

Chapter 1 Introduction

In real world, sound could be from anywhere around us. The psycho-based listening perception gives us the ability to locate the actual position of the sound in 3D world. Therefore, lots of applications, such as multi-sound tracks movies or multi-channel DVDs, take advantage of this human perceptual feeling and give us the illusion that we are in the environment made by the creator. Since every channel has different, or slightly different, content to playback, if we treat them separately, that would require a large amount of memory spaces and make it less possible for applications. Fortunately, ISO/IEC provides a solution to deal with multichannel audio coding, called MPEG Surround (MPS).

MPS [1]-[3] is standardized to exploit the correlation among audio channels to achieve high coding efficiency for multi-channel audio. The concept of MPS is to combine the multichannel signals into a stereo/mono downmixed signal with the spatial parameters including Channel Level Difference (*CLD*) and the Inter-Channel Coherence (*ICC*). The decoder then uses the stereo/mono downmixed signal and the spatial parameters to reconstruct the multichannel signals. There are two key modules affecting the reconstructed audio quality. The first one is the Two-To-One (TTO) downmixer, which combines the stereo audio into mono signal and is illustrated in Figure 1. The other is the decorrelator, which generates the "pseudo-audio" channels for upmixer and is illustrated as the yellow box with "D" in the MPS Decoder shown in Figure 2. This thesis considers the design issues of the TTO downmixer and decorrelator and proposes the novel algorithms for the two critical modules.

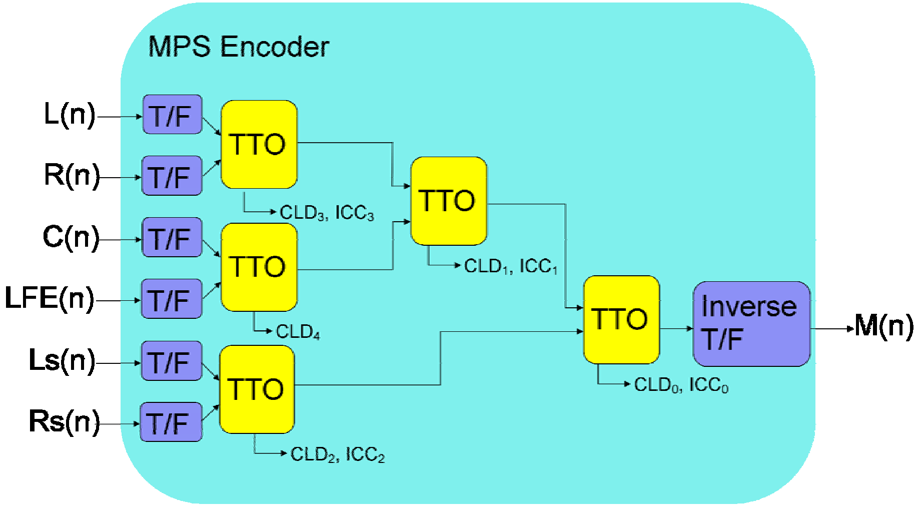


Figure 1 MPEG Surround Encoder (5151)

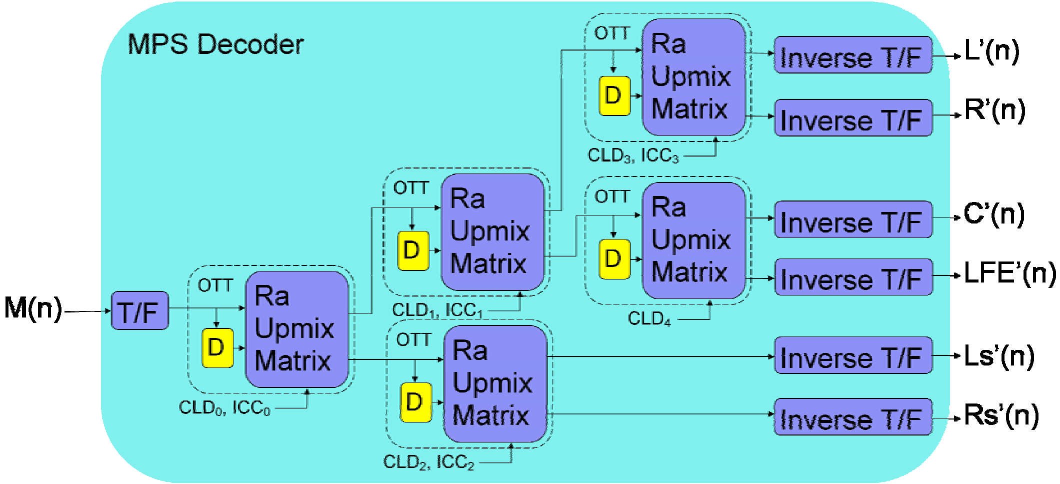


Figure 2 MPEG Surround Decoder (5151)

In MPS decoder, the upmix process is conducted based on the downmixed signals and the decorrelated signals, which are generated by feeding the downmixed signals into a cascade of all-pass filters. The uncorrelation between the decorrelated signal and the downmixed signal are necessary to reconstruct the audio signals preserving the *CLD* and *ICC* of the original signals. However, we will show that the decorrelation cannot be achieved by the current decorrelator in MPS. This thesis proposes a novel decorrelator based on the GS (Gram-Schmidt) orthogonalization process which decomposes a non-ideal decorrelated signal into a correlated component and an uncorrelated component with respect to the downmixed signal. Then the correlated component is turned to be another uncorrelated part by changing the phase of the signal. An adaptive scaling is designed to control the two components to maintain the signal smoothness and the uncorrelation requirement. The proposed decorrelator is compliant to the MPS standard. Experiments show this modification on the decorrelator improves objective difference grade by 0.18 on average.

In an MPS encoder, the TTO downmixer needs to ensure that the energy of the downmixed signal is equal to the sum of the energies of the two input signals so that the upmixed signals in MPS decoder can preserve the energies of the original signals. However, for most of the downmix methods in literature, such as the direct summation of the input signals, the downmixed energy cannot be kept controlled. One straightforward approach is to scale the downmixed signals to fit the energy, but unexpected artifacts might arise due to the possible amplification of the noise. This thesis proposes the GS-based TTO downmixer which can achieve the equality of energies with consideration to the risk and compliant to MPS standard.

This thesis is organized as follows. Chapter 2 provides an overview of the fundamental components related to the main goal of this thesis in MPEG Surround. Chapter 3 demonstrates that three upmix objectives can be achieved by keeping the desired properties on the TTO downmixer and decorrelator. The problem definition and the proposed solutions are presented in Chapter 4, and the experiments are conducted in Chapter 5. Chapter 6 concludes the thesis.

Chapter 2 Backgrounds

MPEG Surround is under the concept of spatial audio coding, so we would introduce the spatial audio coding first and then go through the components, which is related to the goal of this thesis, in MPEG Surround.

2.1 Spatial Audio Coding

The concept of spatial audio coding as employed in MPEG Surround standard is shown in Figure 3. A multi-channel input is converted to a mono/stereo downmixed signal by a MPEG Surround encoder. The properties of the original input signals which might be lost by downmixing are captured and transmitted through a spatial parameter bit stream. The downmixed signal is process by a legacy downmix encoder. The multiplexer combines the downmixed signal from legacy downmix encoder with the spatial parameter bit stream to one output, which would be the input of the decoder side.

After the encoded signal is fed into the decoder shown in Figure 3, demultiplexer separates the downmixed signal and spatial parameter bit stream. The downmixed signal is decoded by the legacy downmix decoder and then is upmixed according to transmitted spatial parameters by MPEG Surround decoder.

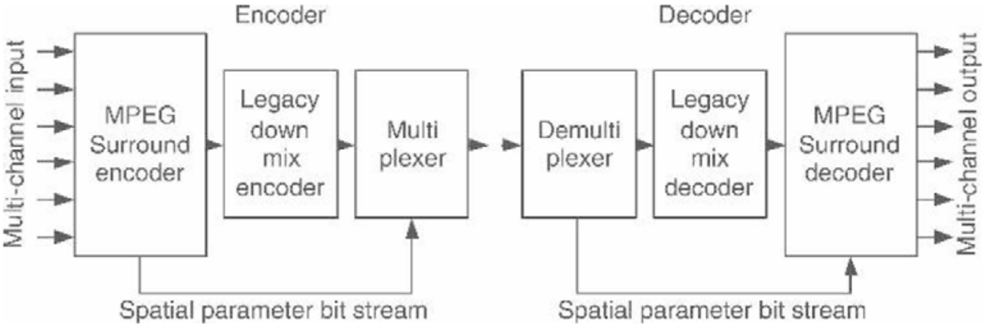


Figure 3　Multichannel encoder and decoder according to spatial audio coding concept [2]

2.2 MPEG Surround

Though Figure 1 and Figure 2 only show one of the coding tree structures provided by MPEG Surround, the components inside the tree structures are the same. Therefore, the introductions for each component are needed for the background knowledge.

Time/Frequency Transformation

The applied filterbank is a hybrid complex-modulated quadrature mirror filterbank (QMF). As shown in Figure 1 and Figure 2, the input signals of the encoder and decoder are passed into time/frequency analysis filterbank, which is composed by a QMF analysis filterbank, subfilters and a delay shown in the left panel of Figure 4. The signals are first fed into the QMF analysis filterbank. Since each subband of QMF analysis filterbank has the same bandwidth, the frequency resolution of QMF analysis filterbank cannot response the sense of hearing model in which the lower frequency part requires higher resolution. Therefore, subfilters are used to provide non-uniform resolution in low frequency and the high frequency part is delayed by the delay module. In Figure 4, the input signal X is fed into the QMF analysis filterbank and turned to be the 64 QMF subband signal, $X\mathrm{q}_{\mathrm{i}}$ with $0 \leq i \leq 64$. Then the 64 QMF subbands would be splitted to 71 hybrid subbands, $X\mathrm{m}_{\mathrm{j}}$ with $0 \leq j \leq 71$. While calculating the spatial parameters at the mixing boxes, the 71 hybrid subbands are rearranged to 28 parameter bands, which is the basic unit of MPS, at most. The inverse time/frequency transformation is combined by the sum modules and a QMF synthesis filterbank, shown in the right panel of Figure 4. The modules of sum and QMF synthesis filterbank are corresponding to subfilters and QMF analysis filterbank respectively. Since the filterbank has the same structure as the one applied in Parametric Stereo, detailed introduction and information can be found in [2]- [4].

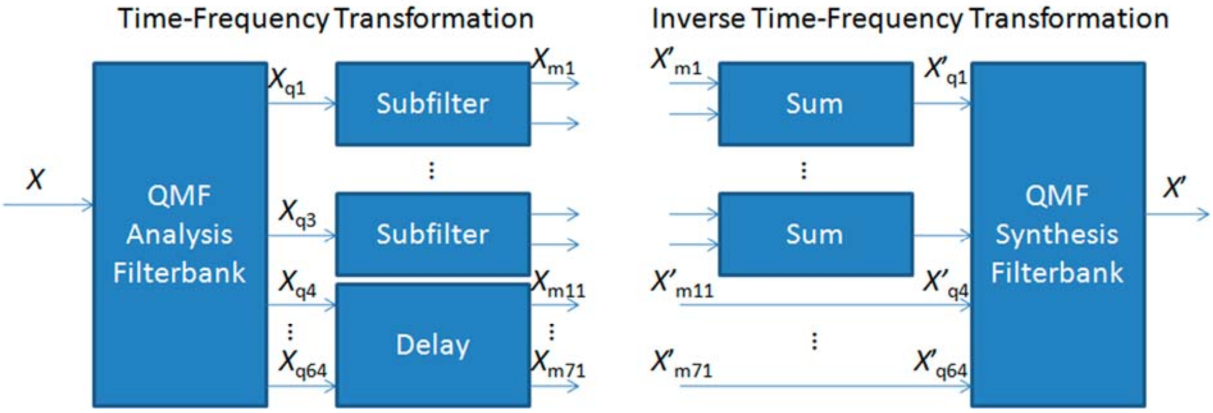


Figure 4 Structure of the time-frequency transformation with its inversion

Framing

There are some restrictions for time resolution. There are 32 time slots in one frame with at most 8 parameter sets, which are the time slots that keep the calculated spatial parameters. The calculated parameter sets can be places at any time slot by variable framing or the time slots with equal distance by fixed framing. No matter which framing is used, at least one set of parameters are placed in the last time slot of a frame. The time slots that do not have the spatial parameters are then interpolated on the R_a matrix coefficients, which would be introduced in the next subsection.

Figure 5 illustrates a frame example with two parameter sets of a parameter band. According to the previous paragraph in this subsection, each column indicates a time slot; the columns with gray grounding are the time slot with spatial parameters; the red line indicates the matrix coefficient value; the bold solid line indicates the frame borders; and the blue line illustrates the interpolation of matrix coefficients.

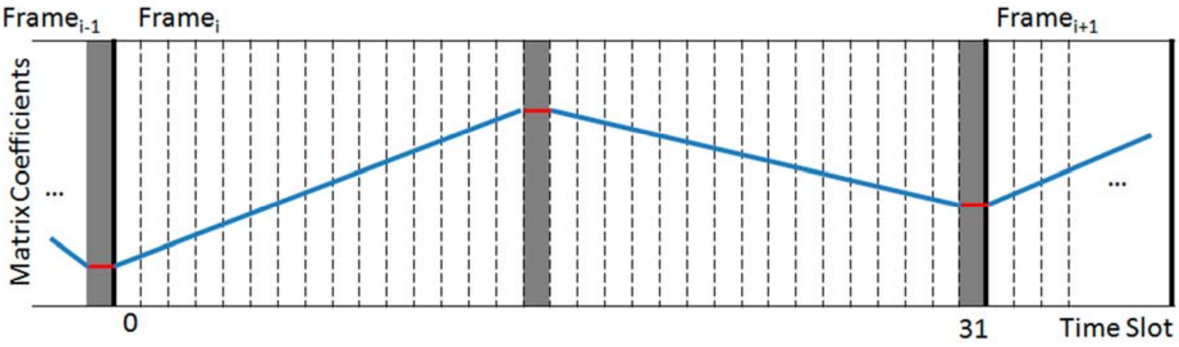


Figure 5 A MPS frame with two parameter sets of a parameter band

Elementary Building Blocks

The common coding blocks for MPEG Surround are One-To-Two (OTT) and Two-To-Three (TTT) at the decoder side, and the corresponding blocks used on the encoder side are Two-To-One (TTO) and Reverse TTT (R-TTT).

- Encoder

TTO coverts a stereo input signal to a mono signal, combined with parameter extraction which represents the spatial parameters between the respective input signals. Since [1] does not detailed specified how to realize downmixing for encoder, we reference the MPS reference software provided by MPEG [5] and know that the downmixing is realized by the direct summation of the input signals as

$$Y_m[n] = X_{1,m}[n] + X_{2,m}[n], \tag{1}$$

where $X_{1,m}$ and $X_{2,m}$ denotes the input signals on the hybrid subband m; Y_m denotes the output signal; n represents the time slot index.

In the appendix of [1], the power ratio of corresponding time/frequency tiles of the input signals, which would be denoted as “Channel Level Difference” or *CLD*, is defined as

$$CLD = 10\log_{10}\left(\frac{\sum_{n=0}^{31}\sum_{m=m_b}^{m_{b+1}-1} X_{1,m}[n]X_{1,m}^{*}[n]}{\sum_{n=0}^{31}\sum_{m=m_b}^{m_{b+1}-1} X_{2,m}[n]X_{2,m}^{*}[n]}\right), \tag{2}$$

where m_b is the hybrid subband boundaries of the *b*th parameter band.

Also, a similarity measure of the corresponding time/frequency tiles of the input signals, which would be denoted as “Inter-Channel Correlation” or *ICC*, is given by the cross correlation as

$$ICC = \mathrm{Re}\left\{\frac{\sum_{n=0}^{31}\sum_{m=m_b}^{m_{b+1}-1} X_{1,m}[n]X_{2,m}^{*}[n]}{\sqrt{\sum_{n=0}^{31}\sum_{m=m_b}^{m_{b+1}-1} X_{1,m}[n]X_{1,m}^{*}[n]\sum_{n=0}^{31}\sum_{m=m_b}^{m_{b+1}-1} X_{2,m}[n]X_{2,m}^{*}[n]}}\right\}. \tag{3}$$

- Decoder

There are two components in an OTT. The first one is the decorrelator and the other is the R_a upmix matrix.

In literature, there are many decorrelation methods [6]-[12]. In [7], decorrelation works on simple nonlinear functions, which is simple in computations. [8] uses interleaving comb filters and frequency shifts to realize decorrelation. [9] and [10] uses time-varying all-pass filters. Boueri et al. [11] proposes a new random time-shifting method on critical band to lower the cross-correlation. A Karhunen–Loève transform based method is also mentioned in [12] for decorrelation. Despite the fact there are many approaches for decorrelation, only the approach used for MPS decorrelator is discussed in this thesis.

Decorrelators are used to generate an uncorrelated signal from the input to simulate the missing channel(s) information for the upmix matrix operations. They are realized by comprising a delay, a lattice all-pass filter, and an energy adjustment stage shown in Figure 6. The configurations for the delay and all-pass filter are controlled by the decorrelator configuration transmitted from encoder. For MPEG Surround, the OTT and prediction-mode TTT have one decorrelator in both them.

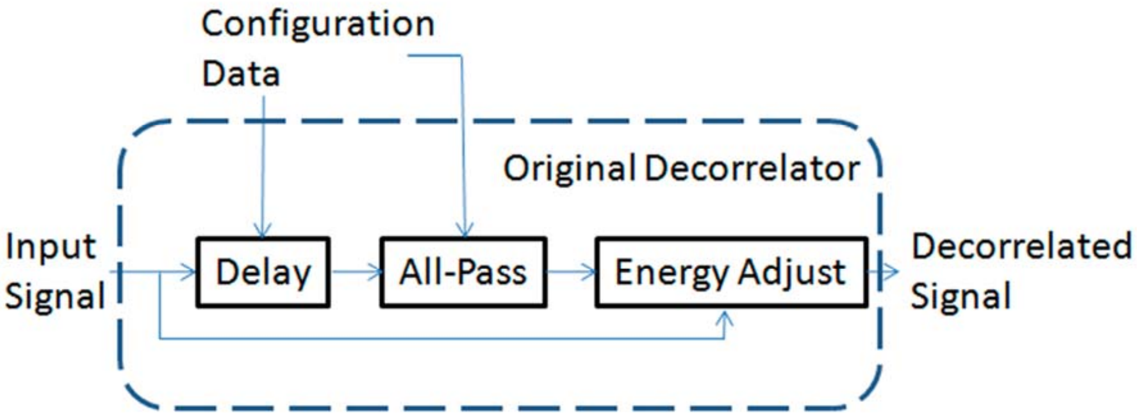


Figure 6 Diagram of the original decorrelator on hybrid QMF domain signals

Let us detailed introduce how the original decorrelator operates. The delayed hybrid subband domain samples $D_m^{delay}[n]$ are obtained as

$$D_m^{delay}[n] = \begin{cases} M_m[n-8] & ,m \in 0-7 \\ M_m[n-7] & ,m \in 8-20 \\ M_m[n-2] & ,m \in 21-29 \\ M_m[n-1] & ,m \in 30-71 \end{cases}, \tag{4}$$

where $M_m[n]$ for $n < 0$ contains the buffered values of the last frame at position $32 - n$. Then the delayed hybrid subband domain samples $D_m^{filt}[n]$ are filters as

$$D_m^{filt}[n] = \frac{1}{a_m[0]} \cdot \left(\sum_{s=0}^{S} b_m[s] \cdot D_m^{delay}[n-1] - \sum_{s=1}^{S} a_m[s] \cdot D_m^{filt}[n-1] \right), \tag{5}$$

where S is the length of the lattice coefficient vector $s[n]$ and the filter coefficients $a_m[n]$ and $b_m[n]$ are derived from the lattice coefficient vector $s[n]$, which is well-defined in the Section 6.6 in [1].

After the $D_m^{filt}[n]$ have been obtained for $0 \le n < 32$, the following energy adjustment procedure is applied. The powers per parameter band *k* of the input samples $E_k^M[n]$ and the filtered samples $E_k^D[n]$ are calculated as

$$E_k^M[n] = \sum_{\forall m \in \kappa(m)=k} \left|M_m[n]\right|^2 , \tag{6}$$

$$E_k^D[n] = \sum_{\forall m \in \kappa(m)=k} \left|D_m^{filt}[n]\right|^2 , \tag{7}$$

with *κ*(*m*) defines in Table A.31 in [1], which is a lookup table for hybridband to parameter band. Then, low-pass filtering on the powers is applied as

$$E_k^{M,Smooth}[n] = \alpha \cdot E_k^{M,Smooth}[n-1] + (1-\alpha) \cdot E_k^M[n], \tag{8}$$

$$E_k^{D,Smooth}[n] = \alpha \cdot E_k^{D,Smooth}[n-1] + (1-\alpha) \cdot E_k^D[n] , \tag{9}$$

with α = 0.8. For the first slot of the first frame both $E_k^{M,Smooth}[n-1]$ and $E_k^{D,Smooth}[n-1]$ with *n* = 0 are initialized as zero vector. For the first slots of all other frames, both $E_k^{M,Smooth}[n-1]$ and $E_k^{D,Smooth}[n-1]$ with *n* = 0 are set to the value of $E_k^{M,Smooth}[n]$ and $E_k^{D,Smooth}[n]$ of the previous frame at *n* = 31. The energy-shaping gain vector is calculated as

$$g_k[n] = \begin{cases} \sqrt{\dfrac{\gamma E_k^{M,Smooth}[n]}{E_k^{D,Smooth}[n] + \varepsilon}} & , E_k^{D,Smooth}[n] > E_k^{M,Smooth}[n]\gamma \\ \min\left(\sqrt{\dfrac{E_k^{M,Smooth}[n]}{\gamma E_k^{D,Smooth}[n] + \varepsilon}}, 2 \right) & , \gamma E_k^{D,Smooth}[n] > E_k^{M,Smooth}[n], \\ 1 & , otherwise \end{cases} \tag{10}$$

with γ = 1.5 and ε = 1e -9. Finally, the decorrelator outputs are constructed as

$$D_m[n] = g_{\kappa(m)}[n] \cdot D_m^{filt}[n]. \tag{11}$$

The OTT matrix operation in Figure 2 is represented as

$$\begin{bmatrix} X'_{1,m}[n] \\ X'_{2,m}[n] \end{bmatrix} = R_a^{n,m} \begin{bmatrix} M_m[n] \\ D_m[n] \end{bmatrix}, \tag{12}$$

where $M_m[n]$ is the downmixed hybrid subband signal; $D_m[n]$ is the output signal of the decorrelator for $M_m[n]$; $X'_{1,m}[n]$ and $X'_{2,m}[n]$ are the upmixed signals. According to the parameters *ICC* and *CLD*, MPS standard specifies R_a as

$$R_a^{l,m} = \begin{bmatrix} H11 & H12 \\ H21 & H22 \end{bmatrix} = \begin{bmatrix} \lambda_1 \cos(\alpha+\beta) & \lambda_1 \sin(\alpha+\beta) \\ \lambda_2 \cos(-\alpha+\beta) & \lambda_2 \sin(-\alpha+\beta) \end{bmatrix}, \tag{13}$$

where

$$\alpha = \frac{1}{2}\arccos(ICC), \tag{14}$$

$$\beta = \tan\left(\frac{\lambda_2-\lambda_1}{\lambda_2+\lambda_1}\arctan(\alpha)\right), \tag{15}$$

$$\lambda_1 = \sqrt{\frac{10^{CLD/10}}{1+10^{CLD/10}}}, \tag{16}$$

$$\lambda_2 = \sqrt{\frac{1}{1+10^{CLD/10}}}, \tag{17}$$

and

$$R_a^{n,m} = \begin{cases} R_a^{l,m}\alpha(n,l)+(1-\alpha(n,l))R_a^{-1,m}, 0 \le n \le \mathrm{t}(l), l=0 \\ R_a^{l,m}\alpha(n,l)+(1-\alpha(n,l))R_a^{l-1,m}, \mathrm{t}(l-1) \le n \le \mathrm{t}(l), 1 \le l < L \end{cases} \tag{18}$$

for $0 \le l < L$ where

$$\alpha(n,l) = \begin{cases} \dfrac{n+1}{\mathrm{t}(l)+1} & , \forall n; l=0 \\ \dfrac{n-\mathrm{t}(l-1)}{\mathrm{t}(l)-\mathrm{t}(l-1)}, & \forall n; l: otherwise \end{cases}, \tag{19}$$

and where L denotes the number of parameter sets, and $\mathrm{t}(l)$ denotes the time slot of parameter set l.

The matrix operations used for TTT have different modes, prediction mode and energy reconstruction mode, which are fully introduced in [3].

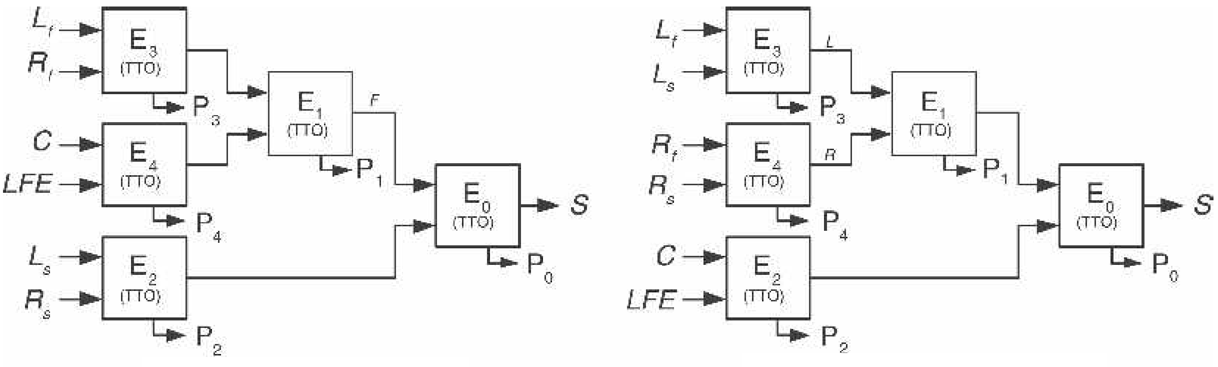
Tree Structures

The common coding blocks for MPEG Surround mentioned in the previous subsection can be cascaded to form the desired spatial coding tree, depending on the specified numbers of inputs and outputs, and additional features. Though MPS can deal with the sequences with 7.1 channels, only the most common tree structures for 5.1 channel input would be described in this subsection.

The first type of tree structures for 5.1 channel input is 5151 and 5152, which supports a mono downmixed signal. Both 5151 and 5152 are shown in Figure 7a and Figure 7b respectively. The six input channels, left front (L), right front (R), left surround (Ls), right surround (Rs), center (C), and low frequency enhancement (LFE) are fed into TTO box pairwisely until a mono downmixed signal is obtained.

To downmix the inputs to the stereo signal, the second type of tree structures is provided, which is commonly known as 525. Figure 8 shows the preferred tree structure for stereo downmix. L and Ls, R and Rs, and C and LFE are processed by R-OTT box separately and then the outputs of the three TTO boxes are fed into a R-TTT box to generate a stereo downmixed signal.

To upmix the downmixed signal, the decoding tree structures are the inversion of the encoding one, which would give the most similar results. For detailed information, please refer to [1]-[3].



(a) 5151 configuration

(b) 5152 configuration

Figure 7 Tree configurations for mono downmix [2]

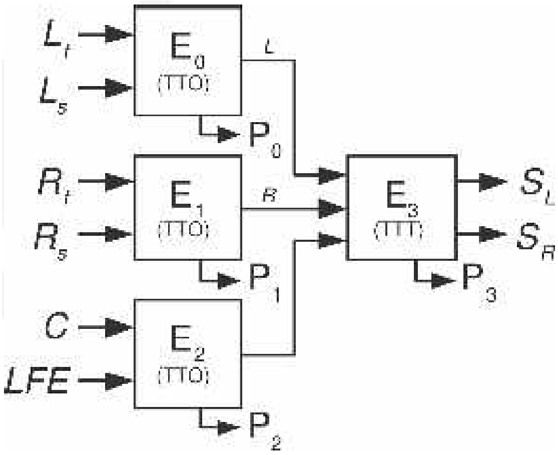


Figure 8 Preferred tree configuration for stereo downmix [2]

Chapter 3 The Objectives of MPS Coding

In order to let the perceptual feeling of decoded output to be the same as the feeling to the original input, there exist some objectives based on the statistical properties between channels. In this chapter, we introduce the objectives which are set for the upmix matrix in an OTT box and derive how they would be held.

Let $\boldsymbol{X}_m$ be a column vector representing a complex hybrid subband signal $X_m[n]$ in a processing frame, for $n = n_0, n_0+1, \ldots, n_1$, i.e., $\boldsymbol{X}_m = [X_m[n_0], X_m[n_0+1], X_m[n_0+2]\ldots, X_m[n_1]]^T$, where superscript T means the transpose operation and $n_0 = 0$ and $n_1 = 31$. An ideal decorrelator should generate an uncorrelated signal with the same energy of its input, i.e. $\mathrm{Re}\{<\boldsymbol{M}_m, \mathbf{D}_m>\} = 0$ and $||\boldsymbol{M}_m||^2 = ||\boldsymbol{D}_m||^2$, where Re{ • } takes the real part of an input; < • > means the inner product and || • || means the vector norm. From (12) to (17), if all the *CLD*s and *ICC*s are the same for the time slot n = n_0, $n_0+1, \ldots, n_1$, the energy of the upmixed signal $X'_{1,m}[n]$, with an ideal decorrelator, can be derived as

$$\begin{aligned}\left\|\boldsymbol{X'}_{\boldsymbol{1,m}}\right\|^2 &= \lambda_1^2 \cdot \begin{pmatrix} \cos^2(\alpha+\beta)\left\|\boldsymbol{M}_{\boldsymbol{m}}\right\|^2 + \sin^2(\alpha+\beta)\left\|\boldsymbol{D}_{\boldsymbol{m}}\right\|^2 \\ +2\cos(\alpha+\beta)\sin(\alpha+\beta)\,\mathrm{Re}\{<\boldsymbol{M}_{\boldsymbol{m}}, \boldsymbol{D}_{\boldsymbol{m}}>\} \end{pmatrix}. \\ &= \lambda_1^2 \cdot \left\|\boldsymbol{M}_{\boldsymbol{m}}\right\|^2 \end{aligned} \tag{20}$$

Likewise, the energy of the upmixed signal $X'_{2,m}[n]$ is reduced to

$$\left\|\boldsymbol{X'}_{2,m}\right\|^2 = \lambda_2^2 \cdot \left\|\boldsymbol{M}_m\right\|^2. \tag{21}$$

The correlation of the upmixed signals is derived as

$$\begin{aligned} &\mathrm{Re}\{<\boldsymbol{X'}_{1,m}, \boldsymbol{X'}_{2,m}>\} \\ &= \lambda_1\lambda_2 \cdot \mathrm{Re}\left\{\begin{matrix} \cos(\alpha+\beta)\cos(-\alpha+\beta)\cdot\left\|\boldsymbol{M}_m\right\|^2 \\ +\cos(\alpha+\beta)\sin(-\alpha+\beta)\cdot<\boldsymbol{M}_m, \boldsymbol{D}_m> \\ +\sin(\alpha+\beta)\cos(-\alpha+\beta)\cdot<\boldsymbol{D}_m, \boldsymbol{M}_m> \\ +\sin(\alpha+\beta)\sin(-\alpha+\beta)\cdot\left\|\boldsymbol{D}_m\right\|^2 \end{matrix}\right\}. \\ &= \lambda_1\lambda_2 \cdot \cos(2\alpha)\cdot\left\|\boldsymbol{M}_m\right\|^2. \end{aligned} \tag{22}$$

Objective I—

The correlations among the upmixed signals from multiple channels are the same as the correlations of the original signals.

According to the definition of the *ICC* parameter in the MPS standard, the *ICC* value of the two upmixed signals in a parameter band is calculated as

$$\Phi_{X_1X_2} = \mathrm{Re}\left(\frac{\sum_{m=m_b}^{m_{b+1}-1} < \boldsymbol{X'}_{1,m}, \boldsymbol{X'}_{2,m} >}{\sqrt{\sum_{m=m_b}^{m_{b+1}-1} \left\| \boldsymbol{X'}_{1,m} \right\|^2 \cdot \sum_{m=m_b}^{m_{b+1}-1} \left\| \boldsymbol{X'}_{2,m} \right\|^2}} \right), \tag{23}$$

From (20)-(22), under a ideal decorrelator, (23) can be reduced to

$$\Phi_{X_1X_2} = \cos(2\alpha). \tag{24}$$

Comparing (24) with (14) shows that Objective I holds with an ideal decorrelator.

Objective II—

The power ratios among the upmixed signals from different channels are the same as those of the original audio signals.

According to the definition of the *CLD* parameter in MPS standard, the *CLD* value of the two upmixed signals in a parameter band is calculated as

$$\Delta L_{X_1X_2} = 10\log_{10}\left(\frac{\sum_{m=m_b}^{m_{b+1}-1} \left\| \boldsymbol{X'}_{1,m} \right\|^2}{\sum_{m=m_b}^{m_{b+1}-1} \left\| \boldsymbol{X'}_{2,m} \right\|^2} \right). \tag{25}$$

By (20) and (21), with an ideal decorrelator, (25) can be reduced as

$$\Delta L_{X_1X_2} = 10\log_{10}\left(\frac{{\lambda_1}^2}{{\lambda_2}^2} \right). \tag{26}$$

From (16) and (17), this result shows that Objective II holds with an ideal decorrelator.

Objective III—

The sum of the energies of the upmixed signals from multiple channels must be equal to the energy of the input signals.

From (16) to (17) and (20) to (21), the energy of the two upmixed signals in a parameter band is calculated as

$$\sum_{m=m_b}^{m_{b+1}-1}\left\|\boldsymbol{X'}_{1,m}\right\|^2+\sum_{m=m_b}^{m_{b+1}-1}\left\|\boldsymbol{X'}_{2,m}\right\|^2=\sum_{m=m_b}^{m_{b+1}-1}\left\|\boldsymbol{M}_m\right\|^2. \tag{27}$$

This means that, if the encoder generates an energy-preserved downmixed signal, Objective III holds with an ideal decorrelator.

The above three objectives are held under the circumstances that Re{<$\boldsymbol{M}_m$, $\mathbf{D}_m$>} = 0 and $||\boldsymbol{M}_m||^2 = ||\boldsymbol{D}_m||^2$ and all the *CLD*s (2) and *ICC*s (3) are kept the same for the time slot n = n_0, n_0+1, …, n_1 while doing (12).

Chapter 4 Design of Downmixer and Decorrelator based on Gram-Schmidt Orthogonal Process

As shown in the previous chapter, to preserve the objectives of MPS coding, it is essential that a decorrelator generates an uncorrelated signal with the same energy of its input. An energy-preserved encoder is also important. We would demonstrate the limits of the current decorrelator and encoder, and then propose our methods based on Gram-Schmidt orthogonal process.

4.1 Decorrelator Issue in Decoder

In this subsection, we analyze the effect of the decorrelator, which is a cascaded all-pass filters in MPS standard, on the basis of [13].

A general subband signal can be approximated as

$$x[n]=\sum_{k=0}^{M-1} a_k \exp(j\omega_k n)+a_n N[n], \tag{28}$$

where the subband signal contains *M* sinusoid signals and an AWGN component *N*[*n*]. The cross power spectral density of *x*[*n*] and the output signal *y*[*n*] from the all-pass filters can be derived as

$$\begin{aligned} S_{YX}(\omega) &= D(\omega)S_{XX}(\omega) \\ &= D(\omega)\left[\sum_{k=0}^{M-1}\left|a_k\right|^2\delta(\omega-\omega_k)+a_n\sigma_N^2\right], \end{aligned} \tag{29}$$

where $\delta(\omega)$ is the Dirac delta function, σ_N^2 is the variance of the AWGN noise signal *N*[*n*], and $D(\omega)$ is the frequency response of the decorrelator. The cross correlation at zero lag is calculated as

$$\begin{aligned} &R_{yx}[0] \\ &= \frac{1}{2\pi}\int_{0}^{2\pi} D(\omega)\left[\sum_{k=0}^{M-1}|a_k|^2\,\delta(\omega-\omega_k)+a_n\sigma_N^2\right]d\omega \\ &= \frac{1}{2\pi}\left[\sum_{k=0}^{M-1}|a_k|^2 D(\omega_k)+a_n\sigma_N^2\int_{0}^{2\pi} D(\omega)d\omega\right] \end{aligned} \tag{30}$$

We can consider (30) from two extreme cases. If the input signal is noise-like (a_k = 0 for all k), the cross correlation value can be approximated as

$$R_{yx}[0] \approx \frac{a_n\sigma_N^2}{2\pi}\cdot\int_{0}^{2\pi} D(\omega)d\omega = a_n\sigma_N^2\cdot d[0]. \tag{31}$$

where d[0] is the impulse response of the all-pass filters at n = 0. The value can be controlled to be zero for the all-pass filters. Hence, the cross correlation value can be zero if the input signal is white. However, if the input signal is purely tonal (a_n = 0 for all n), the cross correlation value is

$$R_{yx}[0] \approx \frac{1}{2\pi}\sum_{k=0}^{M-1}|a_k|^2 D(\omega_k)\,. \tag{32}$$

Thus, we can expect that the output signal of the decorrelator can be highly correlated to the original tonal signal. In other words, the effect of the all-pass filters defined in MPS standard has the decorrelating effect varying with the tonal components in the subband signals.

4.2 Downmixer Issue in Encoder

Most of the current downmix methods, such as the direct summation of the input signals, cannot guarantee that the energy of the output is the same as that of the input. This inconsistent makes the energy summation of the decoded output signals different than that of the original input signals. One straightforward way is to directly adjust the downmixed signals, but unexpected artifacts might arise due to the possible amplification of the noise.

4.3 Gram-Schmidt Orthogonalization Process

Gram-Schmidt[14][15] is a method to construct an orthonormal basis in a k-dimensional inner product space. This process takes a finite, linearly independent set {$\mathbf{v}_1$, $\mathbf{v}_2$, …, $\mathbf{v}_j$} for j $\leq$ k and generates an orthogonal set {$\mathbf{u}_1$, $\mathbf{u}_2$, …, $\mathbf{u}_j$}.

Figure 9 shows the first two steps of Gram-Schmidt process. We can consider $\mathbf{u}_1$ as the normalized vector of $\mathbf{v}_1$.

$$\mathbf{u}_1 = \left(\frac{1}{\|\mathbf{v}_1\|}\right)\mathbf{v}_1 \tag{33}$$

Let $\mathbf{p}_1$ denote the projection of $\mathbf{v}_2$ onto $\mathbf{v}_1$.

$$\mathbf{p}_1 = \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1 . \tag{34}$$

Therefore, we can get the orthogonal vector $\mathbf{v}_2 - \mathbf{p}_1$, which is on the unit vector $\mathbf{u}_2$ orthogonal to $\mathbf{u}_1$.

$$\mathbf{v}_2 \text{ - } \mathbf{p}_1 = \frac{-\langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{v}_1}{\|\mathbf{v}_1\|} + \mathbf{v}_2 . \tag{35}$$

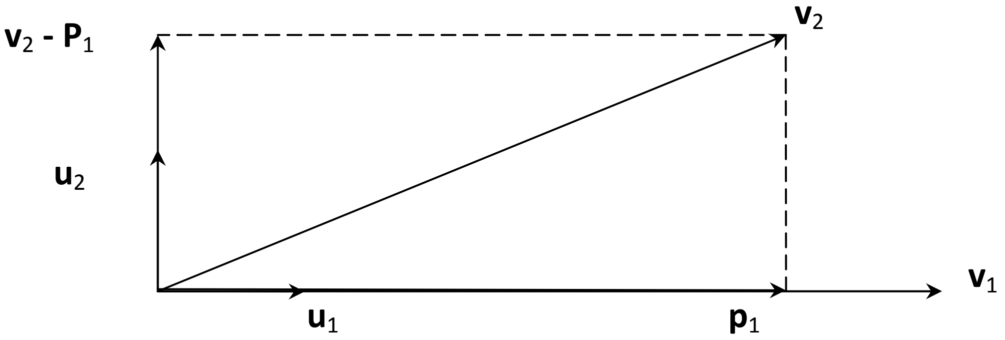


Figure 9 The first two steps of Gram-Schmidt process

The idea of Gram-Schmidt is explained by the first two steps of the process. We can repeat the idea and calculations shown from (33) to (35) to reach the desired k-dimension inner product space.

4.4 The GS-based Decorrelator

To ensure the uncorrelation between the decorrelated signal and the downmixed signal, a Gram-Schmidt based decorrelator (see Figure 10) is proposed to modify the output signal.

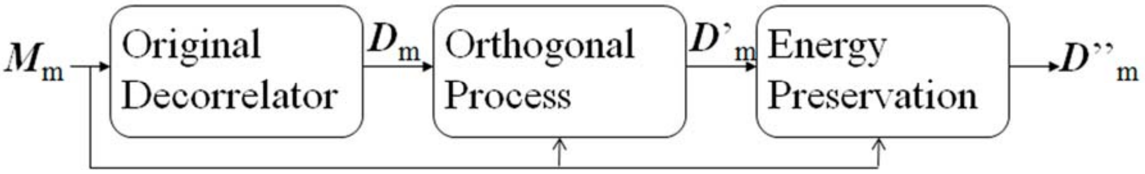


Figure 10 Flowchart of the proposed method to generate proper decorrelated signals

By applying the Gram-Schmidt orthogonalization process to { $\boldsymbol{M}_m$, $\boldsymbol{D}_m$ }, $\boldsymbol{D}_m$ can be decomposed as

$$\boldsymbol{D}_m = < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \boldsymbol{V}_m = \boldsymbol{P}_{1m} + \boldsymbol{V}_m . \tag{36}$$

Although $\boldsymbol{V}_m$ is uncorrelated to $\boldsymbol{M}_m$, the projection part of $\boldsymbol{D}_m$ is correlated to $\boldsymbol{M}_m$, which leads to the decorrelation defect. On second thought that the original decorrelator process changes the phase of the signal to realize orthogonality and only the real-part decorrelation is required, a modified decorrelated vector can be constructed by multiplying scale *j*, which means another 90° shift on the signal, to the projection part:

$$\boldsymbol{D}'_m = j < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \boldsymbol{V}_m . \tag{37}$$

The real part of the inner product of $\boldsymbol{M}_m$ and $\boldsymbol{D'}_m$ is given by

$$\mathrm{Re}(< \boldsymbol{D}'_m, \boldsymbol{M}_m >) = \mathrm{Re}(< \boldsymbol{V}_m, \boldsymbol{M}_m >) = 0, \tag{38}$$

and thus we confirm that $\boldsymbol{D'}_m$ is uncorrelated to $\boldsymbol{M}_m$. (see Figure 11)

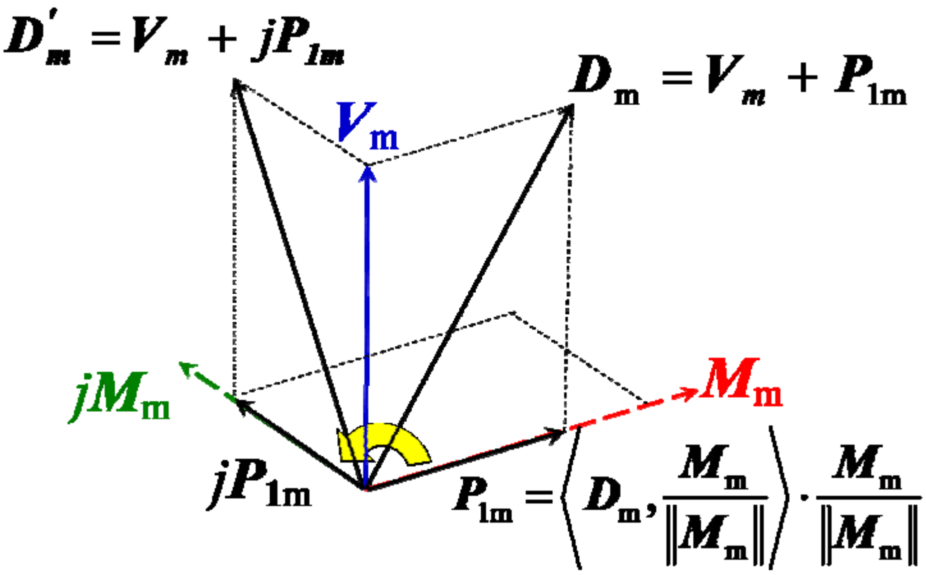

$D'_m = V_m + jP_{1m}$
$D_m = V_m + P_{1m}$
V_m
jM_m
M_m
jP_{1m}
$P_{1m} = \left\langle D_m, \frac{M_m}{\|M_m\|} \right\rangle \cdot \frac{M_m}{\|M_m\|}$

Figure 11 Illustration of the orthogonal process in the GS-based decorrelator

On the other hand, the energy of ***D′****m* is given by

$$\left\|\boldsymbol{D}'_m\right\|^2 = \left\|\boldsymbol{D}_m\right\|^2 = \left|< \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\left\|\boldsymbol{M}_m\right\|} >\right|^2 + \left\|\boldsymbol{V}_m\right\|^2 . \qquad (39)$$

The energy of ***D′****m* is not guaranteed to be the same as the energy of ***M****m* since ***jP****1m* is correlated to ***V****m*. Energy adjustment is required to ensure the equality of the energies of the decorrelated and downmixed signals. From the decoder's point of view, ***M****m* is transmitted from the encoder, which can be considered as the signal similar with the original input signals. Therefore, ***jP****1m* can be amplified without the risk of other artifacts if the energy of ***D′****m* is less than the energy of ***M****m*. In the opposite case, if the energy of ***D′****m* is larger than the energy of ***M****m*, keeping the same relation of ***jP***$_{1\mathrm{m}}$ and ***V***$_{\mathrm{m}}$ which is generated by original decorrelator, is important since reduction can be consider as a suppression on the component.

We verify our thoughts by different method combinations of increasing or decreasing the energy. The two ways to increase the energy, called Increase A and B, and two ways of decrease the energy, called Reduction A and B, are listed below:

- Increase A:

$$\boldsymbol{D}_m'' = \rho\left(j < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \boldsymbol{V}_m \right), \tag{40}$$

with

$$\rho = \|\boldsymbol{M}_m\| / \|\boldsymbol{D}_m\|. \tag{41}$$

- Increase B:

$$\boldsymbol{D}_m'' = j\rho < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \boldsymbol{V}_m, \tag{42}$$

with

$$\rho = \sqrt{\|\boldsymbol{M}_m\|^2 - \|\boldsymbol{V}_m\|^2} \Bigg/ \left| < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \right| . \tag{43}$$

- Reduction A:

$$\boldsymbol{D}_m'' = \rho\left(j < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \boldsymbol{V}_m \right), \tag{44}$$

with

$$\rho = \|\boldsymbol{M}_m\| / \|\boldsymbol{D}_m\|. \tag{45}$$

- Reduction B:

$$\boldsymbol{D}_m'' = j < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} + \rho \boldsymbol{V}_m, \tag{46}$$

with

$$\rho = \sqrt{\|\boldsymbol{M}_m\|^2 \left(1 - \left| \left\langle \boldsymbol{D}_m, \boldsymbol{M}_m / \|\boldsymbol{M}_m\|^2 \right\rangle \right| \right)^2 \Big/ \|\boldsymbol{V}_m\|} . \tag{47}$$

If ρ in (47) is not in the range of 0 to 1, the surplus energy would be reduced from the projection part only as

$$\boldsymbol{D}_m'' = j\rho < \boldsymbol{D}_m, \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|} > \cdot \frac{\boldsymbol{M}_m}{\|\boldsymbol{M}_m\|}, \tag{48}$$

with

$$\rho = 1 \Big/ \left| \left\langle \boldsymbol{D}_m , \boldsymbol{M}_m \Big/ \left\| \boldsymbol{M}_m \right\|^2 \right\rangle \right| . \tag{49}$$

The best combination is the one using Increase B for increasing energy and Reduction A for decreasing energy, which agrees our primary thought on energy adjustment, after we compare the improvements of averages objective quality measurements (ODG) of different combinations, listed in Table 1.

Table 1 ODG improvements of different combinations for GS-based decorrelator

Combinations Comparison	Increase A + Reduction A	Increase A + Reduction B	Increase B + Reduction A	Increase B + Reduction B
ODG Improvement	0.04	-0.4	0.18	-0.28

4.5 The GS-based Downmixer

The advantages of direct summation of the input signals are trivial, but this method does not guarantee that the energy of the output is the same as that of the input. From (1), the energy of ***Y***$_\mathrm{m}$ can be derived as

$$\left\| \boldsymbol{Y}_m \right\|^2 = \left\| \boldsymbol{X}_{1,m} \right\|^2 + \left\| \boldsymbol{X}_{2,m} \right\|^2 + 2 < \boldsymbol{X}_{1,m} , \boldsymbol{X}_{2,m} > . \tag{50}$$

If the correlation exists among ***X***$_{1,\mathrm{m}}$ and ***X***$_{2,\mathrm{m}}$, the term 2<***X***$_{1,\mathrm{m}}$, ***X***$_{2,\mathrm{m}}$> exists and the energy of ***Y***$_\mathrm{m}$ is not equal to the sum of energies of ***X***$_{1,\mathrm{m}}$ and ***X***$_{2,\mathrm{m}}$. Therefore, a Gram-Schmidt based downmix method, based on direct summation of the input signals, is proposed. The flowchart of the proposed downmix method is depicted in Figure 12, where ***Max***$_\mathrm{m}$ is the input signal with larger energy, ***Min***$_\mathrm{m}$ is the input signal with smaller energy, ***P***$_\mathrm{m}$ is the sum of projection part of ***Min***$_\mathrm{m}$ and ***Max***$_\mathrm{m}$, and ***U***$_\mathrm{m}$ is the orthogonal part of ***Min***$_\mathrm{m}$. Based on the same idea of GS-based decorrelator, the Gram-Schmidt process can be applied to the input vector set {***L***$_m$, ***R***$_m$}.

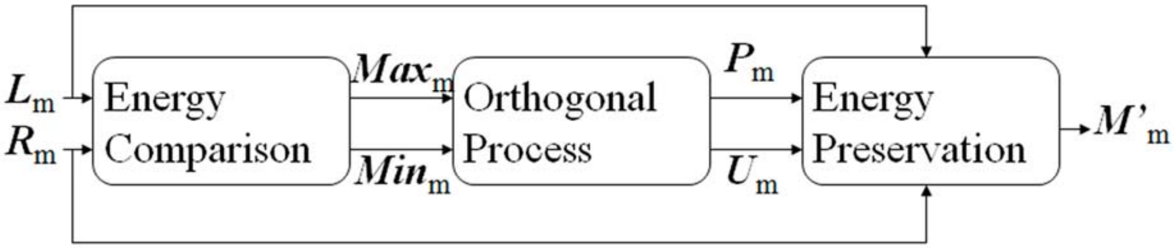


Figure 12 Flowchart of the proposed downmix method

In the beginning, we simply made $\boldsymbol{R}_m$ as the projection basis and decomposed $\boldsymbol{L}_\mathrm{m}$. For most of the test sequences, the ODGs improved. However, there were some results of test sequences, like sm01.wav, were even worse than the one processed by the original encoder. After analyzing the spectrogram of sm01.wav shown in Figure 13, we discovered that the choice of the basis to project on is crucial. The right channel has less energy than the left channel, so the Gram-Schmidt process would decompose the channel with more energy, which might lead to information loss for downmixing. Therefore, the first stage is to compare the energies of two input hybrid subbands and choose the one with larger energy to be the basis of projection. The average improvements for different strategies of projection are listed in Table 2, which is based on the encoder method published in [16].

Table 2 ODG improvements on different strategies of projection

Proj. Stgy. Comparison	Right Channel	Channel with More Energy
ODG Improvement	0.22	0.3

After comparing the energies of $\boldsymbol{L}_m$ and $\boldsymbol{R}_m$, let $\boldsymbol{Max}_m$ represents the one with larger energy and $\boldsymbol{Min}_m$ represents the other one.

The second stage (see Figure 14) perform the Gram-Schmidt process to decompose $\boldsymbol{Min}_m$ as

$$Min_m = < Min_m, \frac{Max_m}{\|Max_m\|} > Max_m + U_m, \tag{51}$$

where $\boldsymbol{U_m}$ is the uncorrelated with respect to the projection basis.

Therefore, the downmixed signal $\boldsymbol{M}_m$ can be expressed as

$$M_m = Max_m + < Min_m, \frac{Max_m}{\|Max_m\|} > Max_m + U_m = P_m + U_m. \tag{52}$$

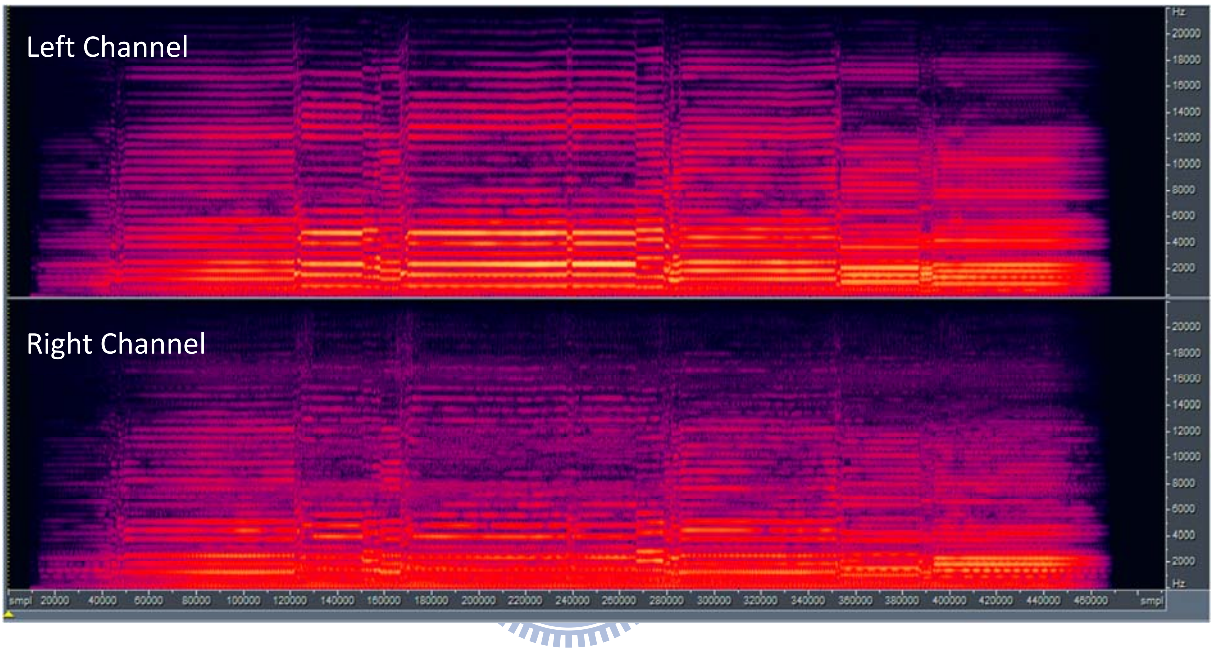


Figure 13 Spectrogram of unprocessed sm01.wav

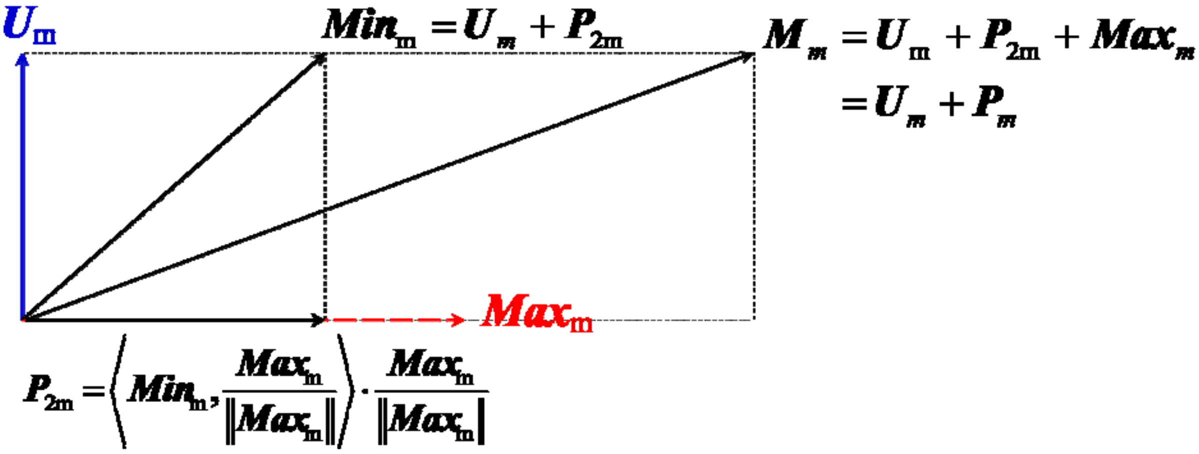

U_m
$Min_m = U_m + P_{2m}$
$M_m = U_m + P_{2m} + Max_m$
$= U_m + P_m$
Max_m
$P_{2m} = \left\langle Min_m, \frac{Max_m}{\|Max_m\|} \right\rangle \cdot \frac{Max_m}{\|Max_m\|}$

Figure 14 Illustration of the orthogonal process in the GS-based downmixer

The energy of $\boldsymbol{M}_m$ is not guaranteed to be the same as the summation of the energies of $\boldsymbol{L}_m$ and $\boldsymbol{R}_m$, which is shown in (50). Energy adjustment is required to ensure the equality of the energies of the downmixed and inputs signals. If the energy of $\boldsymbol{M}_m$ is less than the summation of the energies of $\boldsymbol{L}_m$ and $\boldsymbol{R}_m$, amplifying their same component is a better way for energy adjustment, which is based on the idea that it is important to keep most of the common component in both inputs. If the energy of $\boldsymbol{M}_m$ is more than the summation of the energies of $\boldsymbol{L}_m$ and $\boldsymbol{R}_m$, keeping the different component without losing the same one is better.

We verify the different method combinations for energy adjustment. The ways for increasing the energy of $\boldsymbol{M}_\mathrm{m}$, called Increase 1 and 2, and decreasing the energy of $\boldsymbol{M}_\mathrm{m}$, called Reduction 1 and 2, are listed below:

- Increase 1: the insufficient energy is adjusted by the entire $\boldsymbol{M}_m$

$$\boldsymbol{M}'_m = \rho(\boldsymbol{P}_m + \boldsymbol{U}_m), \tag{53}$$

with

$$\rho = \sqrt{\|\boldsymbol{L}_m\|^2 + \|\boldsymbol{R}_m\|^2 / \|\boldsymbol{M}_m\|^2} \ . \tag{54}$$

- Increase 2: the insufficient energy is adjusted by $\boldsymbol{P}_m$

$$\boldsymbol{M}'_m = \rho\boldsymbol{P}_m + \boldsymbol{U}_m , \tag{55}$$

with

$$\rho = \sqrt{\|\boldsymbol{L}_m\|^2 + \|\boldsymbol{R}_m\|^2 - \|\boldsymbol{U}_m\|^2 / \|\boldsymbol{P}_m\|^2} \ . \tag{56}$$

- Reduction 1: the surplus energy would be reduced from the entire $\boldsymbol{M}_m$ as

$$\boldsymbol{M}'_m = \rho(\boldsymbol{P}_m + \boldsymbol{U}_m), \tag{57}$$

with

$$\rho = \sqrt{\|\boldsymbol{L}_m\|^2 + \|\boldsymbol{R}_m\|^2 / \|\boldsymbol{M}_m\|^2} \ . \tag{58}$$

- Reduction 2: the surplus energy would be reduced from $\boldsymbol{U}_m$ first as

$$\boldsymbol{M}'_m = \boldsymbol{P}_m + \rho \boldsymbol{U}_m , \tag{59}$$

with

$$\rho = \sqrt{\left\|\boldsymbol{L}_m\right\|^2 + \left\|\boldsymbol{R}_m\right\|^2 - \left\|\boldsymbol{P}_m\right\|^2 \Big/ \left\|\boldsymbol{U}_m\right\|^2} \; . \tag{60}$$

If ρ in (58) is not in the range of 0 to 1, the surplus energy would be reduced from $\boldsymbol{P}_m$ only as

$$\boldsymbol{M}'_m = \rho \boldsymbol{P}_m , \tag{61}$$

with

$$\rho = \sqrt{\left\|\boldsymbol{L}_m\right\|^2 + \left\|\boldsymbol{R}_m\right\|^2 \Big/ \left\|\boldsymbol{P}_m\right\|^2} \; . \tag{62}$$

After comparing with the improvements of average ODGs shown in Table 3, the combination of Increase 2 and Reduction 1 gives the best result, which fits the energy adjustment principle.

Table 3 ODG improvements of different reduction method for GS-based downmixer

Combinations Comparison	Increase 1 + Reduction 1	Increase 1 + Reduction 2	Increase 2 + Reduction 1	Increase 2 + Reduction 2
ODG Improvement	0.32	0.3	0.45	0.42

Chapter 5 Experiments and Results

The proposed methods in chapter 4 are implemented on the MPEG Surround reference software and evaluated through the objective and subjective quality tests by plenty of stereo and surround test sequences. In this chapter, we introduce the software, list the test sequences we use, show the differences we observe, and experiment through quality tests in this chapter.

5.1 MPEG Surround Reference Software

The MPEG Surround committee provides a reference software for MPEG Surround [5]. The software is developed in Microsoft Visual Studio environment and is written in the programming language of C. Because the specification defines the MPEG Surround decoding processes, the decoding reference software is supposed to be fully normative with the standard. However, the encoding process is not specified by the standard, so the encoder in the reference software only provides the simplest method that fits the required syntax on the specifications. There are only three downmix tree structures provided: 5151, 5152, and 525. The outputs of the encoder and decoder are not further compressed nor decompressed by any rate control.

5.2 Test tracks

Since the proposed methods are used for the OTT and TTO boxes, creating a stereo condition is also feasible for confirming the effect. We use twelve stereo test tracks recommended by MPEG and listed in Table 4. These tracks include the critical music balancing on the percussion, string, wind instruments, and human vocal. To simplify the effective tree structure to only one pair of TTO and OTT box, the stereo test tracks are extended to six-channel sequences by padding zeros in the C, Ls, Rs, and LFE channels, which would give the same results conducted by Parametric Stereo. There are also four surround test tracks from the MPEG.

Table 4 The twelve tracks recommended by MPEG [4][17]

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)

Remark:

(a) Transients.

(b) Tonal/Harmonic structure.

(c) Natural vocal (critical combination of tonal parts and attacks).

(d) Complex sound.

Table 5　Detail information of the equipments

Equipments		Information
Laptop	CPU	Intel(R) Core(TM)2 Duo CPU T7100 @ 1.80GHz
	Memory	2GB
	Sound Card	Intel 82801H (ICH8 Family) HD Audio Controller
	OS	Windows Vista
Headphone		Grado Prestige SR125i Headphone

5.3 Quality Measurements

The objective quality experiments in this thesis are evaluated by EAQUAL (Evaluation of Audio Quality) which simulates the perception by human ears. It is the realization of BS.1387 [18], recommended by ITU Radio Communication Sector. The range of the objective difference grade (ODG) is from -4 to 0 with 0 the best quality. To evaluate the performance of the proposed methods, four kinds of combinations are compared: Original Encoder (OE) + Original Decoder (OD), OE + Modified Decoder (MD), Modified Encoder (ME) + OD, and ME + MD. However, EAQUAL only takes one-channel or two-channel inputs. Therefore, after the surround tracks are processed, they are separated into stereo/mono tracks, except LFE, to calculate the ODG.

PsyTel Multiple Codec Evaluation Software [19], which follows the test method and impairment scale recommended by ITU-R BS. 1116 [20], is used for subjective listening test. The system allows blind comparison of multiple audio files. Subjects with proper assessment training could give grades in the range from 1 to 5 with 5 means imperceptible difference. This test includes 12 stereo test sequences. Also, the same laptop and earphone (see Table 5) are provided for all the subjects in the stereo condition.

5.4 Proposed Method Verification

It is obvious that the R_a upmix matrix works on every time slots and dissatisfies the requirement that all the *CLD*s and *ICC*s are the same for the time slot n = n_0, n_0+1,..., n_1. Since not all the prerequisites are satisfied, the actual three objectives cannot be held.

To verify whether the objectives mentioned in Chapter 3 would be kept under the proposed method, the interpolation method (18) has to be modified as

$$R_a^{n,m} = R_a^{l,m} \quad ,0 \le n \le t(l),0 \le l < L \tag{63}$$

Figure 15 shows the *CLD* and *ICC* difference percentage between the transmitted parameters and the ones calculated from the upmixed signals. Figure 16 shows the energy ratio between the original inputs and the upmixed outputs. The energy ratios of MEMD are all close to 100%. The reason for not exactly 100% is that the energy loss caused by time-frequency transformation after testing the energy ratios before and after the R_a upmix procedure and the ones from the output and input files. Table 6 shows the energy ratios of the sc02.wav, which losses the most energy in the stereo test sequences. Therefore, both of Figures 15 and 16 tell that using our proposed methods can satisfy the objectives mentioned in Chapter 3. Also, Figure 17 shows after satisfying the objectives, objective quality measurement gives better results for our proposed method.

Table 6 Energy ratios for sc02.wav

Ratio / Domain Test Track	Proposed Encoder		Proposed Decoder	
	Hybridband	Entire File	Hybridband	Entire File
sc02.wav	99.6%	100.0%	99.6%	96.1%

We also test the objective quality measurements of the stereo test sequences by using the original R_a interpolation method, showing in Figure 18. The ODGs show our proposed methods still give better results comparing with the original downmixer and decorrelator.

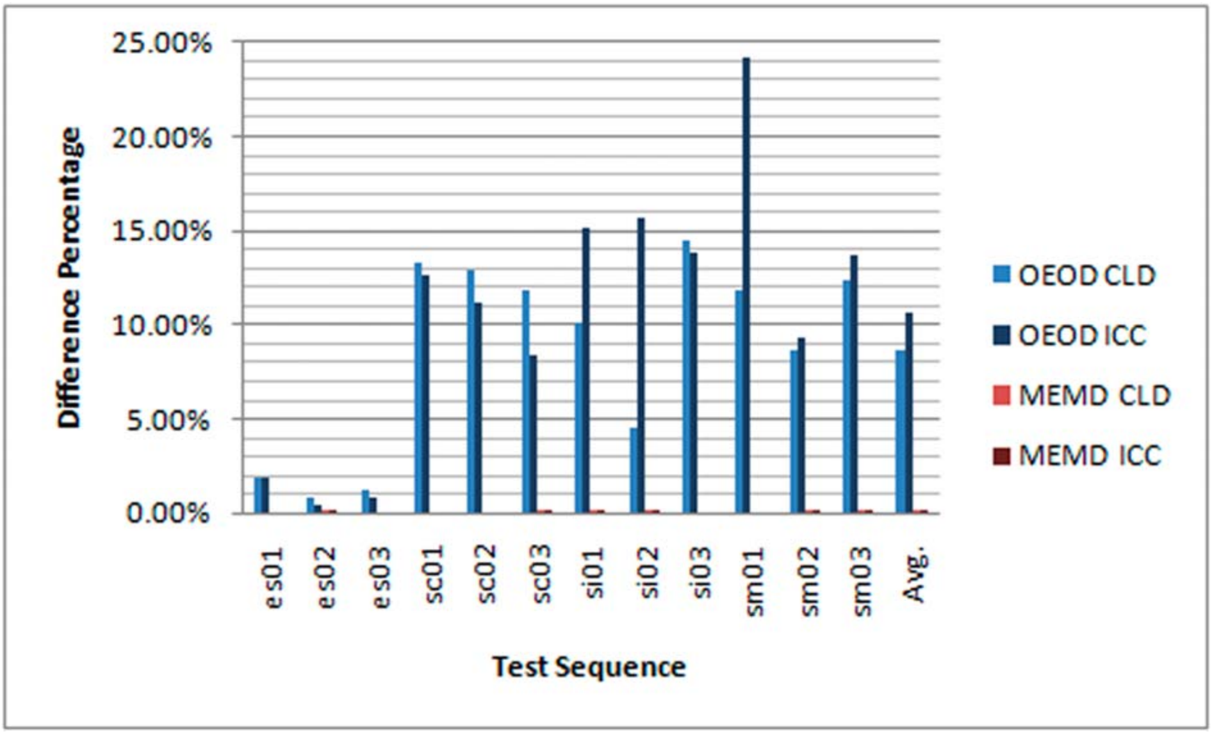


Figure 15 Parameter difference percentage before and after the upmix procedure

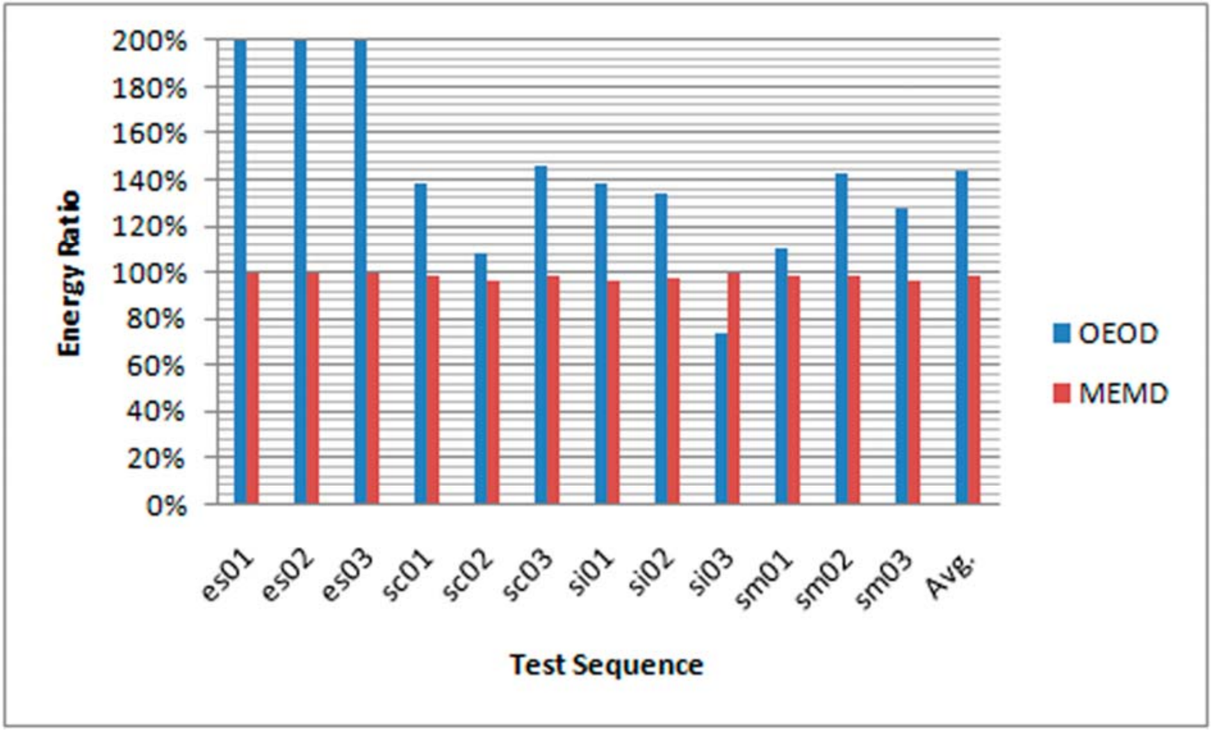


Figure 16 Energy ratio between the original inputs and upmixed outputs

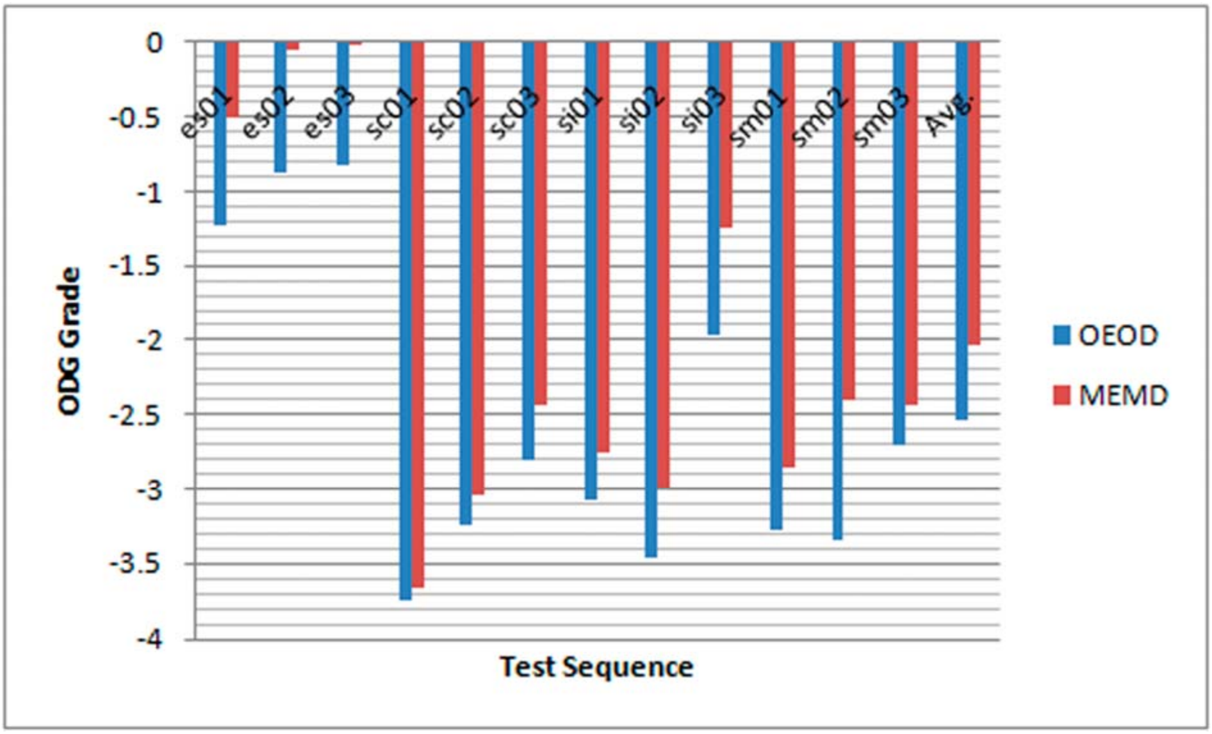


Figure 17 ODGs for stereo sequences by using the same parameter

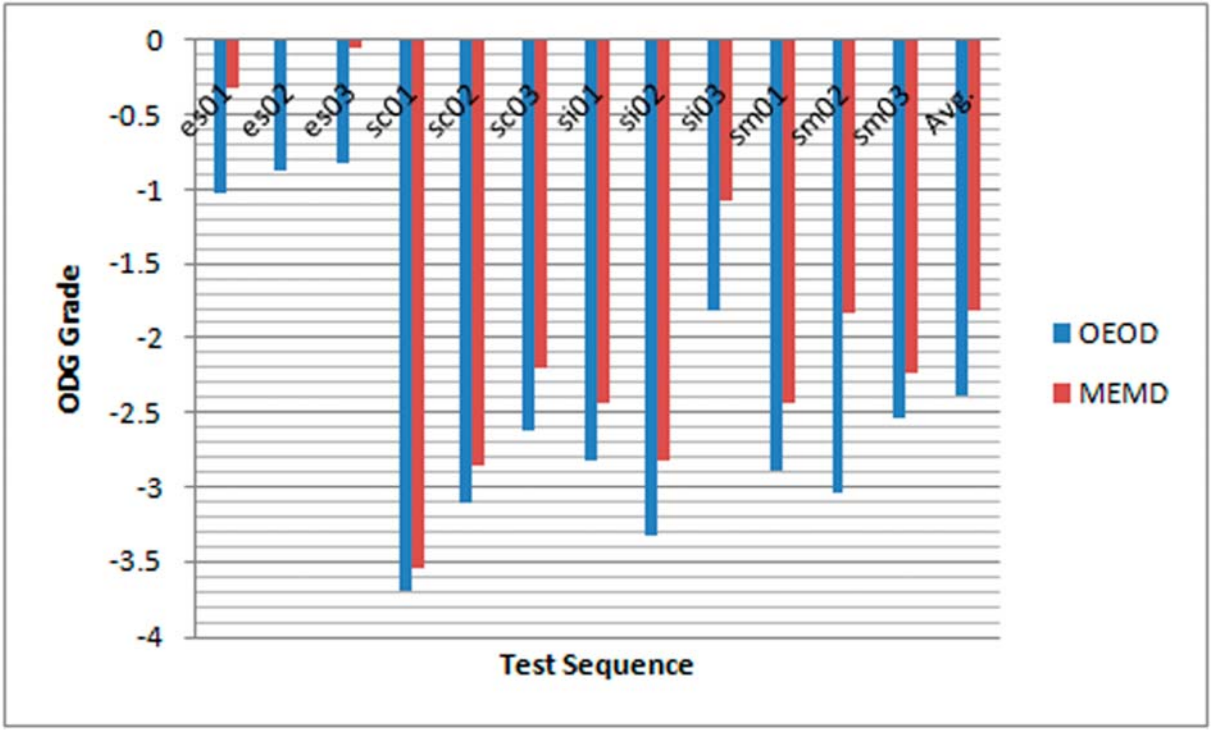


Figure 18 ODGs for stereo sequences by using the original interpolation on upmix coefficients

5.5 Waveform Example

Figures 19-21 show the waveform of the frame, which gives the most energy ratio difference by 201.1%, of es03.wav under different circumstances. Figure 20 shows obvious amplification on the waveform by using the original methods. On the contrary, the waveform shown in Figure 21, which is processed by the GS-based methods, is much similar with Figure 20 with the energy ratio of 100.6% in this frame.

5.6 Spectrogram Example

Figures 22-24 show the spectrograms of sm02.wav under different circumstances. In Figure 23, the spectrogram by the original methods shows some discontinuities. Also, the energy of each tone decays. On the contrary, in Figure 24, the spectrogram by the GS-based methods is smoother and preserves the energy.

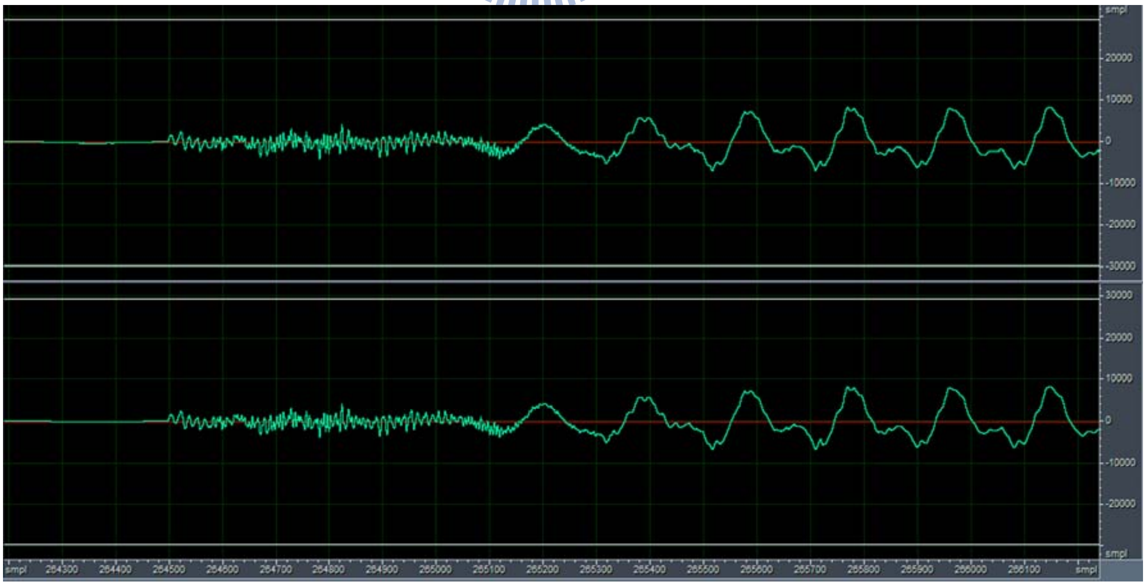


Figure 19 Waveform of original unprocessed es03.wav

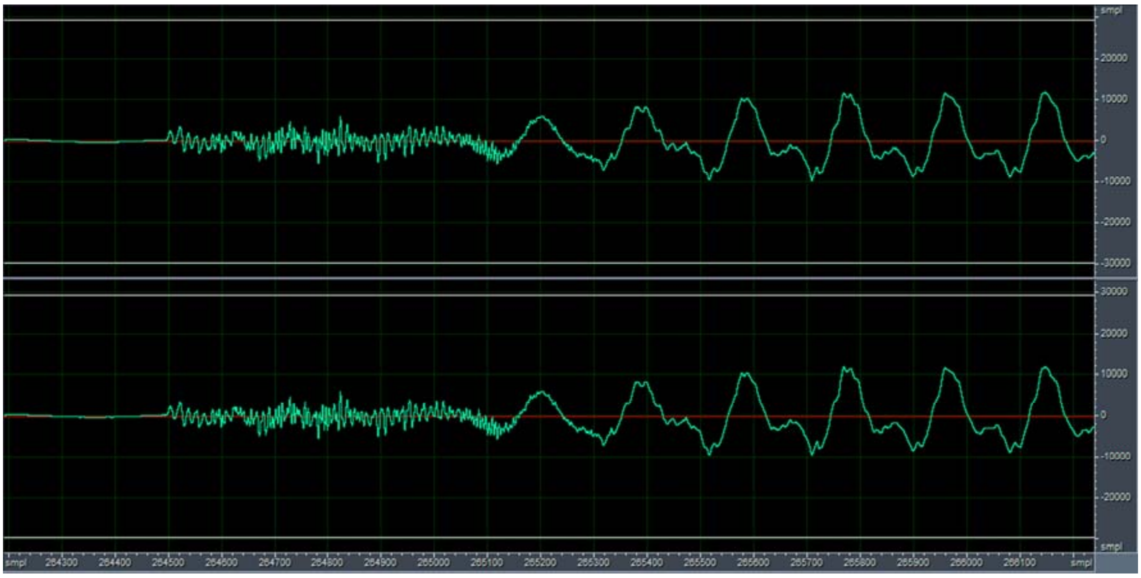


Figure 20 Waveform of es03.wav by the CODEC in standard

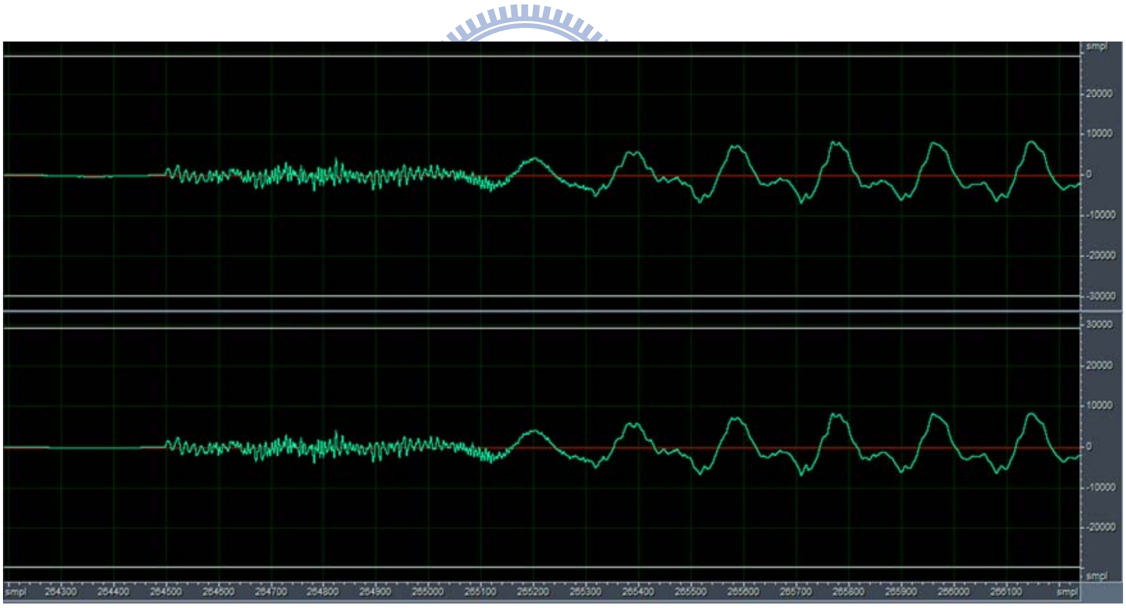


Figure 21 Waveform of es03.wav based on GS-based decorrelator and downmixer

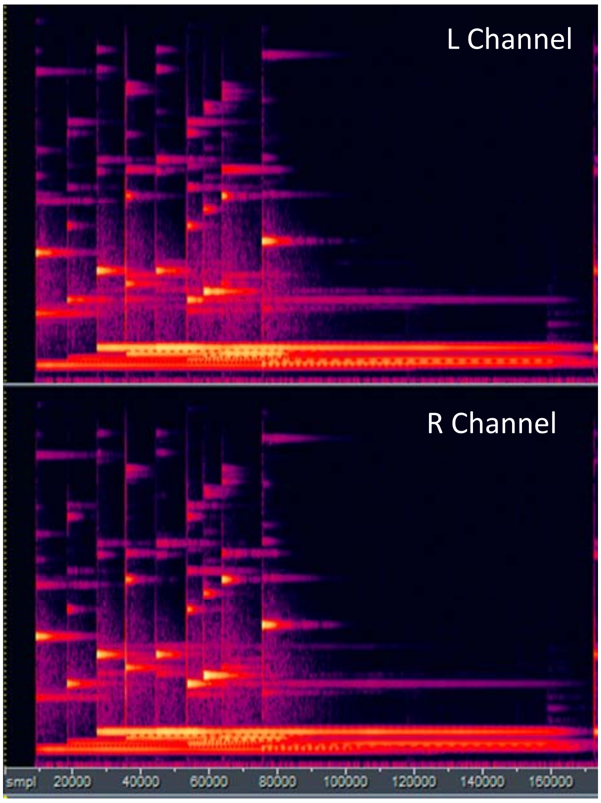


Figure 22 Spectrogram of original unprocessed sm02.wav

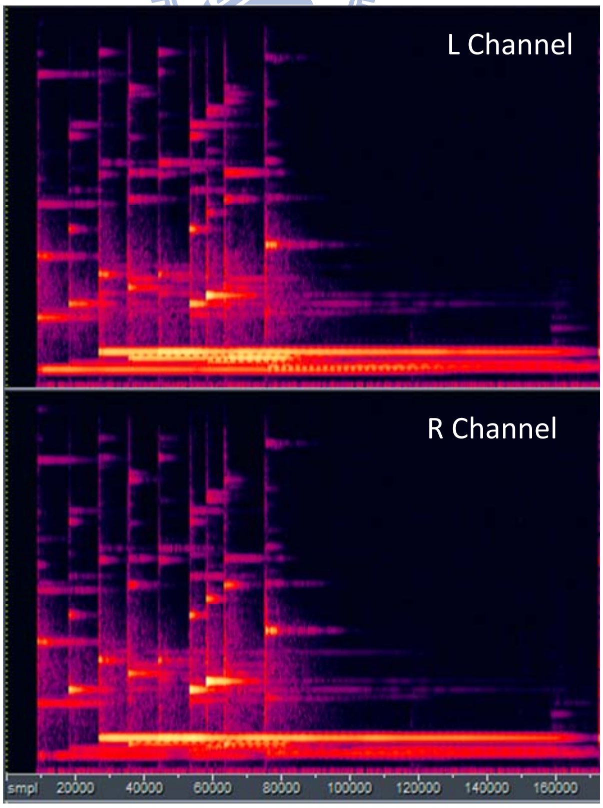


Figure 23 Spectrogram of sm02.wav by the CODEC in standard

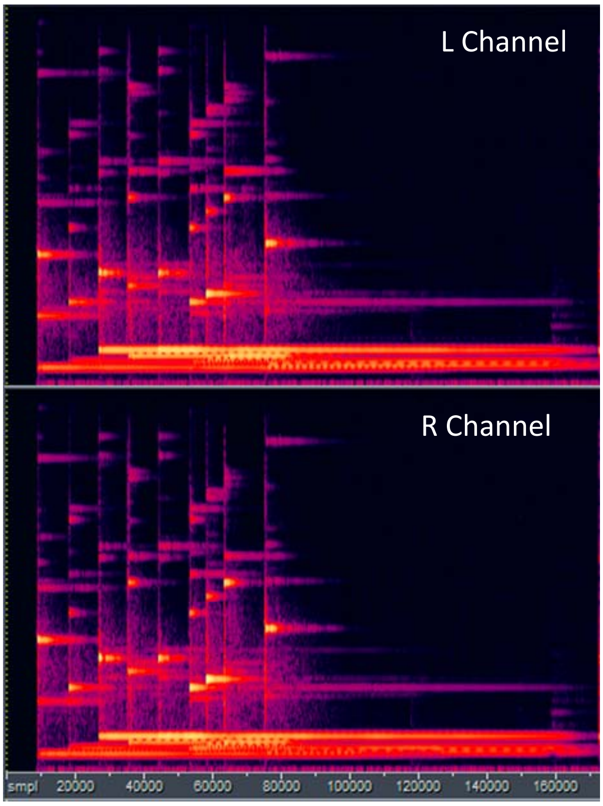


Figure 24 Spectrogram of sm02.wav based on GS-based decorrelator and downmixer

5.7 Results of Objective Quality Measure

The experiment results for the stereo condition are shown in Figure 25. In Figure 26 and Figure 27, four six-channel test sequences are conducted based on different coding tree structure. As shown in Figures 25-27, the proposed methods have improvements over the OE and OD in average. However, the improvement of surround sequences are much less than the ones given by stereo sequences. We erase the content by 2 channels to 4 channels in order to check whether we would get the similar results from the average improvements from the 12 stereo tracks. The corresponding average improvements are listed in Table 7. Even these MPEG four surround tracks are reduced to stereo sequences, the improvement of average ODGs is less than the one from the 12 stereo tracks, which is 0.56 on average. Therefore, these four MPEG surround tracks cannot reflect the benefit of our proposed methods.

Table 7 Improvements of the surround tracks under different erasure (5151)

Track Type	Surround Track			Stereo Track
Erasure	None	C, and LFE	Ls, Rs, C, and LFE	None
ODG Improvement	0.05	0.05	0.19	0.56

From the algorithms of the proposed methods, test tracks with high correlations between different channels are the ones that could give better results. To find the appropriate test tracks, we extract the six-channel audio from DVDs, such as operas, live concerts, and TV series, which are listed in Table 8. We also categorize the extracted audio into two categories: complex sound and natural vocal. The objective quality measurements of these two categories are shown in Figure 28 and Figure 29 with 0.18 and 0.5 ODG improvements on average correspondingly. Though these figures show better results compared with the one from MPEG surround test sequences, we still believe that we could get even better results if the sound tracks of DVD have not been post-produced. Figures 30-32 show the waveform of L and R of Friends.wav, which has the greatest improvement on ODG among surround cases.

We also notice our proposed methods make the ODG of Ls and Rs of pops.wav worse than the one from the original CODEC by 0.03 under the coding tree structure 5151. Since the energy adjustment for either GS-based decorrelator or GS-based downmixer is calculated based on a period of time, we think the problem may be caused by the inconsecutive of the energy adjustment coefficients between two processing periods. We fix this problem preliminary by interpolating the coefficients in the downmixer for the first ten time slots in a processing period and we got positive improvements on ODG of the Ls and Rs of pops.wav. However, this smoothing process causes other surround tracks worse than the one processed by original downmixer, shown in Figure 33. Therefore, this issue is considered as a future work from this thesis.

Table 8 Information about the extracted DVD audio

Track	Time (sec)	DVD Information
Vivaldi.wav	8	Vivaldi - The Four Seasons / Von Karajan, Mutter, Berlin Philharmonic (1987)
Aida.wav	15	Verdi / Pavarotti / Chiara / Dimitrova / Ghiaurov / Pons - Aida (1985)
LaGioconda.wav	15	Ponchielli - La Gioconda / Fischer, Marton, Domingo, Wiener Staatsoper (1986)
Celine.wav	13	Celine Dion: A New Day - Live in Las Vegas (2007)
Brightman.wav	10	Sarah Brightman - Live from Las Vegas (2004)
CelineTalk.wav	15	Celine Dion: A New Day - Live in Las Vegas (2007)
BrightmanTalk.wav	10	Sarah Brightman - Live from Las Vegas (2004)
Friends.wav	15	Friends: The Complete Tenth Season Disk 3 (2003)

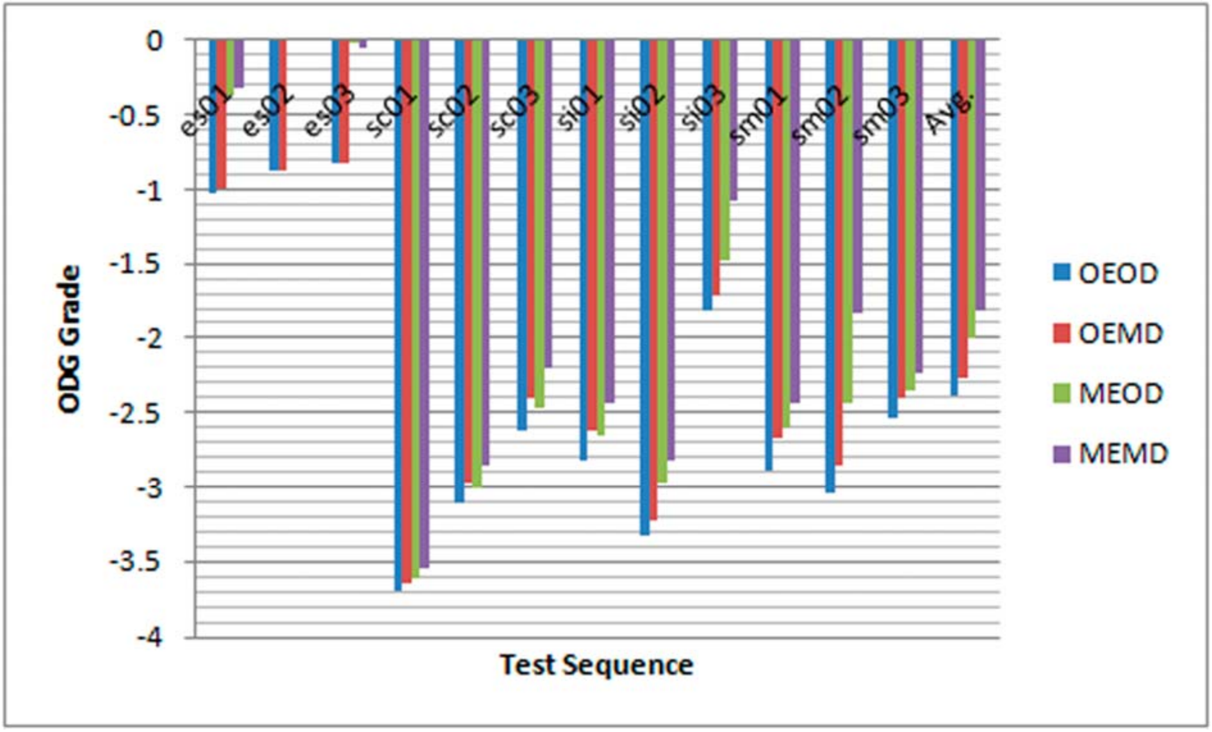


Figure 25 ODGs for stereo sequences by using 5151 tree structure

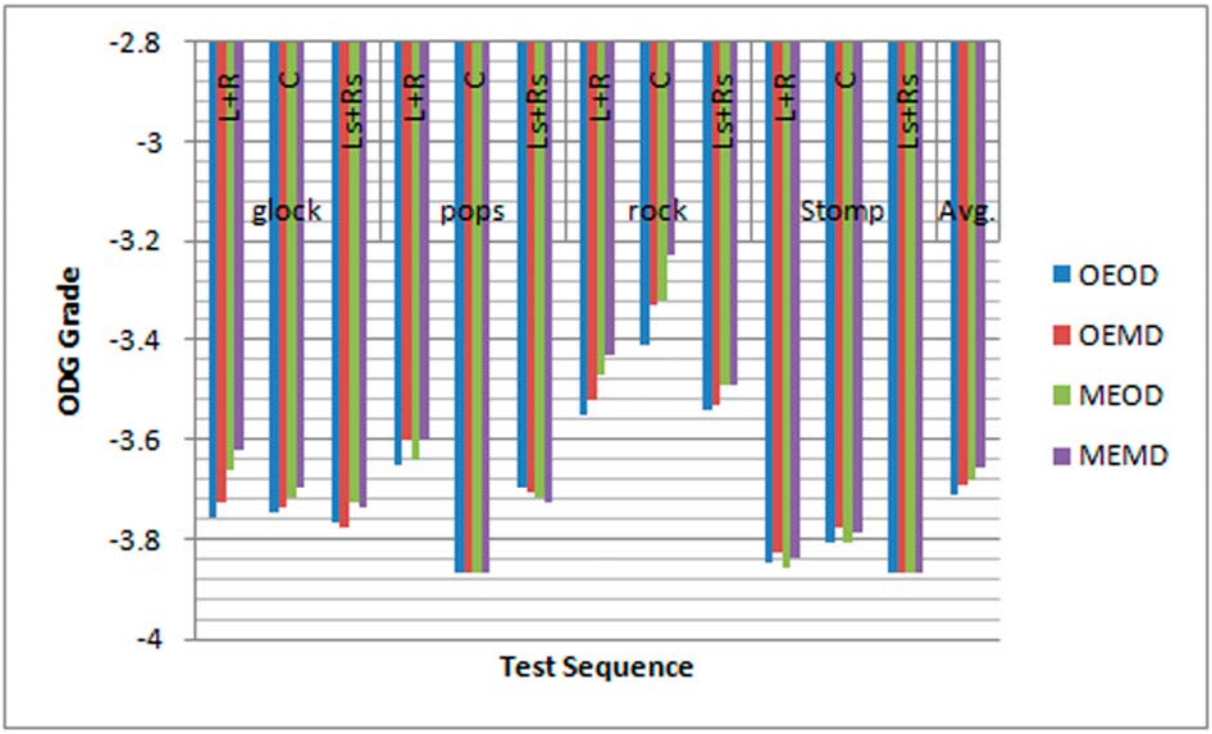


Figure 26 ODGs for MPEG surround sequences by using 5151 tree structure

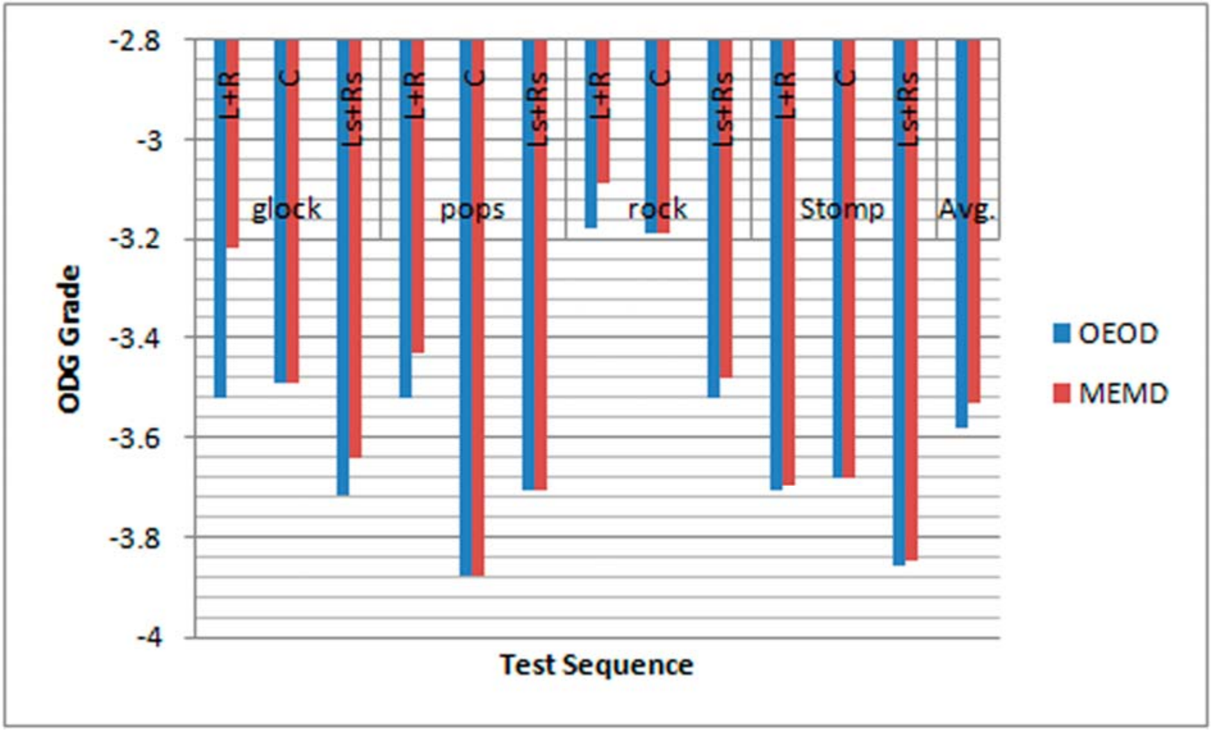


Figure 27 ODGs for MPEG surround sequences by using 525 tree structure

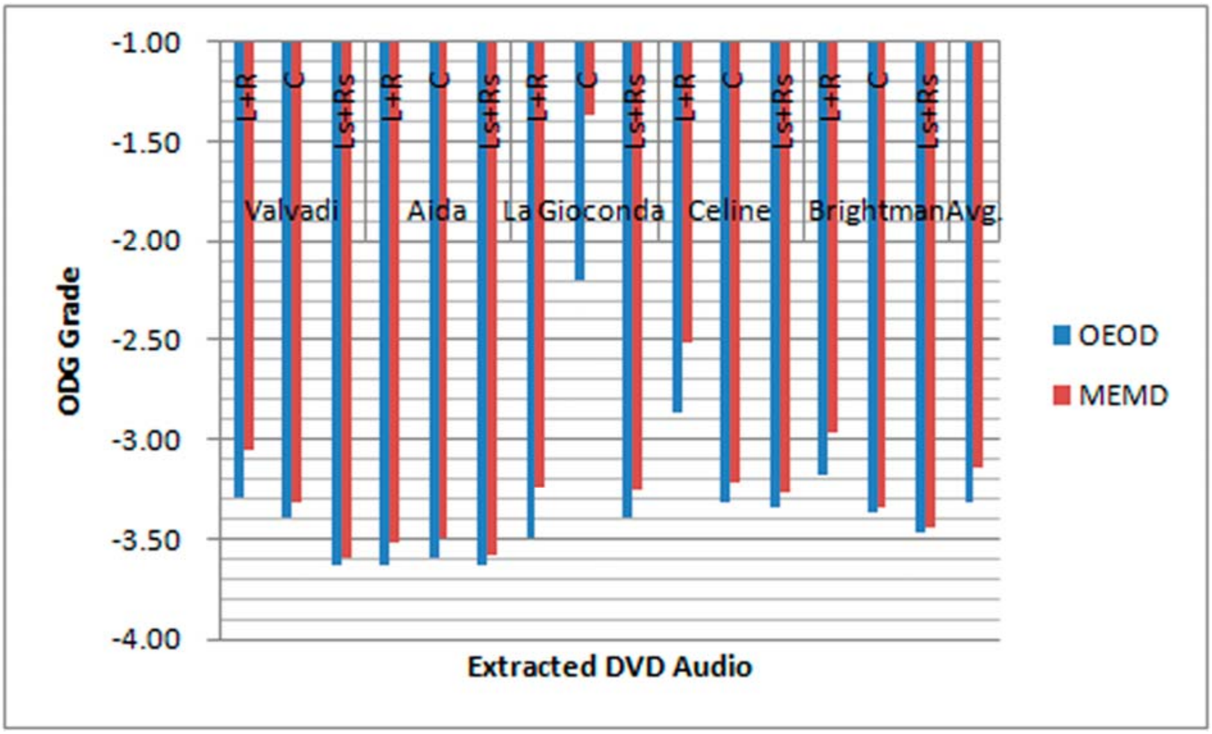


Figure 28 ODGs for complex sound category from extracted audio (5151)

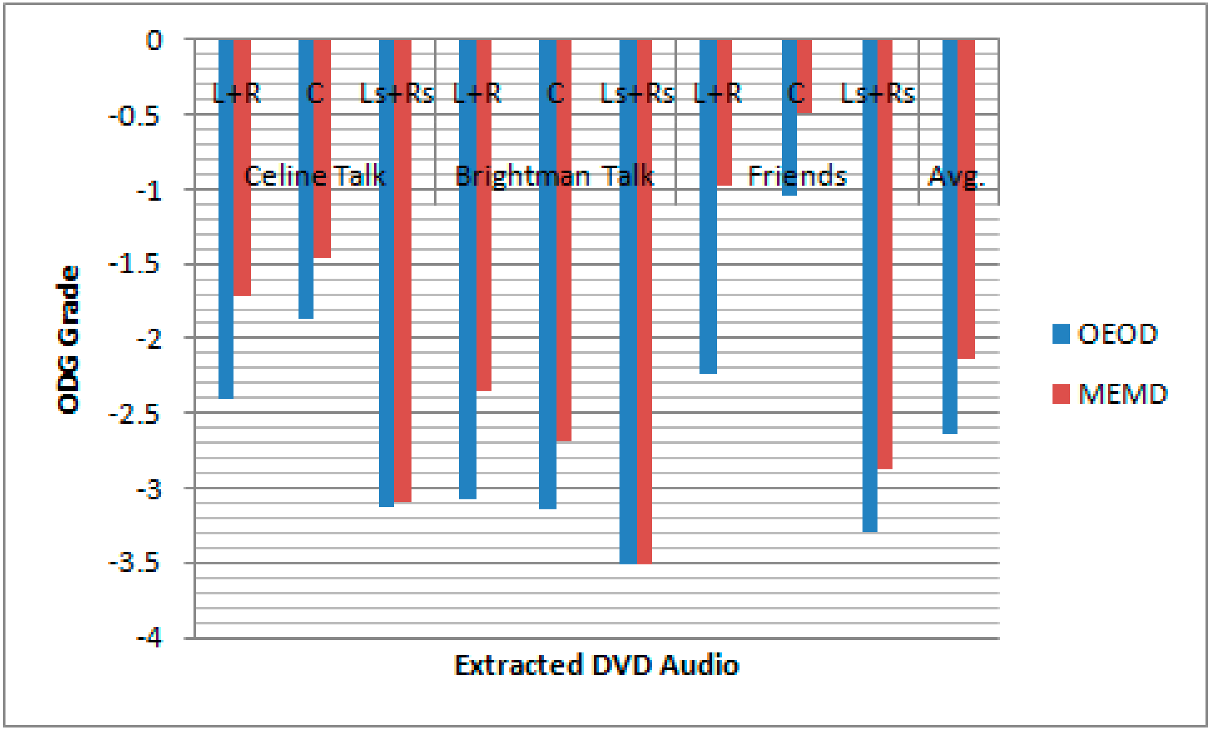


Figure 29 ODGs for natural vocal category from extracted audio (5151)

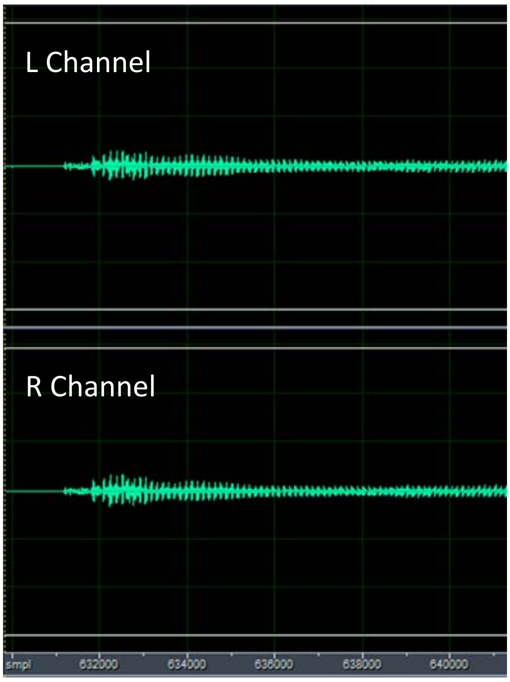


Figure 30 Waveform of original unprocessed L and R of Friends.wav

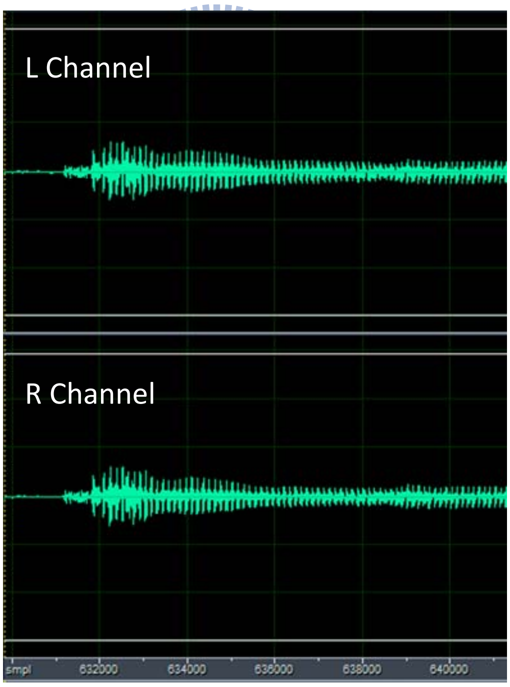


Figure 31 Waveform of L and R of Friends.wav by the CODEC in standard

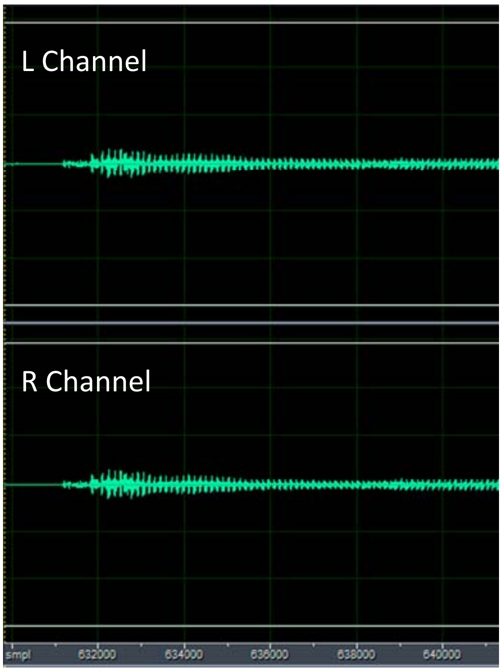


Figure 32 Waveform of L and R of Friends.wav based on GS-based methods

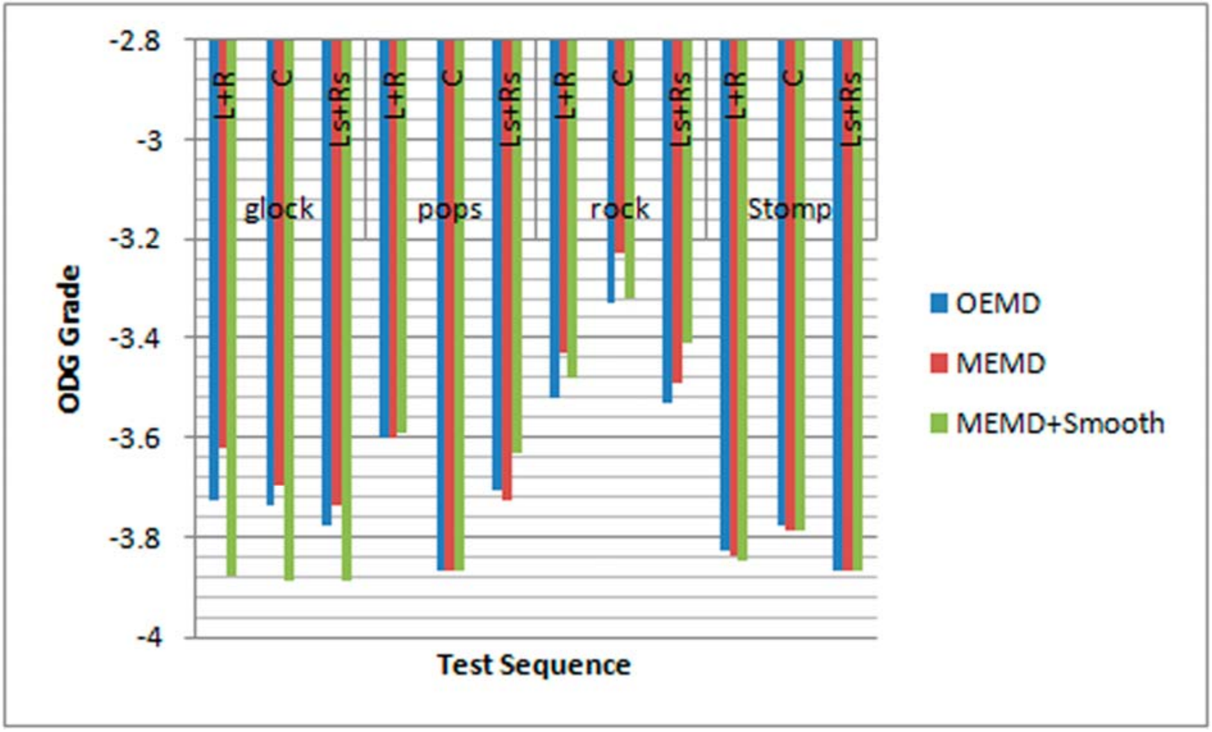


Figure 33 ODGs for MPEG surround sequences by using different CODEC combinations

5.8 Results of Subjective Quality Measure

There are 10 well-trained subjects join this experiment and the results are shown in Figure 34. From the 12 MPEG stereo test sequences, 7 test sequences have better results by the proposed GS-based methods. In average, the proposed methods have better performance.

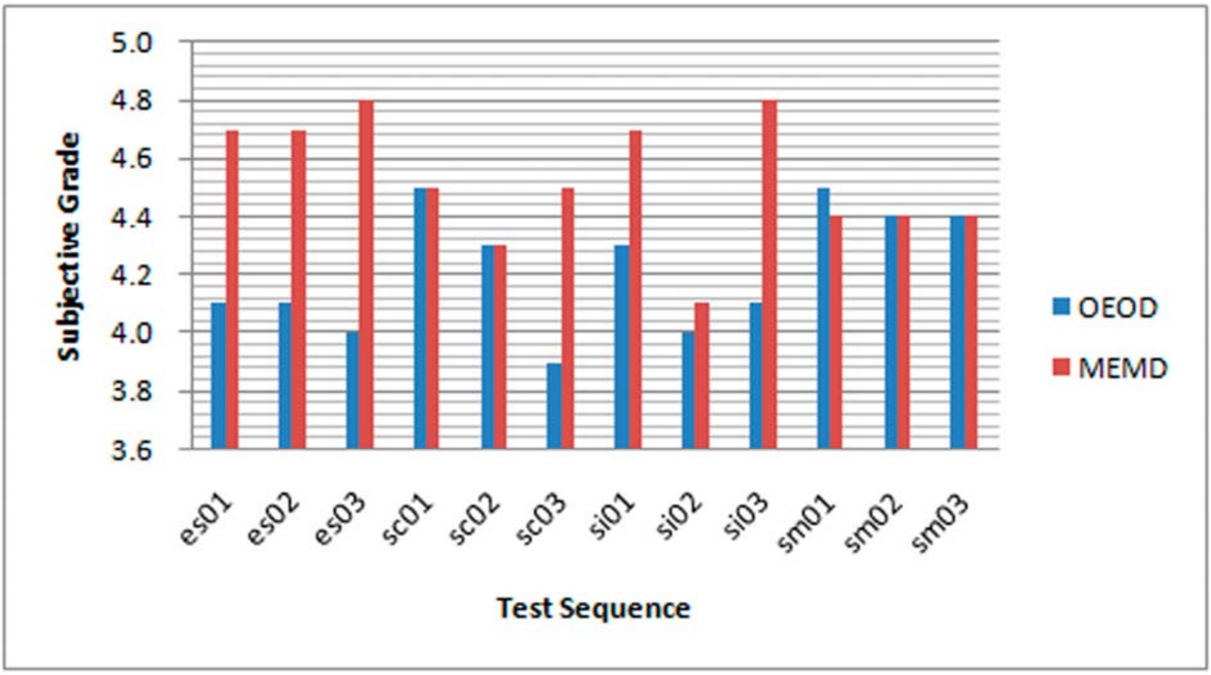


Figure 34 MUSHRA test on stereo tracks

Chapter 6 Conclusion and Future Work

This thesis has two contributions based on the GS process. First, a new decorrelator is proposed to generate the decorrelated signal to that can be uncorrelated with the downmixed signal in OTT box and preserve the energy constraint. Second, a new downmixer is proposed to keep the same energies between the output and the inputs in TTO box. From the derivation and the simulation results presented in this thesis, the GS-based downmixer and decorrelator can improve the methods defined in the MPS specification and can be applied individually to MPS encoder and decoder. Also, the resulted encoder and decoder comply with MPS CODEC, which means that the MPS bit streams can be generated or decoded without needing additional modification.

However, there are still some works in need of confirmation. One is whether the decorrelators should be mutually independent [3]. We have tried to figure out this problem by deriving the difference of the upmix process between the cascaded tree structure and the matrix form, but it seems no direct evidence that the decorrelators should be mutually independent. Another is the time-frequency transformation. From Figure 16, we can tell that the total energy of the outputs is much more similar with the total energy of the inputs, which is still not the same. After testing the energy ratio between the upmix processes, which is nearly 1 by using the proposed method, the energy difference is confirmed to be caused by time-frequency transformation and its inversion procedure. The other is smoothing process of the energy adjustment mentioned in Section 5.7. We believe that the smoothing may lead to better results, but the exact way for smoothing has not been clarified. Therefore, they are considered as the future works of this thesis.

References

[1] Information Technology MPEG Audio Technologies Part 1: MPEG Surround, ISO/IEC 23003-1.

[2] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. Van De Par, "Background, Concept and Architecture for the Recent MPEG Surround Standard on Multichannel Audio Compression," J. Audio Eng. Soc., vol. 55, no. 5, pp. 331–351, May 2007.

[3] J. Breebaart and C. Faler, "Spatial Audio Processing – MPEG Surround and Other Applications," John Wiley & Sons, New York, 2007.

[4] Y. H. Kao, "Design of Parametric Stereo Coding in MPEG HE-AAC," thesis for master, 2009.

[5] ISO/IEC JTC1/SC29/WG11 (MPEG), document N11204, "ISO/IEC 23003-1:2007/Amd.2:2008/Cor.2, MPEG Surround reference software update," Kyoto, Japan, Jan., 2010

[6] Y.W. Liu, and J. Smith, "Perceptually Similar Orthogonal Sounds and Applications to Multichannel Acoustic Echo Canceling," 22nd AES International Conference, 2002.

[7] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 156-165, 1998.

[8] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation – An overview of the fundamental problem," IEEE Signal Processing Lett., vol. 2, pp. 148-151, 1995.

[9] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," Proc. IEEE ICASSP, pp. 3689-3692, 1998.

[10] J. Engdegard, H. Purnhagen, J. Roden, L. Liljeryd, "Synthetic Ambience in Parametric Stereo Coding," Convention Paper 6074, 116th AES Convention, 2004.

[11] M. Boueri, and C. Kyirakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," Convention Paper 6291, 117th AES Convention, 2004.

[12] S. Torres-Guijarro, J. Beracoechea-Alava, I. Perez-Garcia, and F. Casajus-Quiros,"Coding strategies and quality measure for multichannel audio," Convention Paper 6114, 116th AES Convention, 2004.

[13] A. V. Oppenheim and R. W. Schafer, "Discrete Time Signal Processing, 2nd ed.," Prentice Hall, 1999

[14] S. J. Leon, “Linear Algebra with Applications, Sixth Ed.,” Prentice Hall, 2002.

[15] Gram-Schmidt Process, [online] http://en.wikipedia.org/wiki/Gram-schmidt. .

[16] D. P. Chen, H. F. Hsiao, H. W. Hsu, C. M. Liu, “Gram-Schmidt-based Downmixer and Decorrelator in the MPEG Surround Coding,” Convention Paper 8067, 128th AES Convention, 2010.

[17] S. H. Tang, C. M. Liu, and W. C. Lee, “Efficient Design of Time/Frequency Grid in HE-AAC Encoder,” Thesis for Master, 2006.

[18] ITU-R Recommendation BS.1387, “Method for Objective Measurements of Perceived Audio Quality,” Dec. 1998.

[19] PsyTel Multiple Codec Evaluation Software, http://www.psytel-research.co.yu.

[20] ITU Radio communication Sector BS.1116 (rev.1), “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems,” Geneva, 1997.