Analysis by Synthesis Spatial Audio Coding

by
I. Elfitri

Submitted for
the Degree of Doctor of Philosophy
from the University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey

February 2013

© I. Elfitri 2013

ProQuest Number:27558607

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.

ProQuest 27558607

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

Spatial Audio Coding (SAC) is a technique used to encode multichannel audio signals by extracting the spatial parameters and downmixing the audio signals to a mono or stereo audio signal. Recently, various SAC techniques have been proposed to efficiently encode multichannel audio signals. However, all of them operate in open-loop, where the encoder and decoder operate sequentially and independently, and, thus, lack a mechanism for minimising the decoded audio reconstruction error.

This thesis proposes a novel SAC technique that utilises the closed-loop system configuration, termed Analysis by Synthesis (AbS), in order to optimise the downmix signal and the spatial parameters, so as to minimise the decoded signal error. In order to show the effect of the AbS optimisations, the Reverse One-To-Two (R-OTT) module, used in the MPEG Surround (MPS), must first be applied in the frequency domain to recalculate the downmix and residual signals based on the quantised spatial parameters. These parameters show that the AbS scheme can minimise the quantisation errors of the spatial parameters. As the full AbS is far too complicated to be applied in practice, a simplified AbS algorithm for finding sub-optimal parameters, based on the adapted R-OTT module, is also proposed.

Subjective tests show that the proposed Analysis by Synthesis Spatial Audio Coding (AbS-SAC), encoding 5-channel audio signals at a bitrate of 51.2 kb/s per audio channel, achieves higher Subjective Difference Grade (SDG) scores than the tested Advanced Audio Coding (AAC) technique. Furthermore, the objective test also shows that the proposed AbS-SAC method, operating at bitrates of 40 to 96 kb/s per audio channel, significantly outperforms (in terms of Objective Difference Grade (ODG) scores) the tested AAC multichannel technique.

Key words: Spatial audio coding, analysis by synthesis, MPEG Surround, advanced audio coding.

Email: i.elfitri@surrey.ac.uk

Acknowledgements

I would like to sincerely thank my supervisor Professor Ahmet M. Kondo for everything he has done to support, motivate and help me during my PhD study. His assistance, and the opportunities that he gave me, have been crucial components for me in completing my thesis.

I would also like to thank my co-supervisor, Dr. Xiyu Shi, for his support, guidance and sharing experiences. Additionally, I would like to thank Dr. Banu Gunel, who also supervised me during the first and second years of my PhD study.

I would like to sincerely thank the examiners on my PhD examination, Professor Izzet Kale and Dr. Safak Dogan, for taking time to read my thesis and attending the examination as well as for their constructive and valuable comments and suggestions to improve this thesis.

I would like to thank Dr. Janko Calic for his help and discussion about particular topics too. Finally, I owe many thanks to everyone else in the I-Lab Multimedia Communications Research team for their positivity and motivation. And I owe a special thanks to my family and friends for everything.

Contents

List of Figures	vi
List of Tables	vi
1 Introduction	1
1.1 Background	1
1.2 Research Aim and Objectives	3
1.3 Outline of Thesis	4
1.4 Original Contributions	6
2 The Fundamentals of Audio Coding	9
2.1 Basics of Digital Audio	9
2.1.1 Audio Signal	9
2.1.2 Sampling and Quantisation	10
2.1.3 Sound Quality	11
2.1.4 Audio Compression	11
2.1.5 Time-Frequency Transform	12
2.1.6 Perceptual-based Coding	13
2.1.7 Three-Dimensional (3D) Audio	13

2.2	Compression Artifacts	14
2.3	Evaluation of Audio Codec Performance	16
2.4	Applications of Audio Coding Technique	18
2.5	Standardisation Activities for Speech and Audio Coding Technique	20
2.6	The Principle of Analysis by Synthesis	21
2.7	Conclusions	22
3	Spatial Audio Coding	24
3.1	Psychoacoustic Background	24
3.1.1	Human Auditory System	24
3.1.2	Spatial Hearing	26
3.2	Spatial Audio Processing	27
3.2.1	Recording, Transmission, and Reproduction	28
3.2.2	Virtual Source Principle	30
3.3	Multichannel Audio Coding Techniques	31
3.3.1	Matrixed Surround Sound Systems	31
3.3.2	Discrete Multichannel Audio Coding	32
3.3.3	Spatial Audio Coding	34
3.3.4	Spatial Audio Object Coding	38
3.3.5	Other Parametric Techniques	40
3.4	Comparison of Coding Techniques	40
3.5	Conclusions	42

4	Advanced Audio Coding and MPEG Surround	43
4.1	Advanced Audio Coding	43
4.1.1	MDCT Transform	49
4.1.2	Quantisation	50
4.1.3	Noiseless Coding	53
4.2	MPEG Surround	55
4.2.1	Frequency Decomposition	56
4.2.2	Spatial Analysis and Synthesis	60
4.2.2.1	The OTT and R-OTT modules	61
4.2.2.2	The TTT and R-TTT modules	63
4.2.2.3	Parameter Band	65
4.2.3	Quantisation and Coding	66
4.2.4	Rate-distortion Scalability	67
4.3	Conclusions	68
5	A Closed-loop R-OTT Module	69
5.1	Analysis by Synthesis Framework	69
5.2	Implementation of the R-OTT Module within the AbS-SAC Framework	71
5.3	The Closed-loop R-OTT Module in the MPS Architecture	76
5.4	Results	76
5.4.1	Experimental Setup	77
5.4.2	SNR Comparison for Open-loop and Closed-loop Systems	78
5.4.3	Performance Comparison for Various Bitrates	82
5.5	Conclusions	82

6	MDCT-based R-OTT Module	84
6.1	Motivation for the MDCT-Based Implementation	85
6.2	Determining the Parameter Bands	87
6.3	MDCT-Based R-OTT Module for High Bitrate Implementation .	88
6.4	Results	89
6.4.1	Evaluation of the MDCT-based Open-loop R-OTT Module	90
6.4.2	Comparison of the Open and Closed-loop R-OTT Modules over Various Spatial Quantisers	93
6.4.3	Evaluation of the Closed-loop R-OTT Module Using the MPS's Spatial Quantisers	95
6.4.4	Evaluation of the Closed-loop R-OTT Module Encoding 2-channel Audio Signals at High Bitrate	95
6.4.5	Evaluation of the Closed-loop R-OTT Module Encoding 5-channel Audio Signals at High Bitrate	96
6.5	Conclusions	98
7	Simplified AbS Algorithm	99
7.1	Full Search AbS Optimisation	100
7.2	Extended Structure for Multichannel Audio Signals	101
7.3	Simplified AbS Optimisation	105
7.3.1	An Approach for Algorithm Simplification	105
7.3.2	Basic Scheme of the Encoder	107
7.3.3	Algorithm of the AbS Optimisation	108
7.3.4	Complexity of Simplified AbS algorithm	111
7.4	Results	112
7.4.1	Performance for Different Levels of Complexity	112

7.4.2	Performance Comparison at Various Bitrates	115
7.4.3	SNR Comparison with AAC Multichannel	117
7.5	Conclusions	118
8	Audio Codec Evaluation	119
8.1	Experimental Setup	119
8.2	Evaluation of the Closed-loop R-OTT Module	121
8.3	Analysis of Signal Distortion in the QMF Filterbank	122
8.4	Subjective Evaluation	125
8.5	Objective Evaluation	126
8.6	Conclusions	130
9	Conclusions	131
9.1	Research Aim and Objectives	131
9.2	Research Achievements	132
9.2.1	A closed-loop R-OTT module	133
9.2.2	The MDCT-based Implementation of the R-OTT Module .	133
9.2.3	Simplified AbS Optimisation	134
9.2.4	Audio Codec Evaluation	135
9.3	Future Work	136
A	List of Publications	137
B	List of Abbreviations	138
	Bibliography	141

List of Figures

2.1	Block diagram of a typical Analysis by Synthesis technique. . . .	22
3.1	Illustration of microphone beam patterns for 5.1 audio recording .	29
3.2	Illustration of an SAC approach	34
3.3	Block diagram of an SAOC approach	39
4.1	Block diagram of MPEG AAC Encoder	45
4.2	Block diagram of MPEG AAC Decoder	46
4.3	Flowchart of the inner iteration loop	51
4.4	Flowchart of the outer iteration loop	54
4.5	Block diagram of MPEG Surround.	57
4.6	Block diagram of Analysis Quadrature Mirror Filterbank (A-QMF)	59
4.7	Block diagram of (a) OTT module and (b) R-OTT module as used in MPS.	61
4.8	Block diagram of (a) the TTT and (b) the R-TTT modules. . . .	64
5.1	Analysis by synthesis framework for encoding spatial audio. . . .	70
5.2	Block diagram of the closed-loop R-OTT module.	74
5.3	The closed-loop R-OTT modules are employed within a multichan- nel audio encoder. Any type of audio encoder, highlighted as a dashed block, can be employed.	75

5.4	The segSNR per frame	81
5.5	Average SegSNRs of 5-channel audio coders are plotted against the bitrates. At a bitrate of 160 kb/s per audio channel, the improvement achieved by the closed-loop R-OTT module is 5.66 dB . . .	83
6.1	The block diagram of L-channel MDCT-based closed-loop R-OTT.	89
6.2	SegSNR of MDCT-based R-OTT module	91
6.3	SegSNR of MDCT-based closed-loop R-OTT	93
6.4	SegSNR of MDCT-based closed-loop R-OTT module	94
7.1	A tree of R-OTT modules for encoding L-channel audio signals. .	102
7.2	Block diagram of the AbS-SAC encoder.	109
7.3	Performance of the proposed AbS approach using simplified AbS algorithm.	115
7.4	Performance of the AbS-SAC for various bitrates in comparison with AAC and MPS.	117
8.1	A tree scheme of R-OTT modules used in the experiments	121
8.2	The results of the subjective test: (a) the SDG scores of the tested audio codecs averaged over all audio excerpts (b) the SDG scores of all audio excerpts averaged over the tested audio codecs (c) the SDG scores of the tested audio codec for Applause audio excerpt.	127
8.3	Objective difference grade (ODG) of the AbS-SAC for various bitrates in comparison with the tested AAC. The ODG scores of the tested AAC encoding 10-channel audio signals are not plotted as they are similar to the ODG scores of the tested AAC encoding 5-channel audio signals.	129

List of Tables

4.1	The number of coefficients in each scale factor band [1]	50
4.2	Description of AAC Huffman Codebooks	55
4.3	Mapping of 71 hybrid bands to 20 parameter bands	66
5.1	Average SegSNR (dB) of 2 Channel Audio Coders	78
5.2	Average SegSNR (dB) of 5 Channel Audio Coders	80
6.1	Mapping of 49 Scale Factor Bands to 20 Parameter Bands	88
6.2	Average SegSNR (dB) of Various R-OTT Methods	95
6.3	Average segSNRs (in decibels) of MDCT-Based Codecs Compared to FB-Based R-OTT	97
7.1	Complexity of simplified AbS algorithm	112
7.2	Average segSNRs (in decibels) of AbS-SAC	113
8.1	List of Audio Excerpts for Experiments	120
8.2	SegSNRs (in decibels) of the Closed-Loop R-OTT (CL R-OTT) Compared to the Open-Loop R-OTT (OL R-OTT)	122
8.3	SegSNRs (dB) of Various Signal Transformation Schemes Using QMF and MDCT	124

Chapter 1

Introduction

1.1 Background

Having in mind the growing demand for the reliable delivery of high quality multichannel audio in applications such as home entertainment, digital audio broadcasting, computer games, music streaming services as well as teleconferencing, efficient coding techniques [2–4] have become paramount in the audio processing arena. The traditional approach for compressing multichannel audio has been to encode each audio channel using a mono audio coder, such as Dolby AC-3 [5, 6] and MPEG Advanced Audio Coding (AAC) [7–9]. However, for the majority of coders adopting this method, the number of bits to be transmitted tends to increase linearly with the number of channels.

Recently, a new concept for encoding multichannel audio signals has been proposed. It comprises the extraction of the spatial cues and the downmixing of multiple audio channels into a mono or stereo audio signal [10]. The downmix signals are subsequently compressed by an existing audio encoder and then transmitted, accompanied by the spatial cues coded as a spatial parameter. Any re-

ceiver system that cannot handle multichannel audio can simply remove this side information and just render the downmix signals. This provides the coder with backward compatibility, which is important for implementation in various legacy systems. In addition, by utilising the spatial parameters, the downmix signals can be directly upmixed at the decoder side into a multi-channel configuration that may be different from the one used at the encoder side. This technique is known as Spatial Audio Coding (SAC) [10–12].

Various SAC techniques, such as Binaural Cue Coding, (BCC) [13–15] and MP3 Surround [16–18] have been proposed. Inter-Channel Level Difference (ICLD), Inter-Channel Time Difference (ICTD), and Inter-Channel Coherence (ICC) are extracted as spatial parameters that are based on human spatial hearing cues [19]. SAC techniques such as Parametric Stereo (PS) [20–22] and MPEG Surround (MPS) [23–25] may also utilise signal processing techniques, such as decorrelation. The great benefit of these perceptual-based coders is that they can achieve bitrates as low as 3 kb/s for transmitting spatial parameters, as in the case of MPS [24].

There are other techniques based on the extraction of the source location information from multiple channels, thus simplifying the representation of the spatial scene. These techniques calculate the direction vectors representing the directional composition of a scene. At the decoder side, virtual sources are created from the downmix signal at those positions given by the direction vectors. This approach provides flexibility in the decoder output format and the multi-channel configuration. Examples for this technique include Spatial Audio Scene Coding (SASC) [26–28] and Directional Audio Coding (DirAC) [29–32]. DirAC is slightly different in that it incorporates a recording system using a microphone array [33], where the direction vectors are calculated from the microphone signals.

Rather than extracting the spatial parameters, an extension of the SAC method,

called Spatial Audio Object Coding (SAOC) [34–36], analyses object signals instead of channel signals. This offers the end users the flexibility to recompose and reproduce the audio scene. Object Level Difference (OLD) and Inter-Object Cross-coherence (IOC) are extracted as object parameters and transmitted as the side information of a mono or stereo downmix signal. SAOC also provides users with an ability to render either only vocals (to enhance speech intelligibility) or only background music for applications such as karaoke. Its implementation, in combination with DirAC, to help capture audio object has been reported in [35].

As one can observe, each of these coding techniques has its unique advantages and disadvantages. However, they can all be classified as open-loop systems, where the encoders of BCC [15], MP3 Surround [16] and MPS [37], do not consider the decoding process during encoding. The major drawback of an open-loop system is that there is no mechanism employed to reduce the error introduced by quantising the spatial parameters and coding the downmix signals. In this thesis, an Analysis by Synthesis Spatial Audio Coding (AbS-SAC) technique is presented, which provides the advantages of a closed-loop system in order to improve the quality of multichannel audio reproduction at a cost of increasing encoder complexity. It is believed that the AbS-SAC technique can be implemented in any of the recent SAC schemes, even though in this work the AbS-SAC is applied solely in the context of the MPS architecture.

1.2 Research Aim and Objectives

The aim of this work is to develop a new SAC technique for encoding multichannel audio signals that operates in closed-loop to minimise signal distortion due to encoding and quantisation processes. In order to develop a systematic research methodology, this aim can be broken down into the following technical:

-
- To develop a novel SAC approach based on an Analysis by Synthesis (AbS) concept. Various possible ways to perform the spatial audio coding technique in the form of AbS will be investigated, and a system that can be a basis for those various possible implementation will be defined as the AbS-SAC framework.
 - To implement the state-of-the-art spatial audio coding technique, i.e. MPEG Surround (MPS), within the AbS-SAC framework. The Reverse One-To-Two (R-OTT) module that is used in the MPS will be developed in the form of a closed-loop system. Investigation will be focused to find an approach for improving the objective fidelity of the reconstructed audio signals, in terms of segmental Signal-to-Noise-Ratio (segSNR).
 - To investigate the possibility of performing an AbS optimisation to achieve further improvement of segSNR of the reconstructed audio signals. The complexity of the AbS optimisation algorithm will also be investigated and a simplified algorithm will be developed in order to make the proposed technique a practical solution.
 - To perform various methods to evaluate the proposed AbS-SAC approach in comparison to the state-of-the-art techniques. The segSNR measurement is considered to be useful for a comprehensive evaluation, as the AbS approach actually operates in such a way that the output audio signals will be the same as the input signals. In addition, objective and subjective perceptual tests will also be used.

1.3 Outline of Thesis

The rest of this thesis is organised as follows.

Chapter 2: The Fundamentals of audio coding

Chapter 2 introduces briefly the fundamentals of audio coding that include basic principles of digital audio as well as techniques that are closely related to audio coding. Compression artifacts that frequently appear in audio coding as well as the method to evaluate the performance of an audio codec are discussed. Moreover, other related topics are also briefly presented, such as the applications of the audio coding technique and the analysis by synthesis concept that is used as a basis for the proposed AbS-SAC technique.

Chapter 3: Spatial audio coding

Basic knowledge about spatial audio coding is discussed in this chapter. Firstly, a psychoacoustic background of spatial audio coding is illustrated, followed by a discussion about the spatial audio processing, including the recording, transmission and reproduction stages. Subsequently, several multichannel coding techniques are described. Finally, a performance comparison of those coding techniques is presented.

Chapter 4: Advanced audio coding and MPEG surround

In this chapter, two state-of-the-art coding techniques that are currently available for delivering multichannel audio signals are presented. Both techniques that are MPEG AAC and MPEG Surround standards are the basis for the development of the proposed AbS approach.

Chapter 5: A closed-loop R-OTT module

Chapter 5 presents the framework for the analysis by synthesis Spatial Audio Coding (AbS-SAC). The adaptation of the Reverse One-To-Two (R-OTT) module, an elementary block used in MPEG surround, is described. Moreover, the discussion is followed by several experimental results showing the segSNR improvement achieved by the proposed AbS-SAC technique.

Chapter 6: Modified Discrete Cosine Transform (MDCT)-based R-OTT Module

This chapter discusses the MDCT-based closed-loop R-OTT module which is able to further improve the performance of the proposed closed-loop R-OTT module. Additionally, it also simplifies the coding structure. As in the previous chapter, the experimental results are presented to demonstrate that a significant improvement is achieved.

Chapter 7: Simplified AbS algorithm

In chapter 6, the AbS optimisation procedure is discussed. The simplest case of the full search AbS optimisation is presented. However, as complexity is the main challenge of this implementation, a simplified algorithm is proposed in Chapter 7. Experimental results show that the proposed algorithm can increase further the segSNR.

Chapter 8: Audio codec evaluation

In chapter 8, the overall performance of the proposed AbS-SAC technique is evaluated. Several critical test materials are prepared. Objective test methods, such as segSNR measurement, as well as an objective perceptual evaluation using Objective Difference Grade (ODG) metrics, are presented. Furthermore, the results of subjective test are also presented.

Chapter 9: Conclusions

Chapter 9 summarises all the work presented in the thesis. The objectives of the work as well as the proposed solution are restated. In addition, possible future work is presented.

1.4 Original Contributions

- A novel AbS-SAC framework for encoding multichannel audio signals has been proposed. It is believed that a spatial analysis and synthesis technique can be implemented within this AbS-SAC framework. In addition,

any number of input and output channels can be used. Furthermore, any suitable error criteria can be utilised. For taking into consideration the error introduced in the communication channel, an additional block that models the channel error can be included in the AbS-SAC framework.

- An adapted version of the closed-loop R-OTT module has been developed. The downmix and the residual signal are recalculated based on the quantised spatial parameters. Using this method the quantisation error of the spatial parameter can be compensated in the new downmix and residual signals. Implementation of this closed-loop R-OTT module in the MPEG Surround architecture has been investigated. The experimental results show that the closed-loop R-OTT module increases the segSNR. However, it is indicated that the performance can be improved further.
- Based on the previous investigation the MDCT-based R-OTT has been proposed. This is the solution, due to the disadvantage of implementing the closed-loop R-OTT module in the subband domain of the Quadrature Mirror Filterbank (QMF) filterbank which is applied in the MPS. The results of the experiments show that the MDCT-based closed-loop R-OTT module is capable of minimising the error introduced by the quantisation of the spatial parameters.
- An AbS optimisation for finding the reconstructed audio signals with minimum segSNR has been proposed. However, it is shown that the full search AbS optimisation is extremely complex because there are as many as 6.6564×10^{10} combination of parameters that need to be tested in order to select one of them which produces minimum segSNR. To enable low complexity implementation, a simplified AbS optimisation has been proposed. Several possible implementations in terms of algorithm complexity have been investigated. A case of implementation applying 243 iterations, which

is a reasonable level of complexity, has been tested which shows further improvement of the segSNR.

Chapter 2

The Fundamentals of Audio Coding

This chapter provides an overview of fundamental principles in audio coding. It begins with a brief discussion of digital audio that includes audio signal sampling and quantisation, time-frequency transformation and perceptual-based audio coding. It then offers a brief review of compression artifacts, followed by an evaluation of the performance of audio codecs. The applications of and standardisation activities for audio coding are also included in this chapter.

2.1 Basics of Digital Audio

2.1.1 Audio Signal

An audio signal is normally represented in the time domain, describing the magnitude of the signal against the time. It is also useful in some applications, however, to represent the audio signal in the frequency domain. Audible signals

that can be heard by humans such as speech, music, and other similar sounds generally have a frequency range between 20 Hz and 20 KHz. Speech signal have a slightly narrower bandwidth (generally less than 16 KHz) while music signals have a wider frequency component.

2.1.2 Sampling and Quantisation

Audio waveforms are naturally available as an analog signal, but almost every sound processing device is currently working in a digital format. To represent audio signals as digital data for storage and transmission purposes, i.e. Analog to Digital Conversion (ADC), two basic processes are required. These are sampling and quantisation. Firstly, a sampling process is used to convert a continuous-time signal to a discrete-time signal. To sample an audio signal having a 20 KHz frequency bandwidth, for example, a sampling frequency of at least 40 KHz is theoretically required based on Nyquist-Shannon theorem [38, 39], to be able to perfectly reproduce the original signal. If a lower sampling frequency is applied, the reconstructed audio signal may be different from the original (i.e. distortion is introduced). This phenomenon is called aliasing [40, 41]. Normally, however, a sampling frequency of between 44.1 and 48 KHz is applied.

Secondly, a quantisation process converts a continuous-amplitude signal to a discrete-amplitude signal. The magnitude of an audio sample in an analog to digital conversion is normally represented as 16 bits, which means that 2^{16} quantisation levels are available. As with the sampling frequency, above, the larger the number of the quantisation levels the smaller the difference between the continuous and discrete signals. The magnitude difference between the continuous and discrete signals is defined as quantisation error or quantisation noise.

2.1.3 Sound Quality

The bandwidth of the captured audio signal along with the number of bits used to represent each sample of audio signal will determine the quality of the reproduced audio signal. For instance, a speech signal limited in the frequency range between 200 and 3400 Hz and then sampled at 8 KHz [42] is assumed to be acceptable for communication through public telephone networks, even though high frequency parts of the speech signal are discarded. This range was used in the first standardised speech coder. For better rendering, telephony systems are now accommodating a wider speech bandwidth, such as 7 and 15 KHz, which are sampled at 16 and 32 KHz, respectively [43]. A CD quality audio signal, on the other hand, is sampled at 44.1 or 48 KHz in order to capture up to the 20-KHz frequency band of audio signals.

Regarding the number of bits used to represent each sample of audio signal, a Pulse Code Modulation (PCM) [44] represents every sample of speech or audio signal as either 8, 16, or 24 bits. For telecommunication, speech signals sampled at 8 KHz (i.e. 8000 speech samples per second) can be transmitted at 64 kb/s by representing every sample as 8 bits. In contrast, a stereo audio signal sampled at 44.1 KHz, with each sample allocated 16 bits, requires 1.4 Mb/s.

2.1.4 Audio Compression

For the majority of applications it is desirable to have the audio format as compact as possible while trying to maintain the quality of the reproduced audio signal. In order to achieve this compression, an audio encoder is used to remove redundancy in the audio data. At the end of the process an audio decoder is used to recreate the original audio signal. In most cases, however, the audio encoder and decoder introduce some noise into the signal.

PCM is one of the earliest and most straightforward coding approaches that can be applied to speech and audio signals. Later developments of PCM are known; they include Delta Modulation (DM), and Adaptive Differential PCM (ADPCM). In general, all of these approaches exploit the redundancy of the time domain waveform of speech and audio signal. However, modern audio coders employ advanced techniques that rely upon time-frequency decomposition, as well as taking advantage of the psychoacoustic model [45]. The most popular coding technique is MPEG-1 layer 3, commonly known as the MP3. The majority of audio applications available at the moment can support this MP3 audio format. However, there are other coding techniques that are developed for specific purposes. For instance, Code Excited Linear Prediction (CELP) [46] was specifically developed for encoding speech signals.

2.1.5 Time-Frequency Transform

In order to represent the signal in as few bits as possible, various different types of audio coding techniques have been proposed. Recent techniques basically process the time domain audio signal in a frame or a block of audio samples. The frame size, when 48 KHz of sampling frequency is applied, may normally range from 400 samples (8.3 ms) to 4096 samples (85.3 ms) [47–49]. However, a longer or shorter frame size can be applied. Using a longer audio frame potentially gives a higher coding efficiency than a shorter one, however, it can longer the processing delay. Additionally, increasing the audio frame size possibly makes the encoder fail to properly model the characteristic of the audio signal. On the other hand, for applications that require a lower delay, a shorter audio frame, such as 5.8 ms that is used in Constrained-Energy Lapped Transform (CELT) [47], will be useful at a cost of reducing coding efficiency. The audio frame is then either transformed into the frequency domain (a process termed transform coding), or is decomposed

as sub-band signals (subband coding). Transformation or decomposition tools, such as Discrete Fourier Transform (DFT), filterbank, and Modified Discrete Cosine Transform (MDCT), are currently widely used in cutting-edge audio coding techniques. Representing audio signals in the frequency domain instead of in the time domain is generally assumed to be more effective in reducing signal distortion. In addition, processing the audio signal in the frequency domain also allows for the possibility of utilising the psychoacoustic model, which is very important for reducing signal distortion.

2.1.6 Perceptual-based Coding

Knowledge and understanding about human perception of sound signals is also exploited in most modern coding techniques. This kind of approach is called a perceptual-based audio codec. Basically, a psychoacoustic model estimating parameters, such as the hearing and masking threshold, is utilised. The main benefit of the audio codec employing this model is the ability to effectively allocate bits, in the quantisation process, into any part of audio signal that is audible to the human hearing system. Used this way, any imperceptible part of the audio signal will be removed.

2.1.7 Three-Dimensional (3D) Audio

The release of the MPEG-1 standard in 1993 motivated the widespread implementation of digital audio materials. The most popular MPEG-1 layer 3 [50–52], widely known as the MP3, facilitated representation of the audio signal in a more compact way than the conventional audio CD format. Since then, most digital audio materials were available in the MP3 format for both on-line audio services

such as Digital Audio Broadcasting (DAB) and off-line audio services that can be rendered in various portable media players.

In addition to the benefit of providing a compact audio representation, MPEG-1 layer 3 can also support stereo audio format. It is easy now to find audio materials, such as music albums, in the format of the MP3 stereo. Even though the stereo audio format increases the audio data size, because it consists of two audio channels, it is interesting to see that a stereo audio format frequently becomes the first choice for digital audio representation, instead of a mono audio format. This is due to the fact that the stereo audio format will increase the realism of a real acoustic environment. Hence, it is favorable to represent the audio materials in stereo rather than mono formats even though a bigger data size is needed.

Representation of audio materials as stereo formats, or even a larger number of audio channels, has been possible and has been used for many years in multiple applications (particularly in the cinema industry). It is well understood that multichannel audio representation has ability to convey much more information about 3-Dimensional (3-D) audio scenes including the individual sounding objects, their positions as well as the acoustic of the environment. Started by the cinema industry, it is now used in many off-line and on-line applications including home entertainment and the Internet applications. Moreover, the importance of 3-D audio is projected to increase even further, due to the introduction of 3D television services and the spread of 3-D displays.

2.2 Compression Artifacts

An efficient audio coding technique basically employs two approaches. The first approach is to simply capture as narrow a bandwidth as possible. For instance, in an encoder employing Spectral Band Replication (SBR), the high frequency

regions of the audio signal are modeled and parameters are extracted. At the decoder side, the parameters are applied to create the replica of the high frequency parts of the audio signal. The second approach is to represent every audio sample with as few bits as possible. This may be achieved by employing an efficient quantisation technique. Both approaches are performed in such a way as to maintain the reconstructed audio quality. However, both approaches may introduce audio distortion or artifact.

In order to reconstruct an indistinguishable audio signal from compact audio data, attention should be paid not only to the efficient technique of audio coding but also to the artifacts that usually appear in compressing the audio signal. Several major audio artifacts have been identified, such as pre-echo, speech reverberation, aliasing, and birdies [53]. In addition, other artifacts may also be present, due to the process of audio recording as well as format conversion or tandem coding.

Compression artifacts mainly occur due to the processing of audio signal in a frame basis, particularly in encoding a segment of audio signal that contains an abrupt change of magnitude during a short time duration. This event is known as a transient or attack signal. A transient event in a time domain signal will transform to a widespread spectrum in the frequency domain. As a result the psychoacoustic model may fail to calculate the correct allowed distortion over all subbands. It means that quantisation errors may be introduced in the whole frequency spectrum of the frame. At the decoder side, these errors will transform to distortion over all samples in the audio block including in the preceding of the transient event. This distortion that may be audible is known as pre-echo.

Pre-echo occurs during the application of a long length of audio frame, and it can be avoided by shortening the size of the frame. Therefore, advanced audio coders employ an adaptive length of audio frame. This can be realised by predicting the energy of the upcoming frame. If a transient is detected then the frame size

will be switched to a short size. A case similar to pre-echo in encoding speech signal is called speech reverberation. Some parts of speech signal can be classified as a non-stationary signal that has the same properties as a transient signal. In encoding this kind of speech signal, distortion can be introduced in the previous and after the original transient event. However, speech reverberation is usually audible after the sound of the original transient signal. The noise that following the transient event will decrease in the time.

Aliasing artifact is mainly caused by the implementation of critically-sampled filterbank. This type of filter has a set of overlapped bandpass filters. The outputs of the filterbank are down-sampled by a particular factor which is usually chosen to be equal to the number of subband signal. Even though it is designed to obtain a perfect audio reconstruction by eliminating the aliasing, distortion is still introduced due to the imperfect frequency response of the filter. Furthermore, quantisation of subband signals that is performed in between the analysis and synthesis filterbanks also introduce further distortion.

Birdies artifact is discontinuities that exist in the frequency domain. Due to the implementation of psychoacoustic-based bit allocation, specific spectral coefficients may be temporarily allocated sufficient bits, but may not be sufficient in some audio blocks. Consequently, these spectral coefficients would appear in some frames but disappear in the other frames.

2.3 Evaluation of Audio Codec Performance

The quality of the reconstructed audio signal is the most important factor to be considered in designing and developing an audio codec. Other factors include delay and complexity. For this reason, an accurate audio evaluation method is necessary to assess the performance of the audio codec. The most reliable and

relevant method to perform this quality assessment is to undertake a listening subjective test, such as those recommended in the ITU-R BS.1116-1 [54] and the ITU-R BS.1534 recommendations [55]. The ITU-R BS.1116-1 Recommendation is often used in the assessment of audio signal, including multiple audio signals, that introduce small impairments which cannot be detected easily. In order to detect these small impairments it is frequently necessary to repeatedly listen to the audio signal. Thus, it is specified in [54] that to score an audio excerpt several trials of listening may be required. On the other hand, the ITU-R BS.1534 entitled method for the subjective assessment of intermediate quality level of coding systems, is frequently applied to evaluate the quality of the audio signal that contains a larger level of distortion which is easily detectable even in a single listening.

Both of those recommendations specify that a minimum number of five critical audio excerpts, typically having duration of 10 to 25 s, selected from a large number of audio materials, should be used in the assessment. The quality of the audio codec under test in encoding these audio excerpts is judged by a number of expert listeners who have the ability to detect the audio impairments. The expertise of the listeners becomes more important when the impairments are small. It is specified in [54] that an appropriate conclusion of the subjective test is achieved when at least 20 expert listeners participate in the listening test.

As the subjective assessment methods are very expensive and time-consuming, in order to provide a quick estimation of the reproduced audio quality, performing an objective test is preferable. The conventional methods used to perform this objective test include measurements of Signal-to-Noise-Ratio (SNR) and Harmonic-Distortion (HD), where the time and frequency domain of the original and reconstructed audio waveforms are compared. However, these objective metrics are no longer suitable, especially for assessing the quality of the audio signals reconstructed from modern perceptual based coding techniques. The ITU-R

BS.1387-1 [56], recommending a method for objective measurement of perceived audio quality, is assumed to have ability giving a more accurate audio quality assessment, instead of those traditional SNR and HD metrics. Using this objective test, an Objective Difference Grade (ODG) is obtained to represent the quality of the tested audio signals compared to the original signals. Five grades are available in this ODG metric, similar to the Subjective Difference Grade (SDG). They are:

- 0 (imperceptible)
- -1 (perceptible but not annoying)
- -2 (slightly annoying)
- -3 (annoying)
- -4 (very annoying)

2.4 Applications of Audio Coding Technique

MPEG-1 layer 3 (MP3) is a very popular method for compressing audio signals. Its high coding efficiency makes it possible to represent a 3-minute music clip as a piece of data sized around 4 Mb. Since it was standardised, MP3 compressed audio signals have become preferable for many applications, such as those used in home and portable entertainment devices, digital audio broadcasting, music downloading and streaming services, as well as other Internet applications (such as teleconferencing).

Teleconference is a popular way to communicate remotely. Using it, several people can participate in a talk, discussion, or meeting. However, this technology is

still far from being categorised as a comfortable communication system. Many users often find that it is difficult to identify and understand the voice and speech of remote talkers, because all sounds come from a single loudspeaker. This specifically occurs when multiple talkers are speaking simultaneously.

In order to overcome this problem, a multichannel audio system is required where each voice is reproduced by a single loudspeaker at a particular source position. Using this system, user perception increases significantly, particularly in identifying voices from different talkers [57]. However, this additional perceptual realism does not affect the ability to convey the message delivered [58]. For teleconferencing, DirAC [59] and S³AC [60] have been implemented before as spatial audio coding techniques. However, their performance was limited by the performance of the applied audio coding.

Another promising application of a multichannel audio coding is the enablement of personalised music remixes. Users often want to modify rendered music composition. This feature has actually been available for long a time as a limited function, in that the modification is applied into the whole music composition without the capability to adjust any audio object (i.e. singer vocal or musical instruments). With the recent development of Spatial Audio Object Coding (SAOC) technology, it is now possible to provide users with an interaction feature in the rendering stage to independently manipulate every audio object. For instance, with a Karaoke application, users have an option to play either vocals only (A Capella) or to play background music only.

2.5 Standardisation Activities for Speech and Audio Coding Technique

Several standardisation bodies, such as the International Telecommunication Union (ITU), Moving Picture Expert Group (MPEG), European Telecommunication Standards Institute (ETSI), have regular activities in developing new standards of speech and audio coding. Current and future activities show that there is an interesting trend emerging in speech and audio coding standardisation. The MPEG audio group has released various audio codec standards. The first is MPEG-1 audio standard, released in 1993, consists of layers 1, 2 and 3, where the MPEG-1 layer 3 is widely known as the MP3. This codec has the ability to encode mono and stereo audio signals. Moreover, the MPEG-2 audio part has two types of audio codecs that are backward and non-backward compatible. The MPEG-2 non-backward compatible version was released in 1997, and was later called MPEG Advanced Audio Coding (AAC). It is able to encode multiple audio signals in up to 48 channels. Specifically designed for encoding multichannel audio signals, MPEG Surround was subsequently released in 2008, while MPEG Spatial Audio Object Coding (SAOC) was released in 2010. All those audio codec standards typically sample audio signals at 48 KHz while other sampling frequencies are also accommodated. In order to accommodate a speech signal, as it has a specific signal characteristic, Code-Excited Linear Prediction (CELP) is also currently adopted as a tool in the MPEG standard, ISO/IEC 23003-3/FDIS, Unified Speech and Audio Coding (USAC) [61, 62].

On the other hand, ITU has a long history of standardisation activities in speech coding. Several speech codec standards, generally capturing speech signals at 8 KHz, have been released in past years, where one of them was widely implemented, such as CELP. Recently, ITU has also been working to standardise a

new speech coder that can transmit a wider bandwidth speech signal and, thus, make it possible to also encode an audio signal. In order to make it suitable to different networks, backward compatibility, as well as bandwidth and bitrate scalability, are also targeted to be the features of this codec. Moreover, a mono to multichannel upmixing capability is required. It seems that both standardisation bodies represent similar trends in speech and audio coding activities.

New speech coding standards, such as G.729.1 [63, 64], G.711.1 [65] and G.718 [66, 67], have been developed for delivering wideband speech signals sampled at 16 KHz. In addition, the terms superwideband and fullband speech coders [68], sampling speech signals at 32 KHz and 48 KHz respectively, have been introduced. For instance, ITU-T is currently working to extend G.729.1 as a standard for super wideband speech and audio [69]. Furthermore, ITU-T has also released a newly standardised super wideband codec extension for G.718 [70]. Similarly, another extension was jointly developed for both G.711.1 and G.722 [71–73]. These coders are designed to be backward compatible to the previous coders. They also have bandwidth and bitrate scalability. Moreover, a new G.719 has also been standardised as the first ITU-T’s fullband audio coder [74, 75]. For the next standard, ITU-T also has a study group working to develop a new standard that includes a multichannel upmixing capability [43, 76]. These activities seem to accommodate a transmission of multichannel audio signals in-line with MPEG audio activities.

2.6 The Principle of Analysis by Synthesis

Analysis by Synthesis (AbS) technique [77, 78] is a generic method that has already been implemented in many areas, such as estimation and identification [79, 80]. Several decades ago, this concept was proposed as a framework

Figure 2.1: Block diagram of a typical Analysis by Synthesis technique.

for encoding speech signals and determining the excitation signal in a Linear Predictive Coding (LPC)-based speech coder [81, 82]. Since then, many other speech coders have been proposed within this framework, such as the most popular Code-Excited Linear Prediction (CELP) [46], which is currently specified as one of the tools in the MPEG-4 Audio standard [83]. CELP is also currently adopted in development of the MPEG standard, ISO/IEC 23003-3/FDIS, Unified Speech and Audio Coding (USAC) [61, 62]. The AbS technique is currently applied in many applications, including the quantisation of spectral coefficients of the MPEG-AAC audio codec.

An AbS system, depicted in Fig. 2.1, is able to synthesise a signal by a set of parameters where the values of these parameters are usually made variable in order to produce the best matched synthesised signal. The difference between the observed signal and the synthesised signal, called the error signal, is utilised in an error minimisation block. A set of parameters which produce minimum error signal are selected as optimal parameters and sent to the decoder.

2.7 Conclusions

This chapter has given a brief overview of fundamental principles in audio coding that are necessary in understanding the development of any spatial audio cod-

ing technique. A basic concept of audio compression, as well as artifacts, often introduced in recent audio codecs, is presented, followed by a brief discussion about methods used to evaluate the performance of an audio codec. In addition, promising applications of audio coding techniques, as well as related standardisation activities, are also briefly illustrated in this chapter.

Chapter 3

Spatial Audio Coding

This chapter discusses spatial audio coding. A brief overview of psychoacoustics is presented at the beginning of the chapter. The next section presents spatial audio processing methods, including recording, transmission and reproduction, followed by a description of techniques used for spatial audio compression. At the end of this chapter, a comparison of different coding methods is presented.

3.1 Psychoacoustic Background

3.1.1 Human Auditory System

A deeper understanding of the human auditory system has motivated the development of more efficient and powerful quantisation methods employed in an audio coder. The human ear can listen to a very large range of sound levels. The minimum level that can be heard for a particular frequency is defined as the hearing threshold or threshold in quiet [1, 19]. The human hearing system cannot detect any sound level below this threshold, even in a very quiet condition. This

threshold may be different for each frequency. The hearing threshold phenomenon is then taken into consideration when the quantisation process is performed. It suggests that it is not necessary to transmit any frequency component of an audio signal with a sound level below the hearing threshold. In addition, as long as the quantisation noise is below the hearing threshold, the noise will not be heard.

In addition to the hearing threshold, the masking phenomenon [1, 19] should also be considered in developing an effective quantisation technique in an audio coder. When two sounds, with similar frequencies, occur at the same time, the louder sound can cause the other sound to be undetectable, even though its level is higher than that of the threshold in quiet. This phenomenon is termed frequency masking, where the louder sound is named the masking or masker sound, and the softer sound is called the maskee. In the presence of a high-level masker sound, it is useful to determine a specific sound level, because below this level other sounds will be masked by the masker. This level is defined as the masking threshold. The values of the masking threshold are generally high in the frequencies around the masker frequency. However, they tend to decrease as the frequency gets farther away from the masker.

Another type of masking phenomenon is temporal masking. It occurs when two sounds are not presented simultaneously. In this situation the louder sound is still able to mask the other sounds present before and after the masker. Temporal masking is normally effective only for a few milliseconds before and after the masker sound. Exploiting the masking phenomena benefits a quantisation process where it is not relevant to transmit the maskee sound, as it would not be heard by humans.

Further exploration in the frequency masking phenomena led to the understanding of the principle of critical bandwidth [1, 19], which relates to the frequency range where the masking threshold around the masker frequency is flat. In order

to develop an effective audio coder it is useful to perform the quantisation process in the frequency range or frequency subband that is similar to the critical bandwidth, so that the audio signal component below the masking threshold will be discarded.

Based on knowledge of the human hearing system, a psychoacoustic model can be designed. It actually tries to model all relevant factors that influence the hearing process, including the hearing and masking threshold. The goal of the psychoacoustic model is to calculate the Signal to Mask Ratio (SMR), which is a comparison between the level of the audio signals and the masking threshold. Then, the calculated SMR is used to determine how many bits should be allocated to a certain frequency range in order to ensure the quantisation error is imperceptible.

3.1.2 Spatial Hearing

Having two ears, humans can localise where a sound is coming from. Sound waves coming from a sound source may reach the two human ears, the left and the right, at different times, because they travel different distances. Due to these sound propagation properties the human brain detects the time difference between the sound waves from the left and the right ears. This cue is named an Inter-aural Time Difference (ITD) [19, 84]. This may also cause the sound pressure detected by each ear to have a slightly different level. However, shadowing by the human head, reflection and diffraction by surrounding environments can make the differences smaller or larger, creating another cue, Inter-aural Level Difference (ILD) [19, 84]. The human hearing system simply perceives the position of a sound source by processing these inter-aural time and level difference cues. In the case of the sound wave experiencing various reflections, the resulting sound waves arrive at human ears with different amplitudes and different time delays. Consequently,

the sound waves at both ears will be partially incoherent, creating a perception of widening the sound source.

However, there are some positions, such as in front or in the back of the human head [19, 85], which would produce identical ITDs and ILDs. If a sound comes from these places, the human brain cannot precisely determine where the source position is. This phenomenon is known as cone of confusion [86] and can be solved by head movements [87]. Another situation that makes the human brain fail to locate the proper position of the sound source occurs when there are two or more similar sound sources, located in different locations, transmitting sound waves with small time delays. These sounds will be perceived as a single sound, instead of two or more sounds, coming from a position closer to the earlier sound source. Such a situation is called the precedence effect [88].

The human head's size, the shape of pinna, and the reflections off the shoulder and body also affect the human ability to localise a sound source. All of these cues can be represented as a Head-Related Transfer Function (HRTF) [89]. This function is actually a frequency response of the human auditory system. It is measured for a sound signal coming from every possible source direction. All of these cues, ITD, ILD, inter-aural correlation, along with other cues, can be taken into account in developing an effective spatial audio coding technique.

3.2 Spatial Audio Processing

For a long time period, sound waves were captured, for the purpose of storage and transmission, by a single microphone that discarded all those relevant cues that could describe the individual sounding objects, along with their positions, as well as the acoustics of the environment. Consequently, the reconstructed sound reproduced by a loudspeaker lost its spatial impression. In order to maintain its

spatial effect, sound is usually recorded, transmitted and reproduced in several channels (two, five, or more, for instance). The two-channel audio system, widely known as stereo audio, has been a popular option for many applications due to its simplicity. Two channels, the left and right channels, are located in front of the listeners as most listeners focus on the sound in front of them. Significance improvement of spatial impression is provided compared to a single monophonic audio. However, stereo audio system cannot provide surround sound as there is no channel that intended to represent sounds located on the side and at the back of the listeners. Thus, acoustics of the environment particularly at those positions of the listener will not be presented with a stereo audio system.

The most frequently used number of audio channels for surround sound is five. This configuration was recommended by ITU and has been known as the 5.1 audio system [90]. Three channels are located in front of the listener (the center, left and right channels), and two channels are positioned in the back (the rear/side channels). The .1 is for Low Frequency Enhancement (LFE) channel, where the approximately 120-Hz audio bandwidth required for this channel is much lower than the others. This LFE channel is actually intended to give a special impression on a specific low frequency audio content that needs larger sound pressure particularly for audiovisual applications. Other channel configurations are also available, such as 7.1, 10.2, and 22.2 [91,92] audio systems that are useful for rendering more accurate three dimensional surround sound [86]. Principally, all of these systems are intended to give more spatial effects to the listener.

3.2.1 Recording, Transmission, and Reproduction

Each channel of a multichannel audio content is usually recorded separately by a single microphone that is usually intended to capture sound waves coming from a specific area. For example, in Fig. 3.1, a microphone beam pattern for 5.1 audio

Figure 3.1: Illustration of microphone beam patterns for 5.1 audio recording

recording is presented. Otherwise, all channels can be recorded by a microphone array that provides a set of microphone signals such as B-format [93, 94]. Some commercial microphone products such as Holophone [95] and SoundField [96] are available to capture high quality spatial audio and record B-format microphone signals.

For digital transmission and storage, audio signals are generally represented in fewer channels, often two. Conventional techniques, such as Dolby stereo, Surround and Prologic [86], are available by matrixing multichannel signals. Dematrixing is required to reproduce multichannel signals. Otherwise, each channel at multichannel audio content can be coded and transmitted by an audio coder such as one implemented in MPEG-2 AAC. However, this method is clearly not effective in terms of the amount of bits needed to be transmitted. For this reason, Spatial Audio Coding (SAC), where the total bitrate needed for transmission may be less than the discrete multichannel audio coding, is becoming popular.

Three dimensional audio can be reproduced by playing audio signals with a num-

ber of loudspeakers. Generally, audio signals recorded by a specific channel configuration should be reproduced by the same configuration. For mobile applications, headphones can be an interesting option to reproduce multichannel audio signals by trans-coding multichannel audio into binaural audio.

3.2.2 Virtual Source Principle

Due to limitations in the number of loudspeakers, a virtual sound source can be created in the area between two loudspeakers or the area among three loudspeakers. The listener will perceive the sound coming from the point where the virtual source is located. The most popular way to create such a virtual source is by using Vector Based Amplitude Panning (VBAP) [97–101]. The position of the virtual source is represented by a unit vector. The gain of the audio signal fed to each loudspeaker is then determined by the component of the unit vector. In addition, other panning techniques, such as Ambisonic, can be used. For reproducing a virtual source by a headphone, the HRTFs can be utilised and implemented as a filter. Such an application is known as binaural rendering. With this binaural rendering, a listener can perceive that a virtual source is kept in a constant position even though the listener moves his head. The HRTF has been implemented for binaural rendering in MPEG Surround standard [102].

Amplitude panning techniques are not able to create a virtual source outside the area between the loudspeakers. This problem can be solved by Wave Field Synthesis (WFS) [103]. The WFS technique usually requires a large number of loudspeakers. Each loudspeaker is configured to reproduce a different phase of audio signals. Hence, a specific wavefront can be constructed for a sound source at a specific location, even for a position behind the loudspeakers. However, due to the need for a large number of loudspeakers, this WFS technique is not favorable for home consumers' implementation.

3.3 Multichannel Audio Coding Techniques

For the purpose of storage and transmission, it is always desirable to make the multichannel audio signals as compact as possible. Reducing either the number of channels or the number of bits required for representing the audio signals is the basic option. However, it is necessary to maintain the spatial impression that would be created by those compact-represented audio signals. For this reason, efficient coding techniques play an essential role in delivering high quality multichannel audio. In this section, various approaches proposed to encode multichannel audio signals are discussed.

3.3.1 Matrixed Surround Sound Systems

A conventional approach to store or transmit multiple audio signals is to represent the audio signals in fewer channels, such as two. This basic idea is exploited in a matrixed surround technique, where multiple channels are downmixed into mono or stereo audio by a matrix conversion process. Dolby Stereo, intended for cinema sound application matrixing of three or up to six channels into stereo, is among the one of the approaches available using a matrix surround method. Later products from Dolby, such as Dolby Surround, Prologic and Prologic II, and Lexicon Logic 7 [86] are also based on this approach. In order to properly reconstruct the original channels, the dematrixing process should be carefully performed at the decoder side. However, side-effects such as crosstalk, where an audio signal from a particular channel can be present in the other channels, always exist. Among these techniques, Dolby Prologic II is the most popular, as it has a feature optimised for home theater applications.

3.3.2 Discrete Multichannel Audio Coding

The first approach used for digitally compressing multichannel audio involves encoding each audio channel with a mono audio coder such as Dolby AC-3 [5, 6], Digital Theater System (DTS) [104, 105], MPEG Advanced Audio Coding (AAC) [7–9], and High Efficiency AAC (HE-AAC) [8, 9]. It is actually an approach to encode a mono or stereo audio. However, in the case of multichannel signals, the audio signal in each channel is encoded individually and all channels are then transmitted as a single bitstream. For each channel, a psychoacoustic model is used in this approach to provide information about the minimum allowed distortion. Thus, the reconstructed audio signal is supposed to be perceptually lossless, provided that all required bits are sufficiently available. However, for most coders using this method, the number of bits to be transmitted tends to increase linearly with the number of channels.

The audio signal in each channel of AC-3, also known as Dolby Digital, can be sampled either at 32, 44.1, or 48 KHz while the LFE channel is sampled at 240 Hz. The codec can operate at various bitrates, starting at 32 kb/s for a mono audio signal up to 640 kb/s for 5.1 audio signals. However, its typical operating bitrates for encoding 5.1 audio signals are around 320–384 kb/s. It can also carry downmix control information in order to render a stereo audio in case the decoders do not support 5.1 audio rendering.

For high audio quality such as audio for film, DTS accommodates both higher sampling frequencies up to 192 KHz and larger numbers of channels up to 8 channels. Consequently, the DTS codec operates at higher bitrates than AC-3, which is up to 512 kb/s per audio channel. However, a lower audio quality is still possible when encoding the audio signal at bitrates as low as 8 kb/s per audio channel. In addition, DTS also provides an up-matrixing feature, in that new channels are created based on the existing available channels.

On the other hand, MPEG AAC offers three profiles: main, Scalable Sampling Rate (SSR), and Low Complexity (LC). The main profile is the most complex profile, where all tools provided such as Mid/Side (M/S) stereo coding and Temporal Noise Shaping (TNS) are used to achieve the highest audio quality. In contrast, LC-AAC has the lowest complexity. However, it is still able to provide a relatively high quality of audio signals. Up to 48 audio channels are supported. The AAC was initially developed to achieve indistinguishable audio quality for delivering high quality 5.1 audio at a bitrate of 384 kb/s or lower. Assessment using the subjective test reported that this quality was achieved at bitrates of 320 kb/s (BBC, UK) and 256 kb/s (NHK, Japan). The audio signal can be sampled at a frequency as low as 8 KHz, so that low operating bitrates, such as 8 kb/s for a mono audio signal, are supported. Nevertheless, at these low operating bitrates the quality of the reconstructed audio considerably degrades.

High efficiency AAC is designed for operating at low bitrates (i.e. around 20 kb/s per audio channel) to achieve better performance than AAC. It is actually based on LC-AAC for encoding low frequency parts (i.e. half of the bandwidth of the audio signal). A new technology, called Spectral Band Replication (SBR), is introduced to extract control parameters from the high frequency parts of the original audio signals. This control information, mainly consisting of spectral envelope representation, is encoded at a very low bitrate and is then transmitted, along with the LC-AAC encoded bitstream.

At the decoder side, the control parameter can be utilised to create the missing high frequency parts of the reconstructed audio signals. Both the control parameter extraction and the high frequency signal reconstruction are performed in the subband domain of a filterbank called a hybrid Quadrature Mirror Filterbank (QMF). A bitrate of approximately 3 kb/s is assumed to be sufficient to send the control parameters so that the total bitrate of the HE-AAC is lower than the bitrate required when the LC-AAC is used. However, the HE-AAC codec is un-

Figure 3.2: Illustration of an SAC approach

able to achieve transparent quality of the audio signals even though the operating bitrate increases.

3.3.3 Spatial Audio Coding

A recent concept proposed for encoding multichannel audio signals consists of extracting the spatial parameters and downmixing multiple audio channels into a mono or stereo audio signal, as seen in Fig. 3.2. The downmix signals are subsequently compressed by an existing audio encoder and then transmitted, accompanied by spatial cues as side information. Any receiver system which cannot handle multichannel audio can simply remove this side information and just render the downmix signals. This provides the coder with backward compatibility, which is important for implementing various legacy systems. In addition,

by utilising spatial parameters, the downmix signals can be directly upmixed at the decoder side into channel configurations that can be different from the one used at the encoder side. This technique is known as Spatial Audio Coding (SAC) [10–12].

Various SAC techniques have been proposed, such as Binaural Cue Coding (BCC) [13–15] and MP3 Surround [16–18]. Inter-Channel Level Difference (ICLD), Inter-Channel Time Difference (ICTD), and Inter-Channel Coherence (ICC) are extracted as spatial parameters that are basically determined based on the exploitation of human spatial hearing cues [19]. Two types of BCC codecs were proposed for addressing two application scenarios. The first is called BCC for Flexible Rendering and is intended for the transmission of multiple source signals or audio objects. The second is named BCC for Natural Rendering and is needed for the transmission of multiple audio channels. A Discrete Fourier Transform (DFT) is used for the time to frequency domain transformation. The inter-channel parameters are extracted and then transmitted as side information of a mono sum or downmix signal. Any mono speech or audio encoder can be used for further encoding the mono sum signal.

MP3 Surround is developed based on BCC. The basic idea is to benefit from the popular and widespread use of the stereo MP3 player. Instead of a mono sum signal, MP3 Surround creates a stereo sum signal from multiple input signals and uses stereo MP3 encoders for further compression. The reason for transmitting a stereo sum signal arises based on the fact that most audio materials are available in the format of MP3 stereo. With this approach, the stereo MP3 audio codec can be extended for multichannel applications. This approach also takes advantage of the backward compatibility principle. Any MP3 stereo decoder that does not support multichannel applications is still able to play audio materials encoded by MP3 Surround by simply discarding the side information. In addition, a feature to transmit the available artistic stereo signal as the sum signal is offered instead

of creating a new stereo signal, as applied in the BCC encoder. This feature ensures that MP3 decoders that do not support multichannel rendering can still render the best stereo audio composition.

Techniques such as Parametric Stereo (PS) [20–22] and MPEG Surround (MPS) [23–25] may also utilise better time-frequency transformation and signal processing techniques, such as decorrelation. Additionally, near transparent audio reconstruction is possible in MPS by transmitting the residual signal. The great benefit of these perceptual-based coders is that they can achieve bitrates as low as 3 kb/s for transmitting spatial parameters, as in the case of MPS [24].

Parametric stereo is actually a SAC approach for stereo signals only and is not used for multichannel signals. Each of the left and right channels is transformed into subband signals by an effective filterbank called a hybrid Quadrature Mirror Filterbank (QMF). Inter-channel Intensity Difference (IID) and Inter-channel Phase Difference (IPD) are basically similar to ICLD and ICTD, as well as ICC, and are extracted as spatial parameters. A downmix signal is created as a linear combination of both input signals, ensuring that the power of both input signals is equal to the power of the downmix signal. The IID and IPD quantisers are designed based on perceptual criteria, where the introduced quantisation errors are assumed to be inaudible. The PS technique, along with Spectral Band Replication (SBR), support Low Complexity AAC (LC-AAC) for MPEG audio standard known as High Efficiency AAC v2 (HE AAC v2).

The advantages of PS and MP3 Surround are taken into account in the development of MPEG Surround. Based on the PS's spatial analysis and synthesis, an elementary module, named One-To-Two (OTT), is created and utilised in MPS to convert a channel into two channels. For reverse conversion, a Reverse OTT (R-OTT) is also introduced. On the other hand, the Two-To-Three (TTT) module is designed to convert stereo channels to three channels, while the Reverse-TTT

(R-TTT) is intended to do the reverse conversion. Multichannel down/upmixing is performed by a number of OTT modules in a tree scheme. For perfect waveform reconstruction capable of performing near transparent audio reconstruction, residual coding is applied. MPS also offers other features, such as binaural rendering and enhanced matrix mode. Moreover, it includes the artistic downmix feature [106], similar to the one applied in the MP3 Surround.

Other techniques also exist that are based on extracting the source location information from multiple channels, thereby simplifying the representation of the spatial scene. These techniques calculate direction vectors representing the directional composition of a scene. At the decoder side, virtual sources are created from the downmix signal at positions given by the direction vectors. This approach gives the coder the capability to reproduce different number of output channels than input channels. Examples of this technique are Spatial Audio Scene Coding (SASC) [26–28] and Directional Audio Coding (DirAC) [29,30,32].

DirAC is slightly different, in that it incorporates a recording system using a microphone array [33]. Therefore, the direction vectors in this coder are calculated from microphone signals, particularly B-format signals. In addition to direction vector, diffuseness representing the acoustics of the environment is also extracted and transmitted as side information, along with the direction of sound arrival. At the decoder side, the reconstructed audio signals can be rendered by any number of loudspeakers by applying the concept of Vector Based Amplitude Panning (VBAP). The DirAC's spatial parameters can also be translated into CLD and ICC. Moreover, this approach is primarily intended for teleconference application.

Another method is approached by squeezing or mapping the auditory space from 360° into 60° which is practically considered to be good enough for human hearing fidelity. Such a mapping relies on estimating virtual sources, performed by amplitude panning technique, between pairs of input channels. In order to pro-

vide a stereo downmix signal, subsequent panning is applied into estimated virtual sources. This work, known as Spatially Squeeze Surround Audio Coding (S³AC) [107–111], is unique in that it basically does not need to transmit additional side information. The major drawback of S³AC is that it introduces the ambiguity of sound position, caused by representing audio channels as a set of virtual sources.

3.3.4 Spatial Audio Object Coding

An extension of the SAC method is also available today, where the object parameters are extracted, rather than the spatial parameters. It is called Spatial Audio Object Coding (SAOC) [34–36] as it analyses object signals instead of channel signals, offering end users a flexibility to recompose and reproduce the audio scene. The block diagram of an SAOC approach that looks similar to an SAC approach is given in Fig. 3.3. The major difference with SAC is that an audio object analysis block is used to extract object parameters, while in the decoder side an audio object synthesis block is applied for reconstructing back the audio objects. For both rendering with any loudspeaker configuration and remixing the audio scene composition, a mixer/renderer is employed. It is interesting to note that the number of output channels can be different from the number of input channels, making it flexible for different channel configurations.

Using the SAOC approach, both the level and position of each audio object can be adjusted. Any loudspeaker set up is also possible for rendering. Additionally, SAOC also provides users with the ability to render either vocals only for enhancing speech intelligibility or background music only for karaoke. This approach is essential for audiovisual applications that require higher spatial fidelity and synchronisation between the video and audio scenes. However, this approach has difficulty for perfectly capturing audio objects, due to the possibility of interfer-

Figure 3.3: Block diagram of an SAOC approach

ence from other objects, particularly for capturing moving objects. Considering this difficulty, in some cases, audio objects may be recorded separately.

Based on this approach, the MPEG Audio Group has also released a new standard called MPEG SAOC. The hybrid QMF filterbank, as applied in MPS, is adopted for time-frequency decomposition. Object Level Difference (OLD) and Inter-Object Cross-coherence (IOC), along with other optional parameters, such as object energies are extracted as object parameters and are transmitted as side information of a mono or stereo downmix signal. For applications, such as karaoke, that need to completely remove particular audio objects, residual coding is employed.

At the decoder side, a trans-coding block may be utilised for rendering any multi-channel audio, including 5.1, to convert the object parameters to spatial parameters of MPS. An MPS rendering engine is used instead of a pure SAOC decoding block. However, a SAOC decoder without a trans-coding block is applied for single or two-channel rendering, including stereo and binaural rendering. The MPEG SAOC implementation, in combination with DirAC to capture audio object, has been reported in [35]. In this implementation, a DirAC encoding block is applied for capturing multiple single audio objects.

3.3.5 Other Parametric Techniques

Other techniques, such as applying Karhunen-Loève Transform (KLT) [112–115] and Discrete Cosine Transform (DCT) [116] for compressing a number of audio signals, also exist to remove the inter-channel redundancy. Both approaches are included in MPEG-2 Advance Audio Coding (AAC) [1, 117]. A compression can also be realised by exploiting the inter-channel predictability [118]. Coefficients that describe the appropriate prediction parameters are extracted and then sent to the decoder, along with residual signals. Basic configuration of this coder, the TTT module, and the corresponding R-TTT, has also been implemented in MPS for converting three channels into stereo.

3.4 Comparison of Coding Techniques

Some key parameters, such as the bitrate (coding efficiency), perceptual quality, source localisation and complexity can be used in benchmarking existing multi-channel audio coders. Depending on the application, one or more key parameters may be given priority over others, as there is usually a trade-off between them.

For home consumer applications, perceptual quality may be considered to be the most important factor. Coding efficiency and low complexity may be the key parameters for mobile communications, while the source localisation accuracy may be targeted for advanced, spatialised teleconferencing.

When the spatial audio coding technique was first introduced, the key parameter was the coding efficiency as defined in terms of bitrate required for transmission. For instance, MPEG Surround, in combination with AAC stereo, has been shown to perform better in terms of perceptual quality for transmitting 5.1 audio than MP3 Surround, Dolby Prologic II in combination with AAC stereo and AAC multichannel, while maintaining the transmitted bitrate at 160 kb/s [37]. However, for applications where more bandwidth is available, such as the HDTV, the quality of the reconstructed audio is very significant. The perceptual quality of AAC for transmitting 5.1 channel audio at 320 kb/s still outperforms the quality of MPEG Surround in combination with HE-AAC at 160 kb/s [119].

A comprehensive listening test has been conducted for comparing codecs for high quality multichannel audio rendering intended for HDTV applications. Older codecs, such as the Digital Theater System (DTS), which operated at 1.5 Mbps, and Dolby Digital, which operated at 448 kb/s, performed better than other codecs, including MPEG Surround in combination with HE-AAC. It was shown that at present there is no audio codec that can meet the requirement of delivering high quality surround sound at bitrates lower than 448 kb/s [120, 121].

Some experiments have been conducted to evaluate source localisation consistency. It has been reported that MP3 Surround maintains the consistency of the spatial sound image delivered, while the Dolby Prologic II results in changes to the positions of sound components in some cases [16]. Another test showed that S³AC, which is mapping 5-channel into a stereo audio signals without side information, performs better in terms of source localisation than MPEG Sur-

round, which is coding 5-channel signals into a stereo downmix signal with side information. In objective evaluation, MPEG Surround introduced 10.28° average azimuth error while S³AC gave only 1.39° , even though subjective evaluation tests did not show much difference between both audio coders in that localisation error cannot be accurately detected [108].

With regards to complexity in terms of required processor speed, in a metric of Mega Cycles Per Second (MCPS), HE-AAC v1/v2 and MPEG Surround decoders in high quality mode score the highest where each of them requires 150 and 140 MCPS respectively, making them the most complicated when compared to AAC-LC, AC-3 and Enhanced AC-3 decoders [122].

3.5 Conclusions

This chapter has presented a thorough review of spatial audio coding techniques. Starting by discussing on psychoacoustic background such as masking threshold, various techniques on spatial audio processing and coding have been presented. Various techniques that can be used for encoding multiple audio signals have been elaborated including approaches such as matrixed surround, discrete multi-channel and spatial audio object coding. Comparison of different modern coding techniques has been presented at the end of the chapter giving awareness that AAC and MPEG Surround are the state-of-the-art coding techniques in spatial audio coding.

Chapter 4

Advanced Audio Coding and MPEG Surround

This chapter provides a detailed description of MPEG AAC and MPEG Surround as state-of-the-art techniques for delivering multichannel audio signals. Three essential modules in AAC, comprising MDCT transform, quantisation, and noiseless coding, are presented. Additionally, the time-frequency decomposition, the spatial analysis and synthesis as well as the quantisation and coding of parameters employed in MPEG Surround are also discussed in detail.

4.1 Advanced Audio Coding

Following the success of the MPEG-1 audio development, the MPEG-2 Audio committee started work in 1994 for a new audio coding standard intended for multichannel audio. It is targeted to be capable of delivering a higher quality of audio signal than the one achieved by the MPEG-1 backward compatible coding

technique. The work was completed in 1997. It was later renamed MPEG-2 Advanced Audio Coding (MPEG-2 AAC), as it made use of all available advanced techniques at the time of developing. The development of the MPEG-2 AAC standard was aimed at achieving indistinguishable audio quality when encoding 5-channel audio signals at an operating bitrate of 384 kb/s. The term indistinguishable audio quality refers to the one specified in the ITU-R TG10-2. The listening tests undertaken in 1996 at BBC UK and NHK Japan demonstrated that the quality requirement has been met at a bitrate of 320 kb/s.

The MPEG AAC is specified as a modular approach consisting of a series of self-contained modules or tools. As shown in the block diagram given in Fig. 4.1, several tools are included: gain control, Modified Discrete Cosine Transform (MDCT) and Inverse MDCT (IMDCT), Temporal Noise Shaping (TNS), intensity/coupling, prediction, Mid/Side (M/S) stereo coding, and quantisation and coding. Corresponding tools are also employed at the AAC decoder as shown in Fig. 4.2.

The gain control module [1] aims to provide a possibility to reproduce audio signal on the decoder side with lower bandwidth and lower sampling rates. For this purpose, the time domain audio signal is decomposed into 4 equal-bandwidth frequency bands by a Pseudo Quadrature Mirror Filterbank (PQMF). In order to provide minimum error when producing such signal with a lower bandwidth and sampling rate, the output signals of the PQMF may be amplified or attenuated. The amount of amplification or attenuation are transmitted as gain control data. On the decoder side, gain compensation is applied based on the gain control data. Then, either 1, 2, 3 or 4 frequency bands can be transformed back into time domain with a corresponding sampling rate.

The TNS [123] can be applied to reduce the quantisation error due to the presence of transient event. This transient event has a flat frequency spectrum. Thus,

Figure 4.1: Block diagram of MPEG AAC Encoder

Figure 4.2: Block diagram of MPEG AAC Decoder

the spectral coefficients can be effectively predicted. The prediction residual is produced as the output of the TNS and the prediction parameters are transmitted to the decoder. On the decoder side, the prediction parameters are applied to the prediction residual to recreate the spectral coefficients.

The intensity/coupling module [11] consist of two function: intensity coding and coupling. The intensity coding can be used to encode a pair of audio channels such as the left and right channels in 5.1 audio channels while the coupling function can be used to encode two uncorrelated audio channels. A single downmix audio channel accompanied with intensity or coupling parameters is generally produced from the two input channels. The single downmix channel is transmitted to the decoder along with the intensity/coupling parameters. For recreating back two audio channels, intensity/coupling parameters is applied to the single downmix channel.

The prediction module [1] comprises of a set of predictors. Each predictor is used to estimate a spectral coefficient based on the spectral coefficients in the previous frames. The difference between the spectral coefficients and its predicted value, named as residual component, is quantised and sent to the decoder. As the TNS reduces redundancy of spectral coefficients in consecutive frames, it is effective only for audio signal that has strong stationary components.

The M/S stereo coding [11] is used for encoding a pair of audio channels, such as the left and right channels, in encoding multichannel audio signals. The left and right channels are transformed into a middle and a side channels. The middle channel is the normalised sum signals while the side channel is the difference channel. The middle and the side channels are fed to the next module in the encoding process.

As previously explained, the MPEG AAC offers three profiles to provide users with a trade-off between the audio quality and the hardware complexity. The

three profiles are detailed below.

- **Main Profile.** This profile is provided to achieve the best audio quality at the given operating bitrate at a cost of a higher memory and processing power than those required in the Low Complexity (LC) profile. Except the gain control, all other tools, such as TNS, intensity/coupling and prediction, can be operated.
- **Low Complexity (LC) Profile.** A slightly lower performance can be achieved by operating AAC in the LC configuration when the prediction and the gain control are not applied. The TNS order is also limited. For further reducing the complexity the intensity/coupling as well as the M/S tools can also be eliminated.
- **Scalable Sampling Rate (SSR) Profile.** The SSR profile is used to obtain a lower bandwidth and lower sampling rate audio signals. The reconstructed audio output may have a bandwidth of either 18, 12, or 6 KHz. For this profile, the gain control module is necessary as previously discussed.

The process of encoding an audio signal can be explained as follows. The MDCT transform is used to transform the audio signal into spectral components. At the same time, a perceptual model is applied to compute a Signal to Masking Ratio (SMR), which is used in the quantisation process. The spectral coefficients are filtered by the TNS module, producing prediction residuals. For multichannel audio signals, intensity stereo coding can be employed where the energy envelope is calculated and then transmitted. In order to take advantage of subsequent frames, the prediction can be applied while the M/S stereo coding is performed for transmitting the normalised sum and difference signals. In the following sections the MDCT transform, the quantisation process, as well as the noiseless coding process, will be discussed in detail.

4.1.1 MDCT Transform

The process of converting time domain audio signals into spectral frequency representations is a fundamental component in the AAC standard. The time domain audio signal, $x[i]$, is transformed to spectral coefficients, $X[k]$, by means of MDCT as follows:

$$X[k] = 2 \sum_{i=0}^{N-1} x[i] \cos \left(\frac{2\pi}{N} (i + i_0) \left(k + \frac{1}{2} \right) \right) \quad (4.1)$$

where i is the index of audio sample ($i = 0, 1, \dots, N - 1$) with N is the length of input audio frame, k is the spectral coefficient index, and i_0 is a constant ($i_0 = \frac{1}{2} \left(\frac{N}{2} + 1 \right)$) [124]. For transforming the spectral coefficients back to a time domain audio signal, the IMDCT is used as given by,

$$x[i] = \frac{2}{N} \sum_{k=0}^{\frac{N}{2}-1} X[k] \cos \left(\frac{2\pi}{N} (i + i_0) \left(k + \frac{1}{2} \right) \right) \quad (4.2)$$

where the spectral coefficient index, $k = 0, 1, \dots, \frac{N}{2}$.

The length of the input audio frame, N , is normally set to 2048 time samples to achieve high coding efficiency. However, the block of 2048 time samples can cause a problem, especially for an audio frame that has a transient event which causes artifacts. To address this problem, the length of the audio frame can be switched to 256 time samples. This number is chosen as the best compromise value by considering the frequency resolution and the pre-echo suppression at a bitrate of around 64 kb/s per audio channel. A switching mechanism is applied as a function of input audio waveform. The transition from a long to a short frame and vice versa is designed to be seamless by applying a perfect aliasing cancellation. During these transitions, a start and stop window is applied, preserving the time domain aliasing cancellation properties of the MDCT and IMDCT transforms. In addition, it also maintains block alignment. In order to maintain synchronisation

Table 4.1: The number of coefficients in each scale factor band [1]

SFB Index	Number of Coefficients
1 - 10	4
11 - 17	8
18 - 21	12
22 - 23	16
24 - 25	20
26 - 27	24
28 - 29	28
30 - 48	32
49	96

among the audio channels in encoding multichannel audio signals, the short frame should be used in a group of eight frames. Hence, the implementation of a short frame in a channel of multichannel audio signals will not cause a lack of alignment among all audio channels.

4.1.2 Quantisation

Two of the main objectives of the quantisation process are, first, to represent the spectral coefficients the way that the quantisation error meets the requirement that was computed in the psychoacoustic model. The second objective is to encode the audio signal at the desired operating bitrate. To achieve both objectives, the quantisation process involves not only the psychoacoustic model but also a bit reservoir that is needed to effectively distribute bits among consecutive audio frames.

A frame, comprised of 1024 spectral coefficients, is segmented as scale factor

Figure 4.3: Flowchart of the inner iteration loop

bands, where the number of coefficients within each band is determined based on the bandwidth of the critical bands [125,126]. As many as 49 scale factor bands are used as given in Table 4.1 [1].

For each band, a scale factor is determined, and the spectral coefficients are quantised as follows:

$$\bar{X}[k] = \text{sign}(X[k]) \cdot \text{nint} \left[\left(\frac{|X[k]|}{\sqrt[4]{2^{S_F}}} \right)^{0.75} \right] \quad (4.3)$$

where k is the spectral coefficient index, $\bar{X}[k]$ is the quantised spectral coefficient with its value is limited from -8191 to $+8191$ [1]. Consequently, 14 bits are required to represent the index of the quantised value. The *nint* is a nearest integer function, $X[k]$ is the spectral coefficient, and S_F is the scale factor. The

sign is a signum function represented as,

$$\text{sign}(u) = \begin{cases} -1 & \text{if } u < 0 \\ 0 & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases} \quad (4.4)$$

The quantiser may utilise a psychoacoustic model [45] to compute the maximum allowed distortion while an analysis by synthesis procedure is carried out to select the best scale factor and quantised spectral coefficient, resulting in minimal error. The AAC standard does not specify an optimal way for quantisation. However, an iteration process comprising inner and an outer iteration loops is provided for reference. The simplified illustration of the iteration process can be described as follows. The inner iteration loop, as shown in Fig. 4.3, aims to encode the spectral data with the given available bits. To achieve this goal an initial quantiser stepsize is chosen. Using this initial stepsize the spectral data are quantised. The amount of bits required to encode the spectral data is calculated. If the number of required bits exceeds the available bits, then the quantiser stepsize is increased, and the process is repeated. Otherwise the inner iteration stops.

On the other hand, the function of the outer iteration loop, as depicted in Fig. 4.4, is to fulfil as much as possible the allowed distortion computed by the psychoacoustic model. For performing this algorithm an initial scale factor is determined. Then, the inner iteration loop is performed. For every scale factor band, the quantisation error is computed and compared to the allowed distortion supplied by the psychoacoustic model. The parameters are restored if the error is minimum. For every scale factor band with an actual distortion higher than the allowed distortion, the scale factor is amplified. The method used to amplify the scale factor is not specified. It is left up to the encoders implementation. The process is repeated where new amplified scale factors are used. The iteration stops when either one of two conditions apply. The first is the condition when

all scale factors need to be amplified, and the second is if no scale factor needs to be amplified. If one of those conditions applies then the best result is restored. The amplified scale factor obviously requires more bits for encoding. Because the amount of bits available for encoding is constant, the quantiser stepsize in the inner iteration loop has to be adapted.

4.1.3 Noiseless Coding

Huffman coding is then carried out for further compression, where several Huffman codebooks, the details given in Table 4.2, designed for different sets of spectral data, are provided. The quantised spectral coefficients of one or more scale factor bands are grouped and then encoded with an appropriate codebook depending on the maximum absolute value of the quantised spectral coefficients within the group. Groups that have all of their quantised spectral coefficients at zero values are associated to codebook 0, where all of the quantised spectral coefficients within this group do not need to be transmitted. Particular attention is given to the last codebook which is provided for groups with the maximum absolute values greater than or equal to 16 when a special escape sequence is used [124].

A set of n -tuples $[1, 7]$ of quantised spectral coefficients within a group, consisting of either 2 or 4 coefficients, are then represented as Huffman codewords. To keep their size small, most codebooks are given unsigned values. Thus, the sign of each non zero coefficient is represented as an additional bit appended to the unsigned codeword.

Figure 4.4: Flowchart of the outer iteration loop

Table 4.2: Description of AAC Huffman Codebooks

Codebook Index	n-Tuple Size	Maximum Absolute Value	Signed Values
0		0	
1	4	1	yes
2	4	1	yes
3	4	2	no
4	4	2	no
5	2	4	yes
6	2	4	yes
7	2	7	no
8	2	7	no
9	2	12	no
10	2	12	no
11	2	16 (ESC)	no

4.2 MPEG Surround

A block diagram of MPEG Surround is shown in Fig. 4.5. Multiple audio signals are typically downmixed into a mono or stereo signal, which is further encoded by an existing audio coder such as MPEG 1/2 layer 3 (MP3) [50–52] and High Efficiency Advanced Audio Coding (HE-AAC) [8, 9]. Channel Level Differences (CLDs), Inter-Channel Coherence (ICC), and Channel Prediction Coefficient (CPC) are extracted as the spatial parameters and transmitted as well as the downmix signal. In addition the residual signal is computed and transmitted for high quality audio reconstruction. When operating at a low bitrate the residual signal is replaced by a synthetic signal constructed using a decorrelator [127–129].

The audio signal in each channel is decomposed by a filterbank into subband

signals where the spatial analysis is carried out. Every subband signal is segmented into frames and then windowed so that the frame length can be varied, making it adaptable to most existing audio encoders used for compressing the downmix signal. In order to closely match the frequency resolution of the human auditory system, adjacent subband signals are grouped together as a parameter band, sharing a common spatial parameter. Since at lower frequencies a higher resolution is required, a smaller number of subbands are grouped together for a single spatial parameter. Conversely, at higher frequencies a lower resolution is allowed, where a parameter band can comprise a larger number of subband signals.

In order to provide rendering flexibility, several options for the number of parameter bands are available, allowing adjustable frequency-domain resolution in the spatial analysis. The encoder's rate-distortion scalability [119, 128] also provides the flexibility required to operate at various bitrates. Moreover, MPS encoder offers an option to transmit an artistic downmix signal [106], available from the content provider.

4.2.1 Frequency Decomposition

MPEG Surround uses a hybrid Quadrature Mirror Filterbank (QMF) [37, 83, 119] for performing the frequency decomposition. An Analysis QMF (A-QMF) is used to decompose the audio signal in each channel into 71 hybrid (non-uniform) subband signals, while a Synthesis QMF (S-QMF) is used to transform them back into a time domain signal. The A-QMF consists of 2 stages as given in Fig. 4.6. The first stage is a filterbank applied to decompose a fullband audio signal into 64 uniform (equal-bandwidth) subband signals. The audio signal in each channel, $s[i]$, (the channel index is ignored for simplicity) is firstly convolved with a set of

Figure 4.5: Block diagram of MPEG Surround.

impulse responses, $G_{0,m_0}[i]$ [130], given as,

$$G_{0,m_0}[i] = h_0[i] \exp \left[j \frac{\pi}{2M_0} (m_0 + 0.5)(2i - 1) \right] \quad (4.5)$$

where $h_0[i]$ is the prototype filter with a filter length of 640, $m_0 (m_0 = 0, \dots, M_0 - 1)$ is the index of the uniform subband, and $M_0 = 64$ is the total number of the uniform subbands. Furthermore, each uniform subband signals is proceeded to a down-sampler with a factor of M_0 to obtain lower rate signals, hence, can simplify the subsequent processing of the subband signals. Thus, the output of the down-sampling process can be represented as,

$$x_{m_0}[n] = (s[i] * G_{0,m_0}[i])[M_0 n] \quad (4.6)$$

where $x_{m_0}[n]$ is a set of down-sampled subband signals.

The second stage of the A-QMF aims to further enhance the spectral resolution of the subband signals at low-frequency regions. The first subband signal is decomposed into 8 sub-subbands, named hybrid bands, by convolving it with a set of impulse responses, $G_{1,m_1}[n]$ [130], given by,

$$G_{1,m_1}[n] = h_1[n] \exp \left[j \frac{\pi}{4} (m_1 + 0.5)(n - 6) \right] \quad (4.7)$$

where $m_1 (m_1 = 0, 1, \dots, 7)$ is the index of the convolution output, and $h_1[n]$ is the prototype filter with a length of 13. The second and the third subband signals are convolved with another filterbank with impulse responses, $G_{2,m_2}[n]$ [130], given by,

$$G_{2,m_2}[n] = h_2[n] \cos [\pi (m_2 - 1)(n - 6)] \quad (4.8)$$

where $m_2, (m_2 = 0, 1)$ in order to obtain 2 hybrid bands from each subband. The prototype filter, $h_2[n]$, also has a filter length equal to 13. Since some of the passbands of the second stage filterbank coincide with the corresponding stopbands of the first stage filterbank, 10 hybrid bands are created instead of 12. As the 61 remaining subband signals are not decomposed any further and

Figure 4.6: Block diagram of Analysis Quadrature Mirror Filterbank (A-QMF)

in order to make them synchronised to the 10 hybrid band signals, they are just delayed by the same time as the delay introduced in the second stage filterbank. Hence, the 10 hybrid band signals, along with the 61 delayed subband signals, form the 71 hybrid bands of A-QMF.

The S-QMF is performed to convert the 71 hybrid bands back into a fullband audio signal. The first step of the S-QMF is the summation of the first 10 hybrid band signals to transform them back into their 3 original subband signals. Subsequently, the newly formed 3 subband signals, along with the remaining 61 hybrid band signals, representing the original 64 subband signals, are convolved with a second stage filterbank having impulse responses, $H_{m_0}[i]$ [130], as follows:

$$H_{0,m_0}[i] = \frac{1}{M_0} h_0[i] \exp \left[j \frac{\pi}{2M_0} (m_0 + 0.5)(2i - 255) \right] \quad (4.9)$$

where $h_0[i]$, m_0 , and M_0 are defined similar to those in (4.5).

4.2.2 Spatial Analysis and Synthesis

The MPEG Surround system comprises two pairs of elementary building blocks for channel conversion and the reverse process: One-To-Two (OTT) and Two-To-Three (TTT) modules. The OTT module is used to convert a single channel to two channels while the TTT module is used to convert two channels to three channels. The reverse conversions are done by the Reverse OTT (R-OTT) module and the Reverse TTT (R-TTT) module. CLDs, ICCs, and the residual signal are extracted from the R-OTT module, whereas CPCs, ICCs and the residual signal are calculated from the R-TTT module. The whole process in the encoder and decoder is built up by combining several OTT and TTT modules in a tree structure. This section describes the process of extraction of the CLD, ICC as well as the residual signal, as they are implemented within the proposed framework.

Figure 4.7: Block diagram of (a) OTT module and (b) R-OTT module as used in MPS.

4.2.2.1 The OTT and R-OTT modules

The schematic of the OTT and R-OTT modules are depicted in Fig. 4.7. The R-OTT converts two input channels into one output channel. CLD and ICC are extracted as spatial parameters in a parameter band which can be a subband or a group of subbands. Conversely, the OTT re-synthesises two channels from one channel, utilising the spatial parameters. A single value for CLD and ICC parameters is calculated for each parameter band, b . However, the downmix signal and the residual signal are calculated for each subband, s .

The CLD, denoted as C , is defined as the ratio between the energies of the signals in the first and second channels for a parameter band b :

$$C^b = \frac{e_{x_1}^b}{e_{x_2}^b} = \frac{\sum_n x_1^b[n] x_1^{b*}[n]}{\sum_n x_2^b[n] x_2^{b*}[n]} \quad (4.10)$$

where the $x_1^{b*}[n]$ and $x_2^{b*}[n]$ represent the complex conjugate of $x_1^b[n]$ and $x_2^b[n]$, respectively. For transmission, the quantised logarithmic values of the CLDs are

conveyed. The parameter ICC, denoted as I , is determined by,

$$I^b = \text{Re} \left\{ \frac{\sum_n x_1^b[n] x_2^{b*}[n]}{\sqrt{e_{x_1}^b e_{x_2}^b}} \right\} \quad (4.11)$$

describing the degree of correlation between the input channels.

The downmix signal $y[n]$ is a scaled sum of the input signals within each subband. One possible representation of the downmix signal can be written as

$$y^s[n] = \frac{x_1^s[n] + x_2^s[n]}{a^s + b^s} \quad (4.12)$$

where the energy constants a and b calculated as,

$$(a^s + b^s)^2 = \frac{e_{x_1}^s + e_{x_2}^s + 2I\sqrt{e_{x_1}^s e_{x_2}^s}}{e_{x_1}^s + e_{x_2}^s} \quad (4.13)$$

representing the energy preservation constraint [37].

Furthermore, the residual signal $r[n]$ in each subband is determined from the following decomposition:

$$x_1^s[n] = a^s y^s[n] + r^s[n] \quad (4.14a)$$

$$x_2^s[n] = b^s y^s[n] - r^s[n] \quad (4.14b)$$

which produces a single residual signal for reconstructing both $x_1^s[n]$ and $x_2^s[n]$.

At the decoder side, both audio signals are recreated by estimating a and b as follows:

$$\hat{a} = X \cos(A + B) \quad (4.15a)$$

$$\hat{b} = Y \cos(A - B) \quad (4.15b)$$

where the X, Y, A , and B variables given as

$$X = \sqrt{\frac{\hat{C}}{1 + \hat{C}}} \quad (4.16a)$$

$$Y = \sqrt{\frac{1}{1 + \hat{C}}} \quad (4.16b)$$

$$A = \frac{1}{2} \arccos(\hat{I}) \quad (4.16c)$$

$$B = \tan \left[- \left(\frac{X - Y}{X + Y} \right) \arctan(A) \right] \quad (4.16d)$$

are determined from the quantised values of CLD, $\hat{C}$, and the quantised values of ICC, $\hat{I}$. Hence, both signals can be reconstructed as

$$\hat{x}_1[n] = \hat{a}\hat{y}[n] + \hat{r}[n] \quad (4.17a)$$

$$\hat{x}_2[n] = \hat{b}\hat{y}[n] - \hat{r}[n] \quad (4.17b)$$

which are similar to (4.14) but use the decoded downmix and residual signals, $\hat{y}[n]$ and $\hat{r}[n]$, respectively. The indices for the subbands and parameter bands have been ignored for notation simplicity.

4.2.2.2 The TTT and R-TTT modules

The TTT and R-OTT modules are shown in Fig. 4.8. The TTT module is designed to convert a stereo audio (two channels) to three audio channels. Reciprocally, the R-TTT module is used to convert three audio channels into a stereo audio channel. Two encoding modes are available: the prediction and the energy modes. Using the prediction mode, the Channel Prediction Coefficient (CPC) is transmitted as a spatial parameter. Additionally, Channel Level Difference (CLD) is calculated and transmitted as side information.

Let the left, right, and center channels, $x_L[n], x_R[n], x_C[n]$, become the inputs of the R-TTT module while the other left and the right channel, $y_L[n], y_R[n]$,

Figure 4.8: Block diagram of (a) the TTT and (b) the R-TTT modules.

are the outputs. The two outputs, along with an auxiliary signal, $y_C[n]$, can be represented as linear combinations of the input signals as below

$$y_L[n] = x_L[n] + \frac{1}{2}\sqrt{2} x_C[n], \quad (4.18a)$$

$$y_R[n] = x_R[n] + \frac{1}{2}\sqrt{2} x_C[n], \quad (4.18b)$$

$$y_C[n] = x_L[n] + x_R[n] - \frac{1}{2}\sqrt{2} x_C[n]. \quad (4.18c)$$

When a prediction mode is used, two CPC coefficients are calculated in such a way that the auxiliary signal, $y_C[n]$, is as close as the linear combination of the two outputs according to

$$\hat{y}_C[n] = \gamma_1 y_L[n] + \gamma_2 y_R[n]. \quad (4.19)$$

The residual signal is defined as the difference between the auxiliary signal, $y_C[n]$, and the predicted auxiliary signal, $\hat{y}_C[n]$, as

$$r_C[n] = y_C[n] - \hat{y}_C[n]. \quad (4.20)$$

In case the residual signal is not transmitted to the decoder (this is the case for low bitrate transmission) the corresponding energy loss can be described by transmitting the Inter-Channel Coherence (ICC).

Using the prediction mode, a reliable estimation of the auxiliary signal is required for the decoder. If it is difficult to guarantee the reliability of the prediction mode then the TTT, using an energy mode, can be applied. The energy mode is performed by describing the relative energy distribution among the three input channels. Two CLD parameters can be transmitted. The prediction and energy mode can be performed independently for each parameter band.

In the decoding process using the prediction mode, the TTT module is employed to recreate three channels from a stereo channel as below

$$\hat{x}_L[n] = \frac{2}{3}y_L[n] - \frac{1}{3}y_R[n] + \frac{1}{3}\hat{y}_C[n], \quad (4.21a)$$

$$\hat{x}_R[n] = -\frac{1}{3}y_L[n] + \frac{2}{3}y_R[n] + \frac{1}{3}\hat{y}_C[n], \quad (4.21b)$$

$$\hat{x}_C[n] = \frac{1}{3}\sqrt{2}y_L[n] + \frac{1}{3}\sqrt{2}y_R[n] - \frac{1}{3}\sqrt{2}\hat{y}_C[n]. \quad (4.21c)$$

where the predicted auxiliary signal is calculated from the CPCs and the residual signal as in (4.19) and (4.20). If the residual signal is not available (i.e. not transmitted from the encoder side), the TTT module has two options to recreate the three outputs. The first one is to apply a suitable gain to each output based on the CPC and ICC as below

$$\hat{x}_L[n] = \left(\frac{\gamma_1 + 2}{3\text{ICC}} \right) y_L[n] - \left(\frac{\gamma_2 - 1}{3\text{ICC}} \right) y_R[n], \quad (4.22a)$$

$$\hat{x}_R[n] = \left(\frac{\gamma_1 - 1}{3\text{ICC}} \right) y_L[n] - \left(\frac{\gamma_2 + 2}{3\text{ICC}} \right) y_R[n], \quad (4.22b)$$

$$\hat{x}_C[n] = \left(\frac{\sqrt{2}(1 - \gamma_1)}{3\text{ICC}} \right) y_L[n] - \left(\frac{\sqrt{2}(1 - \gamma_2)}{3\text{ICC}} \right) y_R[n]. \quad (4.22c)$$

The second is to create a synthetic residual signal by means of a decorrelator.

4.2.2.3 Parameter Band

MPS offers a scalable spatial resolution, where several maximum numbers of parameter band: 4, 5, 7, 10, 14, 20, and 28, can be used. For instance, when 20

Table 4.3: Mapping of 71 hybrid bands to 20 parameter bands

PB Index	HB Index	PB Index	HB Index
1	2-3	11	12
2	1	12	13
3	4	13	14
4	5	14	16
5	6	15	17-18
6	7	16	19-21
7	8	17	22-25
8	9	18	26-30
9	10	19	31-42
10	11	20	43-71

parameter bands are used, the 71 hybrid bands are grouped into 20. Table 4.3 provides the mapping of 71 Hybrid Bands (HB) to 20 Parameter Bands (PB) [83].

4.2.3 Quantisation and Coding

The logarithmic value of the extracted channel level difference, calculated as $10 \cdot \log_{10}(C)$, is represented using one of the following non-uniform quantisation values [130],

$$\text{CLD} = [-150, -45, -40, -35, -30, -25, -22, -19, -16, -13, -10, -8, \\ -6, -4, -2, 0, 2, 4, 6, 8, 10, 13, 16, 19, 22, 25, 30, 35, 40, 45, 150]$$

where 5 bits are allocated to send the index of this quantised CLD. Additionally, the extracted inter-channel coherence, I , is represented by one of the following

non-uniform quantisation values,

$$\text{ICC} = [-0.99, -0.589, 0, 0.36764, 0.60092, 0.84118, 0.937, 1]$$

with 3 bits are allocated for transmitting the index of the quantised ICC.

The residual signal is encoded in the same way as the LC-AAC [7]. The MPS standard specifies the transformation of the residual signal from the subband domain to the spectral coefficients of the MDCT transform.

4.2.4 Rate-distortion Scalability

For a trade-off between the bitrate and the quality of the reconstructed audio signals, the residual signals can be transmitted in three different ways [119]. First, the bandwidth of each residual signal can be varied between zero and full signal bandwidth. Second, the bitrate of each residual signal can be made variable. Third, there is an option to transmit the residual signal only from a single or more particular R-OTT modules in the tree structure. For the two most extreme cases, zero and the full bandwidth of the residual signals can be transmitted. In case no residual signal is transmitted, the MPS works in an enhanced matrix mode [130], while full waveform reconstruction is achieved when full bandwidth of residual signals are transmitted to the decoder.

The other options to increase or decrease the bitrates, to achieve the desired quality of reconstructed audio signals, are to choose appropriate spatial, time and frequency resolutions of the spatial parameters. For instance, to achieve low bitrate transmission, a single ICC parameter can be transmitted [119].

4.3 Conclusions

This chapter has provided a detailed review of MPEG Advanced Audio Coding (AAC), as well as MPEG Surround, as cutting-edge technology in the area of spatial audio coding. The MPEG AAC was developed based on a discrete multichannel approach where each audio channel is encoded separately, even though inter-channel redundancy can be exploited. Additionally, MPEG Surround was developed based on a spatial audio coding approach where channel level difference and inter-channel coherence are extracted as spatial parameters. Then, multichannel audio signals are downmixed into a mono or stereo audio signals. The downmix signals are encoded further by an audio encoder and transmitted along with the spatial parameters as side information.

Chapter 5

A Closed-loop R-OTT Module

This chapter presents the framework for an Analysis by Synthesis Spatial Audio Coding (AbS-SAC) approach. First the implementation of the R-OTT module within the AbS-SAC framework is presented, resulting in a new adapted version called the closed-loop R-OTT module. This closed-loop R-OTT module is then evaluated by measuring the SNR. The experimental results are compared with the MPS, which uses the conventional open-loop R-OTT module. The results clearly show that significant SNR improvement is achieved.

5.1 Analysis by Synthesis Framework

The framework of the AbS-SAC [131, 132] is given in Fig. 5.1. A spatial synthesis block, similar to that performed at the decoder side, is embedded within the AbS-SAC encoder as a model for reconstructing multichannel audio signals. Assuming that there is no channel error, the audio signals synthesised by the model in the encoder will be exactly the same as the reconstructed audio signals at the decoder side. The error minimisation block is used to compare the input signals with the

Figure 5.1: Analysis by synthesis framework for encoding spatial audio.

reconstructed signals based on a suitable criteria such as Mean Squared-Error (MSE) or other perceptual relevant criteria. The resultant downmix and residual signals as well as the optimal spatial parameters are then transmitted to the decoder.

Based on this framework, various implementations are possible. They are listed as follows:

- Any approach of spatial analysis and synthesis can be implemented.
- Different numbers of input and output channels can also be used.
- Various types of suitable error criterion can be utilised.
- For taking into consideration the error introduced in the communication channel, a block modeling the channel error can be inserted between the spatial analysis and synthesis block.
- The AbS-SAC approach can be implemented for only a single parameter recalculation without multiple iteration. This can be considered as a modification to the original open-loop spatial analysis block.

- The implementation can be intended to find the optimal synthesised signals by performing the trial and error procedure. Either the original blocks or the adapted version of the spatial analysis and synthesis can be applied.

5.2 Implementation of the R-OTT Module within the AbS-SAC Framework

The OTT and R-OTT modules, as used in MPS, can be implemented within the AbS-SAC framework, where two channels of original audio signals are fed to an R-OTT module as the spatial analysis block. On the other hand, the OTT module is performed as the spatial synthesis block for reconstructing two channels of synthesised audio signals, so that (4.17) becomes the formula of the model. Assuming that the MSE is used as an error criterion, the error signals that are the differences between the original audio signals, $x_1[n], x_2[n]$, and the synthesised signals, $\hat{x}_1[n], \hat{x}_2[n]$, can be written in vector form as,

$$\mathbf{e}_1 = \mathbf{x}_1 - \hat{\mathbf{x}}_1 \quad (5.1a)$$

$$\mathbf{e}_2 = \mathbf{x}_2 - \hat{\mathbf{x}}_2 \quad (5.1b)$$

and then represented as,

$$\mathbf{e}_1 = \mathbf{x}_1 - \hat{a}\hat{\mathbf{y}} - \hat{\mathbf{r}} \quad (5.2a)$$

$$\mathbf{e}_2 = \mathbf{x}_2 - \hat{b}\hat{\mathbf{y}} + \hat{\mathbf{r}} \quad (5.2b)$$

by substituting (4.17). Furthermore, the MSEs for both channels can be represented as,

$$\text{MSE}_1 = (\mathbf{x}_1 - \hat{a}\hat{\mathbf{y}} - \hat{\mathbf{r}}) \cdot \mathbf{e}_1^T \quad (5.3a)$$

$$\text{MSE}_2 = (\mathbf{x}_2 - \hat{b}\hat{\mathbf{y}} + \hat{\mathbf{r}}) \cdot \mathbf{e}_2^T \quad (5.3b)$$

where $\mathbf{e}_1^T$ and $\mathbf{e}_2^T$ being the transpose operations of $\mathbf{e}_1$ and $\mathbf{e}_2$, respectively.

Assuming that each error signal does not have all-zero values, so that the minimum mean squared-errors for both channels, minimum of MSE_1 and MSE_2 , can be obtained by requiring every component of the error signals to be orthogonal to the corresponding transposed error signal which means that there is no correlation between each component of the error signals with the corresponding transposed error signal. These are described as follows:

1. First component: both input signals, $\mathbf{x}_1$ and $\mathbf{x}_2$, have to be orthogonal to the corresponding transposed error signals, $\mathbf{e}_1^T$ and $\mathbf{e}_2^T$. These can be written as,

$$\mathbf{x}_1 \mathbf{e}_1^T = \mathbf{x}_1 \mathbf{x}_1^T - \hat{a} \mathbf{x}_1 \hat{\mathbf{y}}^T - \mathbf{x}_1 \hat{\mathbf{r}}^T = 0 \quad (5.4a)$$

$$\mathbf{x}_2 \mathbf{e}_2^T = \mathbf{x}_2 \mathbf{x}_2^T - \hat{b} \mathbf{x}_2 \hat{\mathbf{y}}^T + \mathbf{x}_2 \hat{\mathbf{r}}^T = 0 \quad (5.4b)$$

which can be simplified as

$$\hat{\mathbf{r}} = \mathbf{x}_1 - \hat{a} \hat{\mathbf{y}} \quad (5.5a)$$

$$\hat{\mathbf{r}} = \hat{b} \hat{\mathbf{y}} - \mathbf{x}_2 \quad (5.5b)$$

where the decoded residual signal is represented as two different expressions.

2. Second component: The decoded downmix signal $\hat{\mathbf{y}}$ has to be orthogonal to both transposed error signals $\mathbf{e}^T$ (i.e., $\hat{\mathbf{y}} \mathbf{e}_1^T = 0$ and $\hat{\mathbf{y}} \mathbf{e}_2^T = 0$). Both can be represented as,

$$\hat{\mathbf{y}} \mathbf{e}_1^T = \hat{\mathbf{y}} \mathbf{x}_1^T - \hat{a} \hat{\mathbf{y}} \hat{\mathbf{y}}^T - \hat{\mathbf{y}} \hat{\mathbf{r}}^T = 0 \quad (5.6a)$$

$$\hat{\mathbf{y}} \mathbf{e}_2^T = \hat{\mathbf{y}} \mathbf{x}_2^T - \hat{b} \hat{\mathbf{y}} \hat{\mathbf{y}}^T + \hat{\mathbf{y}} \hat{\mathbf{r}}^T = 0 \quad (5.6b)$$

and then simplified as

$$\mathbf{x}_1 = \hat{a} \hat{\mathbf{y}} + \hat{\mathbf{r}} \quad (5.7a)$$

$$\mathbf{x}_2 = \hat{b} \hat{\mathbf{y}} - \hat{\mathbf{r}} \quad (5.7b)$$

where they are actually the same equations as (5.5).

3. Third component: The decoded residual signal $\hat{\mathbf{r}}$ has to be orthogonal to both transposed error signals $\mathbf{e}^T$ (i.e., $\hat{\mathbf{r}}\mathbf{e}_1^T = 0$ and $\hat{\mathbf{r}}\mathbf{e}_2^T = 0$). Using the same method as in the previous component, these representations,

$$\hat{\mathbf{r}}\mathbf{e}_1^T = \hat{\mathbf{r}}\mathbf{x}_1^T - \hat{a}\hat{\mathbf{r}}\hat{\mathbf{y}}^T - \hat{\mathbf{r}}\hat{\mathbf{r}}^T = 0 \quad (5.8a)$$

$$\hat{\mathbf{r}}\mathbf{e}_2^T = \hat{\mathbf{r}}\mathbf{x}_2^T - \hat{b}\hat{\mathbf{r}}\hat{\mathbf{y}}^T + \hat{\mathbf{r}}\hat{\mathbf{r}}^T = 0 \quad (5.8b)$$

can therefore be simplified as

$$\hat{\mathbf{r}} = \mathbf{x}_1 - \hat{a}\hat{\mathbf{y}} \quad (5.9a)$$

$$\hat{\mathbf{r}} = \hat{b}\hat{\mathbf{y}} - \mathbf{x}_2 \quad (5.9b)$$

which are also similar to those obtained from the first and the second components.

From (5.5), further simplification can be made where the decoded downmix signals are represented as,

$$\hat{\mathbf{y}} = \frac{\mathbf{x}_1 + \mathbf{x}_2}{\hat{a} + \hat{b}} \quad (5.10)$$

which is a requirement that has to be met in order to obtain the minimum MSE. It means that the decoded downmix signal, $\hat{\mathbf{y}}$, has to be equal to the sum of both input signals, $\mathbf{x}_1 + \mathbf{x}_2$, divided by the sum of the estimated energy constant, $\hat{a} + \hat{b}$. Based on (5.5) and (5.10), assuming that the error due to the encoding process of the downmix signal is small and can be neglected, the downmix signal transmitted from the encoder will be identical to the decoded downmix signal. Thus the new downmix signal can be calculated as,

$$\mathbf{y}_{new} = \frac{\mathbf{x}_1 + \mathbf{x}_2}{\hat{a} + \hat{b}} \quad (5.11)$$

and this approximation will strongly affect the quality of the reconstructed audio signals.

Figure 5.2: Block diagram of the closed-loop R-OTT module.

Moreover, based on the new optimised downmix signal, (4.14) can be used to obtain the expression for the new optimised residual signal as below,

$$\mathbf{r}_{new} = \mathbf{x}_1 - \hat{\mathbf{a}}\mathbf{y}_{new} = \hat{\mathbf{b}}\mathbf{y}_{new} - \mathbf{x}_2 \quad (5.12)$$

where either $\mathbf{x}_1 - \hat{\mathbf{a}}\mathbf{y}_{new}$ or $\hat{\mathbf{b}}\mathbf{y}_{new} - \mathbf{x}_2$ can be used to determine $\mathbf{r}_{new}$. If both input signals have the exact same magnitude but opposite phases (i.e. $\mathbf{x}_1 = -\mathbf{x}_2$), then the downmix signal has all-zero values, $\mathbf{y}_{new} = 0$. Consequently, the residual signal can be determined as $\mathbf{r}_{new} = \mathbf{x}_1 = -\mathbf{x}_2$ and a specific information has to be transmitted to the decoder conveying this information.

Given the quantised spatial parameters, downmix and residual signal recalculations are performed using (5.11) and (5.12). The new optimised downmix and residual signals are then transmitted, replacing the original downmix and residual signals. The block diagram of this scheme is given in Fig. 5.2. Both input signals are fed to the R-OTT module. The spatial parameters are extracted. However, the downmix and residual signals are not computed, as they will not be transmitted but will be replaced by new, optimised signals. The R-OTT module is

Figure 5.3: The closed-loop R-OTT modules are employed within a multichannel audio encoder. Any type of audio encoder, highlighted as a dashed block, can be employed.

only used for the purpose of parameter extraction, so that in the implementation, the R-OTT module can be replaced with a simple CLD and ICC extractor. The OTT module is performed just to calculate the estimated energy constants. Furthermore, the downmix and residual signal recalculation block is supplied with the estimated energy constants to calculate the new optimised downmix signal using (5.11) and then compute the new optimised residual signal using (5.12).

5.3 The Closed-loop R-OTT Module in the MPS Architecture

Fig. 5.3 shows the implementation of the closed-loop R-OTT module in the MPS architecture where more than two channels (i.e., $L > 2$) of input and output exist. As in the MPS, each audio signal is decomposed by a A-QMF to obtain 71 hybrid band signals. Subsequently, multiple audio signals are processed by the MPS spatial analysis that is the tree of R-OTT modules. The spatial parameters are extracted, and downmix signals are created from multiple input signals. Moreover, residual signals can be calculated. In the original MPS scheme, the downmix and the residual signals are produced as shown in the figure. However, for this closed-loop implementation, all signals will be discarded and replaced by new, optimised downmix and residual signals.

The signal recalculation block is embedded for constructing a new, optimised downmix and new, optimised residual signals for each corresponding R-OTT module in the tree structure. Every new optimised intermediate downmix signal is fed back to the tree of R-OTT modules, replacing the original signal, which will be used by the corresponding R-OTT module in the next layer within the tree structure.

5.4 Results

Since the proposed analysis by synthesis system attempts to reduce errors by trying to make the output waveform similar to the input waveform, an objective performance metric, rather than a subjective one, has been found to be more suitable, particularly for comparing it to the conventional open-loop system. Due to its simplicity, Signal-to-Noise Ratio (SNR) measurement has been chosen.

In terms of multichannel audio coding, the SNR is defined as,

$$\text{SNR}_k = 10 \log_{10} \left[\frac{\sum_n (x_k[n])^2}{\sum_n (x_k[n] - \hat{x}_k[n])^2} \right], \quad (5.13)$$

where $x_k[n]$ and $\hat{x}_k[n]$ are the original and reconstructed audio signals, respectively for the k^{th} channel and for a frame length L where $n = 0, \dots, L - 1$, and $k \in \{1, 2\}$ and $k \in \{1, 2, 3, 4, 5\}$ for the 2-channel and the 5-channel encoders, respectively. The SNR values are then averaged for all frames of the channel k , which gives the segmental SNR, segSNR_k .

5.4.1 Experimental Setup

A number of low-correlated and high-correlated audio signals sampled at 44100 Hz have been provided for the experiments, called as Speech+ and Talk, respectively. For 5-channel input, the Speech+ (low-correlated signals) consisted of individual audio objects that were female and male speech, cello, trumpet and percussion music, while the Talk (high-correlated signals) consisted of the Left (L), Right (R), Center (C), Left surround (Ls), and Right surround (Rs) channels of 5.1 recordings containing panned mixtures of the individual audio objects. For 2-channel input, low-correlated signals consist of the female speech and cello while high-correlated signals use only L and R channels. Each of these signals was fed into a hybrid filter bank decomposing the signal into 71 hybrid bands. In calculating CLDs and ICCs 20 parameter bands were used. Segmentation and overlap-add windowing were performed in the subband domain. In this experiment, each audio segment consisted of 32 subband samples which is equivalent to the effective length of 1024 time domain samples.

Furthermore, for encoding the mono downmix signal, AAC was used at the bitrates of 64, 80, 96, 112, 128, 144 and 160 kb/s. Each residual signal was encoded with the bitrates starting from 16 to 160 kb/s. To grade the performance of the

audio coders employing the closed-loop R-OTT module, an audio coder applying the open-loop R-OTT module has also been implemented.

Two other configurations of audio codecs, used in the experiments for investigating the performance of the proposed closed-loop R-OTT module, can be explained as below

- ContSP R-OTT. This is an open-loop R-OTT module transmitting unquantised spatial parameters intended to determine how much the close-loop R-OTT module reduces the error introduced by the quantisation and coding processes.
- Unquantised R-OTT. This is another open-loop R-OTT module transmitting unquantised downmix signals, residual signals and spatial parameters.

5.4.2 SNR Comparison for Open-loop and Closed-loop Systems

Table 5.1: Average SegSNR (dB) of 2 Channel Audio Coders

Type of Input	Audio Channel	Open-loop R-OTT	Closed-loop R-OTT	ContSP R-OTT
Low-correlated signals	1	24.52	30.92	31.08
	2	33.99	37.23	37.35
Average		29.26	34.08	34.22
High-correlated signals	L	28.83	31.55	35.66
	R	24.21	29.50	34.18
Average		26.52	30.52	34.92
Overall average		27.89	32.30	34.57

The experiments in this section are aimed at demonstrating that the closed-loop algorithm is able to improve the segSNR achieved by the R-OTT module. Additionally, the experiments using low-correlated audio signals as inputs are intended to investigate the effect of the correlation of input signals on the segSNR of the reconstructed audio signals. In this experiment, full waveform reconstruction was performed by transmitting all residual signals from each R-OTT encoder at 160 kb/s.

Table 5.1 shows the results for a 2-channel audio codec for low-correlated and high-correlated inputs. The average segSNR measured on the closed-loop R-OTT is 32.30 dB, which is 4.41 dB higher than the conventional open-loop R-OTT. Moreover, Table 5.2 shows the results for the 5-channel coder. The overall average segSNR for the closed-loop R-OTT is 30.49 dB, which is 6.13 dB higher than that of the conventional open-loop R-OTT. It clearly indicates that the closed-loop algorithm applied on R-OTT module can minimise errors introduced by the quantisation process of the spatial parameters at a cost of additional parameter recalculation in the encoder side.

Moreover, the results also show that the average segSNRs achieved by all the tested audio codecs when encoding 5-channel audio signals are lower than when encoding 2-channel audio signals. In particular, this is clearly shown by the segSNRs of the audio codecs encoding the low-correlated audio signals. For instance, Table 5.2 shows that the average segSNR achieved by the closed-loop R-OTT when low-correlated signals are used is 29.85 dB, which is lower than 34.08 dB; which was achieved by the 2-channel closed-loop R-OTT for low-correlated signals as shown in Table 5.1. It indicates that structuring the R-OTT modules in a tree scheme may introduce further signal distortion particularly when low-correlated audio signals are used as inputs.

In order to evaluate the maximum segSNR improvement that can be achieved in

Table 5.2: Average SegSNR (dB) of 5 Channel Audio Coders

Type of Input	Audio Channel	Open-loop R-OTT	Closed-loop R-OTT	ContSP R-OTT
Low-correlated signals	1	21.86	26.47	26.92
	2	21.72	26.21	26.56
	3	27.19	31.72	32.19
	4	28.61	32.05	32.32
	5	22.53	32.83	33.39
Average		24.38	29.85	30.27
High-correlated signals	L	26.38	32.06	35.31
	R	25.49	31.27	34.62
	C	20.64	28.63	33.12
	Ls	22.25	31.98	34.20
	Rs	26.92	31.70	34.32
Average		24.34	31.13	34.31
Overall average		24.36	30.49	32.29

the closed-loop R-OTT module, the segSNR measured for the ContSP R-OTT should be observed. The results in both Table 5.1 and 5.2 show that the segSNRs of the ContSP R-OTT are higher than the segSNRs of the audio codec applying the closed-loop R-OTT module, especially when high-correlated signals are used as inputs. As can be seen in Table 5.1, when a 2-channel audio coder is used, the average segSNR measured on the ConstSP R-OTT for high-correlated signals is 34.92 dB, which is much higher than the 30.52 dB measured on the closed-loop R-OTT. The same case applied when a 5-channel audio coder was used. As shown in Table 5.2, the segSNR measured for 5-channel ContSP R-OTT for high-correlated signals is 34.31 dB, which is 4.04 dB higher than that achieved by the closed-loop R-OTT. It indicates that even though the proposed closed-loop R-OTT module

Figure 5.4: The segSNR per frame

is capable of minimising signal distortion, the segSNR improvement is not as high as expected. It suggests that performing further investigation is necessary to discover why the closed-loop R-OTT module cannot achieve the segSNR as high as the segSNR of the ConSP R-OTT scheme.

For a closer look into the SNR improvement, the segSNR for several audio frames for one of the channels of the low-correlated signals has been plotted in Fig. 5.4. It shows the segSNRs of the conventional open-loop R-OTT and the closed-loop R-OTT modules at the same operating bitrate. For comparison, the segSNR of the unquantised R-OTT was shown, for which all of the signals and parameters are unquantised. This figure shows that significant SNR improvement can be achieved with the closed-loop algorithm, although there is still a high margin before reaching the maximum segSNR that is supposed to be achieved, as shown by the unquantised R-OTT.

5.4.3 Performance Comparison for Various Bitrates

In this section, the performance of the 5-channel closed-loop R-OTT audio coder is evaluated for various bitrates. In these experiments, lower bitrates were achieved by limiting the bandwidth of the transmitted residual signal. The spatial parameter resolution was kept constant on 20 parameter bands, and the time resolution was also fixed at the effective frame length of 1024 samples. At the decoder side, the decorrelator was not performed as it provides a synthetic residual signal particularly when all residual signals are not transmitted from the encoder side.

The results are given in Fig. 5.5. The segSNR in dB is plotted against the bitrate per audio channel in kb/s. The result of this experiment shows that the proposed closed-loop R-OTT module outperforms the open-loop system for all bitrates. The highest SNR improvement is achieved when all the residual signals are transmitted. However, it decreases as the bitrate allocated for the residual signal becomes lower. The test results suggest that the proposed closed-loop R-OTT module, operating at bitrates between 40 and 160 kb/s per audio channel, can improve the segSNR even though the highest improvement is achieved at 160 kb/s per audio channel.

5.5 Conclusions

A framework for spatial audio coding based on the analysis by synthesis technique has been presented in this chapter. It is applicable to MPEG Surround. An adapted version of the R-OTT module is introduced. Comparisons were made with the conventional open-loop MPEG Surround system using objective metrics. The results show that the proposed closed-loop R-OTT module can improve 5.66 dB of the average segSNR of the reconstructed audio signals at a cost of additional

Figure 5.5: Average SegSNRs of 5-channel audio coders are plotted against the bitrates. At a bitrate of 160 kb/s per audio channel, the improvement achieved by the closed-loop R-OTT module is 5.66 dB

parameter recalculation in the encoder side. However, the closed-loop R-OTT module implemented in the subband domain cannot achieve as high segSNR as expected. It is suggested that further investigation is necessary to discover why the segSNR improvement is not as high as expected and then to propose a solution for further improving the segSNR.

Chapter 6

MDCT-based R-OTT Module

The previous chapter has presented the description of the new closed-loop R-OTT module that is implemented in the context of MPS architecture. It has been shown that significant error reduction was achieved by applying this module. However, there is an indication that the closed-loop R-OTT module cannot reduce the error introduced by the spatial parameter quantisers as much as expected. In this chapter the disadvantages of employing the closed-loop R-OTT module in the subband domain of QMF filterbank is discussed, leading to the proposal for applying the frequency domain closed-loop R-OTT module.

The beginning part of this chapter discusses the disadvantage of applying the closed-loop R-OTT module in the subband domain of QMF filterbank, so that it is motivated to propose the MDCT-based implementation of the closed-loop R-OTT module. The adaptation of the R-OTT module for implementation in the frequency domain is described, as is the method needed to determine the parameter bands. At the end of this chapter the results of the experiments are given, supporting the proposal. The results show that applying the closed-loop R-OTT module in the frequency domain is capable of minimising further the

distortion of the reconstructed audio signals.

6.1 Motivation for the MDCT-Based Implementation

The proposed closed-loop R-OTT algorithm relies on an approximation in (5.11) when recalculating the downmix signal as a new, optimised signal, $\mathbf{y}_{new}$. Therefore, the signal distortion reduction process is based on how to create the new optimised downmix signal on the encoder side, such that the synthesised downmix signal on the decoder side fulfils the criteria in (5.10). In practice, this is achieved by ensuring that the approximation error, which is the difference between the synthesised downmix signal, $\hat{\mathbf{y}}$, and the new optimised downmix signal, $\mathbf{y}_{new}$, is minimised. In order to obtain the minimum approximation error, both the synthesised and the approximated signals should be synchronised and compared.

The closed-loop R-OTT method is ideally capable of considerably minimise the error introduced by the quantisation process of the spatial parameters. This is because the new optimised downmix and residual signals, $\mathbf{y}_{new}$ and $\mathbf{r}_{new}$, are computed based on estimated energy constants, $\hat{a}$ and $\hat{b}$, so that the quantisation errors of CLD and ICC are now compensated for through the newly optimised signals. Consequently, the quantisation errors of CLD and ICC no longer affect the overall distortion of the synthesised audio signals.

However, the operation of the closed-loop R-OTT method in the subband domain of the QMF filterbank is not always appropriate for applying the proposed closed-loop R-OTT module. The QMF transform has a disadvantage when applied to the closed-loop R-OTT module, because the impulse responses of the analysis filterbank, $G_{0,s}$, are different from the impulse responses of the synthe-

sis filterbank, $H_{0,s}$, leading to a dissimilarity in the length of delay introduced. The analysis filterbank introduces a delay of 704 samples, which is equivalent to 11 samples in the QMF subband domain. On the other hand, the synthesis filterbank introduces a delay of 257 samples, which is equal to 4.015625 samples in the QMF subband domain [130]. As a result of fractional sample delay of the 0.015625 sample, the synthesised signals cannot perfectly match the target signals, leading to a less than perfect comparison.

In order to compensate for such delay, the target signals can be delayed by a fractional delay filter [133] making both synthesised and target signals become perfectly synchronised. However, in this work, it is proposed to use another time-frequency transformation tool for employing the closed-loop R-OTT module. Although there are various signal transformation tools available that may be suitable, it is proposed to apply the closed-loop R-OTT module in the MDCT domain. In addition to its ability to meet the requirement in (5.11) and (5.12), applying the MDCT-based R-OTT module avoids the need for the transformation of the residual signals from the subband domain of filterbank to spectral coefficients for the purpose of quantisation, as in the case of MPS. Thus, it can simplify the structure of the encoder. Moreover, the MDCT transform also provides a perfect signal reconstruction [134, 135].

Please note that applying the open-loop R-OTT module in the MDCT domain possibly leads to strong artifacts in the reconstructed audio signals. It is only the closed-loop version of the R-OTT module that is proposed for implementation in the MDCT domain. It is shown in the result section that the closed-loop R-OTT module is capable of providing a perfect signal reconstruction, even though the spectral coefficients are scaled in the downmixing process, and then the quantised spatial parameters are transmitted, provided that the downmix and residual signals are not encoded.

6.2 Determining the Parameter Bands

In order to provide a scalable spatial resolution, the MDCT-based closed-loop R-OTT module can be applied using a number of parameter bands which is proposed, for simplicity, to be equal to the number of parameter bands of the MPS. The purpose is to make the frequency resolution of each parameter band in the frequency domain implementation equal to the implementation in the subband domain. In the subband domain implementation, as applied in the MPS, a single or a number of hybrid bands can be simply grouped as a single parameter band. However, grouping the spectral coefficients in the frequency domain implementation of the R-OTT module is not as easy as grouping the hybrid bands in the subband domain implementation. It is necessary to take into consideration the quantisation process of the spectral coefficients as previously explained (see chapter 4 that briefly discusses the MPEG AAC and MPS).

In performing quantisation, the spectral coefficients are grouped into 49 scale factor bands. It is therefore more effective to group a single or a number of spectral bands as a parameter band, instead of directly grouping the spectral coefficients so that there is no spectral band that will be divided into different parameter band. In order to take this into consideration, the parameter bands for performing spatial analysis and synthesis in the frequency domain can be determined by mapping the 71 hybrid bands of the QMF filterbank to the 49 scale factor bands of the spectral coefficients. Depending on the number of the parameter bands, various methods of mapping the hybrid bands to the scale factor bands can be applied.

For instance, such a mapping for grouping 49 Scale Factor Bands (SFB) to 20 Parameter Bands (PB) is given in Table 6.1. For this mapping the frequency resolution of every scale factor band (see Table 4.1) and every parameter band used in the MPS (see Table 4.3) are firstly calculated. Then, all scale factor bands

Table 6.1: Mapping of 49 Scale Factor Bands to 20 Parameter Bands

PB Index	SFB Index	PB Index	SFB Index
1	1	11	16-17
2	2	12	18-19
3	3	13	20-21
4	4	14	22-23
5	5-6	15	24-25
6	7-8	16	26-27
7	9-10	17	28-29
8	11	18	30-31
9	12-13	19	32-37
10	14-15	20	38-49

are grouped into 20 parameter bands in such a way that the frequency resolution of each parameter band is as close as possible to the frequency resolution of each corresponding parameter band in Table 4.3.

6.3 MDCT-Based R-OTT Module for High Bitrate Implementation

Fig. 6.1 shows a block diagram of an L-channel audio encoder using the MDCT-based closed-loop R-OTT module. When compared to the corresponding subband domain structure, it can be seen that all FB blocks are replaced with MDCT blocks. In addition, the trans-coding process required to transform the subband domain residual signals into spectral coefficients has been removed. Instead, spectral quantisers that are actually AAC quantisers are performed. In case the AAC is used for encoding the downmix signal, the spectral coefficients of

Figure 6.1: The block diagram of L-channel MDCT-based closed-loop R-OTT.

the downmix signal are directly fed to the spectral quantiser. This is the same implementation case when MPEG Surround is combined with HE-AAC, where the subband domain signals may be fed directly to the spectral band replication (SBR) by bypassing the synthesis and analysis filterbank conversion.

6.4 Results

In order to show further improvement achieved by applying the MDCT-based R-OTT module the Talk audio excerpt (high-correlated audio signals), which were also used in chapter 5, were used as inputs for the experiments. However, the Speech+ (low-correlated audio signals) were not tested because the improvement achieved by the QMF-based closed-loop R-OTT module, has been very close to the expected segSNR (see Table 5.1 and Table 5.2). As discussed in Chapter 5, the

Talk audio excerpt consisted of the Left (L), Right (R), Centre (C), Left surround (Ls), and Right surround (Rs) channels of 5.1 recordings containing panned mixtures of the individual audio objects. For 2-channel input, the left (L) and right (R) channels were used as inputs. In calculating CLDs and ICCs, 20 parameter bands were used in both subband and frequency domain implementation.

6.4.1 Evaluation of the MDCT-based Open-loop R-OTT Module

Before investigating the proposed MDCT-based closed-loop R-OTT module, it is good to have a better understanding of the implementation of the open-loop R-OTT module in the MDCT domain. Therefore, the performance of the MDCT-based open-loop R-OTT module, in terms of segSNR, is evaluated in this section. The experiments were conducted by operating a number of audio codecs, employing the MDCT-based open-loop R-OTT module, using different quantisers. All audio codecs encoding 2-channel audio signals operated at various bitrates ranging from 50 to 150 kb/s per audio channel. The CLDs and ICCs were quantised uniformly. As many as 5 different quantisation tables, each with its own stepsize, were tested. For benchmarking, an audio codec employing the subband domain open-loop R-OTT module (i.e. as implemented in MPS), operating at the same range of bitrates, was also tested. In addition the MDCT-based R-OTT module was also tested using the MPS's spatial quantiser. The average segSNRs measured on the experiments are plotted in Fig. 6.2.

The results show that the audio codecs applying the MDCT-based open-loop R-OTT module quantising and representing CLD and ICC with 7, 8, 10 and 12 bits achieve higher segSNRs than the audio codec employing the subband domain open-loop R-OTT module at almost all operating bitrates. On the other hand, the

Figure 6.2: SegSNR of MDCT-based R-OTT module

audio codecs applying the MDCT-based open-loop R-OTT module, quantising each CLD and ICC using 6-bit uniform quantiser, as well as the MPS' spatial quantiser, have lower segSNRs than the 2-channel MPS at all operating bitrates. The results clearly indicate that the implementation of the MDCT-based open-loop R-OTT module at the tested bitrates can achieve a higher performance than the MPS, provided that appropriate spatial quantisers are applied. It suggests that the spatial quantisers play an important role in enhancing the performance of the audio codec applying the open-loop R-OTT module.

Furthermore, it is interesting to note that the implementation of the MDCT-based R-OTT module employing MPS's spatial quantisers cannot achieve better performance than the MPS itself. It clearly indicates that the MPS's spatial quantisers are not appropriate for the implementation of the MDCT-based open-loop R-OTT module. It suggests that for applying the R-OTT module in the frequency domain, an investigation should be conducted for implementation over

various spatial quantisers. Then proper spatial quantisers should be chosen for a particular range of operating bitrates in order to achieve the optimal performance of the MDCT-based R-OTT module.

It can also be seen that at a bitrate of 150 kb/s per audio channel, approximately 9 dB of SNR improvement, compared with MPS, is achieved by the codec employing the MDCT-based R-OTT module using 12-bit uniform spatial quantiser. Conversely, a slightly different performance is shown by the audio codec when applying a 10-bit uniform spatial quantiser. However, the segSNR improvement decreases as the bitrates become lower. It also shows that more than 3 dB segSNR improvement is achieved by the audio codecs using both 10 and 12-bit uniform spatial quantisers operating at bitrates between 80 and 150 kb/s per audio channel. Furthermore, the segSNR improvement continues to decrease and becomes insignificant (i.e. improvement is around 1 dB) at lower bitrates, particularly at around 65 kb/s per audio channel. In contrast, at operating bitrates lower than 60 kb/s per audio channel, the MPS achieves the highest segSNR among all tested audio codecs.

The results indicate that the MDCT-based open-loop R-OTT module is not appropriate for low bitrate implementation below 60 kb/s per audio channel. It suggests that the performance of the R-OTT module applied in the frequency domain is subject to the operating bitrate. Moreover, it also suggests that the open-loop R-OTT module performs better in the subband domain than in the frequency domain, when the bitrate is lower than 60 kb/s per audio channel.

In addition, Fig. 6.2 also shows a boundary of the segNSRs that can be achieved by the audio codec applying any number of bits applied in the uniform spatial quantiser to represent the CLD and ICC. This boundary may be useful to grade the performance of the proposed MDCT-based closed-loop R-OTT module.

Figure 6.3: SegSNR of MDCT-based closed-loop R-OTT

6.4.2 Comparison of the Open and Closed-loop R-OTT Modules over Various Spatial Quantisers

As an early step in investigating the operation of the MDCT-based closed-loop R-OTT module, the experiments were conducted by measuring the segSNR for different quantisation errors. The audio codecs employing the MDCT-based closed-loop R-OTT module using 3 different uniform spatial quantisers representing both the CLDs and ICCs with 6, 7, and 8 bits were tested. For comparison, the codecs applying the MDCT-based open-loop R-OTT using similar spatial quantisers were also included in the experiments. The results were plotted and depicted in Fig. 6.3. It shows that the closed-loop technique significantly improves the average segSNR for every type of the tested quantisers. As expected, the results demonstrate that the performances of the closed-loop R-OTT modules using different spatial quantisers are similar. As previously explained, the

Figure 6.4: SegSNR of MDCT-based closed-loop R-OTT module

closed-loop R-OTT module operates by minimising the errors contributed during the quantisation of the spatial parameters. Hence, the stepsize of the spatial quantiser, which leads to the amount of the quantisation error, does not affect the performance of the closed-loop R-OTT module. It indicates that the closed-loop R-OTT module is capable of greatly minimise the error introduced by the spatial quantiser regardless of the amount of error introduced. It therefore suggests that the best spatial quantiser, in terms of the smallest number of bits required, among the tested quantisers to be applied in the MDCT-based R-OTT module, is the MPS's spatial quantisers.

6.4.3 Evaluation of the Closed-loop R-OTT Module Using the MPS's Spatial Quantisers

In order to investigate further the performance of the MDCT-based closed-loop R-OTT module using the MPS's spatial quantisers, experiments have also been conducted for encoding 2-channel audio signals over various operating bitrates between 50 and 150 kb/s per audio channel. The results of the experiments are given in Fig. 6.4. For comparison, the upper bounds of the segSNRs achieved by the MDCT-based open-loop R-OTT module, termed as boundary, are plotted. Moreover, the segSNRs of the FB-based open and closed-loop R-OTT modules are also included. The results demonstrate that the closed-loop R-OTT module can improve segSNR in the subband and even further in the frequency domain. Moreover, it is also shown that the MDCT-based closed-loop R-OTT module can exceed the upper bound of the segSNRs that can be achieved by the MDCT-based open-loop R-OTT module. It indicates that the closed-loop R-OTT module applying the MPS's quantisers consistently has a higher segSNR than the open-loop R-OTT module, regardless of the spatial quantiser applied.

6.4.4 Evaluation of the Closed-loop R-OTT Module Encoding 2-channel Audio Signals at High Bitrate

Table 6.2: Average SegSNR (dB) of Various R-OTT Methods

Audio Channel	FB-Based OL R-OTT	FB-Based CL R-OTT	MDCT-Based CL R-OTT
Left	28.83	31.55	38.48
Right	24.21	29.50	36.64
Average	26.52	30.52	37.56

The experiments in this section aim to demonstrate the ability of the proposed MDCT-based closed-loop R-OTT module to improve performance, in terms of segSNR. The non-uniform MPS quantisers were used in the quantisation of the spatial parameters. Table 6.2 shows the measured segSNRs of both the filterbank (FB)-based and the MDCT-based closed-loop R-OTT modules operating at 160 kb/s per audio channel. For a more comprehensive comparison and result analysis, the segSNRs achieved by the FB-based open-loop R-OTT module are also presented.

The results in Table 6.2 demonstrate that the proposed closed-loop R-OTT module is capable of improving the segSNR. It shows that the FB-based closed-loop R-OTT can improve the segSNR by approximately 4 dB. Moreover, performing the MDCT-based closed-loop R-OTT module is able to improve further the segSNR. About 11 dB of segSNR improvement is achieved in comparison to the FB-based open-loop R-OTT. Even though the experiments were performed only for a single block of R-OTT module (not performed in a tree structure) it is strongly suggested that significant segSNR improvement would also be achieved for the implementation with larger number of input channels when a tree of R-OTT modules is applied. As an example, the next section discusses the experiments for encoding 5-channel audio signals.

6.4.5 Evaluation of the Closed-loop R-OTT Module Encoding 5-channel Audio Signals at High Bitrate

The evaluation of the closed-loop R-OTT module in this section aims to demonstrate that the closed-loop R-OTT algorithm is also able to improve the segSNR when encoding 5-channel audio signals. In addition the evaluation is also intended to show that the closed-loop R-OTT module is also capable of distributing the

Table 6.3: Average segSNRs (in decibels) of MDCT-Based Codecs Compared to FB-Based R-OTT

Audio Channel	FB-Based OL R-OTT	FB-Based CL R-OTT	MDCT-Based CL R-OTT
Left	26.38	32.06	38.00
Right	25.49	31.27	38.02
Centre	20.64	28.63	36.65
Left surround	22.25	31.98	38.15
Right surround	26.92	31.70	38.27
Mean	24.34	31.13	37.82
STD	2.75	1.43	0.66

error among the audio channels. The audio codecs operate at 160 kb/s per audio channel. Table 6.3 shows the results of the comparison between average segSNRs in multichannel audio encoders employing the open-loop and closed-loop R-OTT modules in both FB-based and MDCT-based scenarios.

It is shown in Table 6.3 that the MDCT-based closed-loop R-OTT is able to reach the average segSNR of 37.82 dB. Approximately 13 dB of SNR improvement is obtained, compared to the conventional FB-based open-loop R-OTT module. Benchmarking to the FB-based closed-loop R-OTT module shows that the segSNR increases by more than 6 dB. The results indicate that the closed-loop R-OTT algorithm is also capable of minimising signal distortion at all channels of 5-channel signals resulting in segSNR improvement.

Interesting results were produced in situations where the segSNR achieved by the MDCT-based closed-loop R-OTT in each channel had similar values. The average segSNR is 37.82 dB, while the standard deviation is equal to 0.66 dB. This standard deviation is significantly lower than 2.75 dB which is the standard deviation of the segSNR measured on the FB-based open-loop R-OTT module.

It is also considerably lower than 1.43 dB: the standard deviation of the segSNR achieved by the FB-based closed-loop R-OTT module. The results indicate that the encoder performing the MDCT-based closed-loop R-OTT module does not only significantly improve the segSNR but also makes the segSNR of audio signals very close to each other. It suggests that the coder is able to distribute the distortion uniformly across the channels.

6.5 Conclusions

The MDCT-based closed-loop R-OTT module has been proposed in this chapter. The reason why the QMF filterbank is not suitable for this closed-loop R-OTT module has been explained at the beginning of this chapter. Furthermore, the adaptation that required for implementation of the R-OTT module in the MDCT domain has also been described including the way to determine the parameter band as well as the possibility to simplify the coding structure. Significant segSNR improvement has been demonstrated in the result section. Moreover, the results also indicate that the proposed closed-loop R-OTT module can also distribute the error across the audio channels.

Chapter 7

Simplified AbS Algorithm

In this chapter the implementation of the analysis by synthesis (AbS) optimisation procedure in a tree of R-OTT modules to find the optimal signals and parameters is discussed. The simplest case of the AbS optimisation procedure where a single R-OTT module is applied for encoding 2-channel audio signals is presented first, followed by a more complex implementation when a multichannel scenario is performed. Furthermore, it is explained that the main challenge of adopting this procedure for a single or even a larger number of R-OTT modules is the huge number of possible parameter combinations leading to impractical implementation.

A simplified AbS algorithm, proposed to address the practicality problem, is then discussed. This algorithm is developed based on the MDCT-based closed-loop R-OTT module presented in the previous chapter. Three steps of simplifications are applied to make the optimisation procedure more straightforward. Additionally, the simplified AbS algorithm can be performed with various complexity levels which means that it can be adjusted for different hardware specification. The experimental result shows that the proposed simplified AbS algorithm with various

complexity levels can improve further the performance of the R-OTT module. Moreover, experiments at various operating bitrates also shows that the proposed simplified AbS algorithm with a reasonable complexity level significantly improves the segSNR.

7.1 Full Search AbS Optimisation

Referring to the proposed AbS-SAC framework for spatial audio coding as given in Fig. 5.1, a case of the simplest AbS implementation to encode 2-channel audio signals can be illustrated. In order to find the optimal downmix and residual signals, as well as the optimal parameters, a full search AbS optimisation should be applied. An OTT module is used as a model for reconstructing 2 channels of audio signals. As the full searching procedure is performed, there is no need for applying the spatial analysis block.

An AbS optimisation procedure can be carried out in such a way that the inputs of the optimisation procedure are the quantisation values of the downmix and residual signals, as well as the spatial parameters. All of these inputs can be varied to reconstruct various forms of audio signals. The purpose of the AbS optimisation procedure is to examine all possible outcomes obtained by combining all inputs in every possible way i.e. all possible combinations of every variable. For each combination, the OTT module reconstructs audio signals and the error minimisation block then computes signal distortion. Any combination that obtains minimum error is chosen as the optimal one.

The quantisation values of the spectral coefficient quantiser of the downmix and residual signals that are becoming the inputs of the AbS optimisation procedure range from -8191 to $+8191$, meaning that there are 16383 quantisation values for each spectral coefficient of the downmix signals. On the other hand, 31 and 8

quantisation values of the CLD and ICC, respectively, are also available, provided that the MPS's quantisers are used. As a result of combining all those quantisation values, for each index of the spectral coefficient the number of available combinations can be computed as $16383 \times 16383 \times 31 \times 8 = 6.6564 \times 10^{10}$. Note that for simplifying the calculation, the spatial parameters are assumed to be calculated for every spectral coefficient. The scale factor band is ignored.

7.2 Extended Structure for Multichannel Audio Signals

In order to illustrate the AbS optimisation procedure for a larger number of audio channels, it is useful to extend the tree structure of the R-OTT and OTT modules in the encoder and decoder respectively, comprising a number of layers with a number of R-OTT and OTT modules in each layer. The extended encoder structure is given in Fig. 7.1

Let L denote the number of input channels, P the number of R-OTT modules at the first layer of the tree scheme, and Q the number of layers in the tree scheme. For simplicity, the tree structure is assumed to be symmetric which means that $L = 2^Q = 2P$. Using (4.12), the intermediate downmix signal, $y_{p,q}[k]$, where p ($p = 0, 1, \dots, P - 1$) is the index of R-OTT at a particular layer, q ($q = 0, 1, \dots, Q - 1$) being the index of layer in the tree scheme, can be written as,

$$y_{p,q}[k] = \frac{y_{2p,q+1}[k] + y_{2p+1,q+1}[k]}{a_{p,q} + b_{p,q}} \quad (7.1)$$

where $y_{2p,q+1}[k]$ and $y_{2p+1,q+1}[k]$ are the intermediate downmix signals for a particular p and q . For $q = Q - 1$, the intermediate downmix signals are calculated from the input signals as,

$$y_{p,Q-1}[k] = \frac{x_{2p}[k] + x_{2p+1}[k]}{a_{p,Q-1} + b_{p,Q-1}} \quad (7.2)$$

Figure 7.1: A tree of R-OTT modules for encoding L-channel audio signals.

At the decoder side, the audio signals can be reconstructed using (4.17) as,

$$\hat{x}_l[k] = \hat{a}_{p,Q-1}\hat{y}_{p,Q-1}[k] + \hat{r}_{p,Q-1}[k] \quad (7.3a)$$

$$\hat{x}_{l+1}[k] = \hat{b}_{p,Q-1}\hat{y}_{p,Q-1}[k] - \hat{r}_{p,Q-1}[k] \quad (7.3b)$$

where l ($l = 2p + 1 = 1, 2, \dots, L$) is the channel index and $\hat{y}_{p,Q-1}$ is the synthesised intermediate downmix signal at layer of $Q - 1$. For an arbitrary OTT module, the estimated intermediate downmix signal can be represented as,

$$\hat{y}_{2p,q+1}[k] = \hat{a}_{p,q}\hat{y}_{p,q}[k] + \hat{r}_{p,q}[k] \quad (7.4a)$$

$$\hat{y}_{2p+1,q+1}[k] = \hat{b}_{p,q}\hat{y}_{p,q}[k] - \hat{r}_{p,q}[k] \quad (7.4b)$$

and then, for $p = 1$ and $q = 1$, (7.4) can be written as,

$$\hat{y}_{0,1}[k] = \hat{a}_{0,0}\hat{y}_{0,0}[k] + \hat{r}_{0,0}[k] \quad (7.5a)$$

$$\hat{y}_{1,1}[k] = \hat{b}_{0,0}\hat{y}_{0,0}[k] - \hat{r}_{0,0}[k] \quad (7.5b)$$

where $\hat{y}_{0,0}[k]$ is the synthesised downmix signal.

Since the encoding of L -channel audio signals (for $L > 2$), where L denotes the maximum number of input channels, incorporates $L - 1$ R-OTT modules, it is possible to achieve better error reduction by applying the AbS optimisation procedure to the whole structure of the encoder optimising all signals and parameters at all R-OTT modules. Thus, the overall performance of the coder can be improved. This is the basic idea used for encoding L -channel audio signals.

The error signals that are defined as the differences between the original audio signals $x_l[n]$ and the synthesised signals $\hat{x}_l[n]$, can be written in the vector form as,

$$\mathbf{e}_l = \mathbf{x}_l - \hat{\mathbf{x}}_l \quad (7.6)$$

where $l = 1, 2, \dots, L$ is the index of audio channel with a maximum value of L . Assuming that the transmission channel is lossless, the error for channel L can

be written by taking $l = L$ as,

$$\mathbf{e}_L = \mathbf{x}_L - \hat{b}_{P,Q} \hat{\mathbf{y}}_{P,Q} + \hat{\mathbf{r}}_{P,Q} \quad (7.7)$$

by substituting the synthesised audio signals in (7.3). Furthermore, this error can be represented as follows:

$$\begin{aligned} \mathbf{e}_L = & \mathbf{x}_L - \hat{b}_{P,Q} \hat{b}_{P/2,Q-1}, \dots, \hat{b}_{4,3} \hat{b}_{2,2} \hat{b}_{1,1} \hat{\mathbf{y}}_{1,1} \\ & - \hat{b}_{P,Q} \hat{b}_{P/2,Q-1}, \dots, \hat{b}_{4,3} \hat{b}_{2,2} \hat{\mathbf{r}}_{1,1} \\ & - \hat{b}_{P,Q} \hat{b}_{P/2,Q-1}, \dots, \hat{b}_{4,3} \hat{\mathbf{r}}_{2,2} \\ & \vdots \\ & - \hat{b}_{P/2,Q-1} \hat{\mathbf{r}}_{P/2,Q-1} \\ & - \hat{\mathbf{r}}_{P,Q} \end{aligned} \quad (7.8)$$

by including all the corresponding synthesised intermediate downmix signals derived from (7.4) as well as the corresponding estimated energy constants, $\hat{a}_{p,q}, \hat{b}_{p,q}$, for each layer (i.e. from layer $q = 2$ up to $q = Q - 1$) and the synthesised downmix signals, $\hat{\mathbf{y}}_{1,1}$, in the expression by substituting $\hat{\mathbf{y}}_{P,Q}$. This expression can be generalised for all the channels that contribute to the quality of the reconstructed audio signal on channel l by jointly selecting the optimal downmix signal, a set of optimal spatial parameters, and a set of optimal residual signals.

As in the case of encoding 2-channel audio signals, the AbS optimisation is performed to find the optimal signals and parameters from all possible combination. For each combination, the tree of OTT modules reconstruct the audio signals and the error minimisation block then computes the signal distortion. Any combination that obtains minimum error is chosen as the optimal one. However, the number of possible combinations significantly increases, because there are as many as Q channels of residual signals for each input audio as shown in (7.8). Furthermore 31 and 8 quantisation values of the CLD and ICC, respectively, are available for each OTT module, provided that the MPS's quantisers are used.

Note that there are as many as Q OTT modules involved in reconstructing each audio channel.

As a result of combining all the quantisation values of spectral coefficients and spatial parameters, a huge number of possible combinations exist. For instance, for each index of the spectral coefficient when encoding 5-channel audio signals, a total of 4 OTT modules are employed. Thus, the number of available combinations can be computed as $16383^4 \times 31^4 \times 8^4 = 2.7251 \times 10^{26}$. Consequently, a total of 2.7251×10^{26} loops must be performed to select optimal coefficients and parameters making the encoding an extremely complex task, if not an impossible one.

7.3 Simplified AbS Optimisation

7.3.1 An Approach for Algorithm Simplification

A simplified trial and error procedure can be applied in order to find sub-optimal signals and parameters as a solution for the impractical implementation requirements of the full search AbS procedure. Three steps of simplifications are applied to make the algorithm simple. First, the number of parameters and spectral coefficients involved in the searching procedure are significantly reduced. For instance, rather than finding an optimal spectral coefficient from all quantisation values, a sub-optimal coefficient is simply chosen from a limited number of quantisation values which are assigned based on decoded spectral coefficients. Considering the trade-off between the complexity and the degree of sub-optimality, the number of coefficients and parameters involved in the searching procedure can be made variable.

Second, the main AbS-SAC algorithm is performed as a sequential process in that

the sub-optimal signals and parameters are not selected at the same time but one after the other. The sub-optimal spectral coefficients of the downmix and residual signals can be determined first. Once the sub-optimal downmix and residual signals are found, the sub-optimal CLDs and ICCs can be selected. Alternatively, sub-optimal spatial parameters are selected first followed by choosing sub-optimal spectral coefficients. Performing the searching algorithm in a sequential process will significantly reduce the number of possible combinations to be examined.

Finally, the sequential process is performed iteratively until an insignificant error reduction is achieved. The reason for performing the iteration process is that the sub-optimal spatial parameters are found based on the selected sub-optimal downmix and residual signals. Additionally, the sub-optimal spectral coefficients of the downmix and residual signals are selected based on the chosen sub-optimal spatial parameters. Consequently, it is possible to re-optimize the downmix and residual signals after determining sub-optimal spatial parameters. In contrast, it is also possible to re-select new sub-optimal spatial parameters once sub-optimal downmix and residual signals are found. It is expected that undertaking the iteration process will gradually reduce signal distortion.

The limited number of quantisation values, where the sub-optimal spectral coefficient is selected from, can be determined as follows: The spectral coefficients decoded by the spectral decoder become the inputs to the algorithm. A number of quantisation values defined as the candidates for sub-optimal spectral coefficient named as predetermined vector, $ix_p[k]$, can then be assigned based on the decoded spectral coefficients, $ix[k]$, as

$$ix_p[k] = [ix[k] - v, ix[k] - v + 1, \dots, ix[k] - 1, ix[k], ix[k] + 1, \dots, ix[k] + v - 1, ix[k] + v] \quad (7.9)$$

where $ix[k]$ is the decoded spectral coefficient as in (4.3), k is the index of spectral coefficient, and $2v + 1$ is the size of the predetermined vector, $ix_p[k]$, with

v an integer number reflecting the computational complexity of the searching procedure.

A limited number of quantisation values of spatial parameters are determined in a similar way. The CLDs and ICCs obtained from the quantiser are used as the initial values. A set of predetermined values of channel level differences, C_p , and a set of predetermined values of inter channel coherences, I_p , are determined using,

$$P_p = \begin{cases} [P(1), P(2), \dots, P(p_b - 1), P(p_b), P(p_b + 1), \\ \dots, P(2w), P(2w + 1)] \\ \text{if } P(p_b) \leq w + 1; \\ [P(p_b - w), P(p_b - w + 1), \dots, P(p_b - 1), \\ P(p_b), P(p_b + 1), \dots, P(p_b + w - 1), P(p_b + w)] \\ \text{if } w + 1 < P(p_b) \leq P_{max} - w; \\ [P(P_m - 2w), P(P_m - 2w + 1), \dots, P(p_b - 1), \\ P(p_b), P(p_b + 1), \dots, P(P_m - 1), P(P_m)] \\ \text{if } P_m - w < P(p_b); \end{cases} \quad (7.10)$$

where $P(p_b)$ is the initial decoded parameter, P_m is the size of the codebook, P is the spatial parameter which is either C , or I , p_b is the index of the initial decoded parameter, w is an integer number reflecting the complexity of the procedure. For CLD $w \leq 15$ and for ICC $w \leq 3$ while b is the index of the parameter band.

7.3.2 Basic Scheme of the Encoder

The simplified AbS optimisation is performed based on the proposed MDCT-based closed-loop R-OTT module. The downmix and residual signals, as well as the spatial parameters extracted from the closed-loop R-OTT module, are used

as the initial input of the simplified AbS optimisation. Practically, the AbS-SAC encoder is implemented, as shown in Fig. 7.2. The audio signal in each channel is transformed to spectral coefficients by means of MDCT. A tree scheme employing the closed-loop R-OTT modules is then performed in order to extract the spatial parameters. The new optimised downmix and residual signals are calculated based on the quantised spatial parameters. The downmix signal, as well as the residual signals, are then encoded by the spectral encoders, which actually consist of spectral coefficient quantisation and the noiseless coding scheme. Prior to being supplied to the tree of OTT modules, all signals are decoded back by the spectral decoders. The decoded downmix and residual signals, as well as the quantised spatial parameters, are then used by the tree of OTT modules to upmix the audio signals. The error minimisation block compares the spectral coefficients of the reproduced audio signals with those of the original signals and then computes the errors.

A closed-loop optimisation procedure, utilising the error minimisation block and the tree of OTT modules, can be carried out. The inputs of the optimisation loop are the decoded spectral coefficients of the downmixed and residual signals, as well as the decoded spatial parameters. The purpose of the closed-loop approach is to examine all possible outcomes obtained by combining all quantisation values of the spectral coefficient and spatial parameter quantiser in every possible way (i.e. all possible combinations of every variable in (7.8)).

7.3.3 Algorithm of the AbS Optimisation

The operation of the algorithm can be explained as follows. For each index of the spectral coefficient, the candidates for sub-optimal spectral coefficients of the downmix and residual signals are combined. Then, all resulting possible combinations are examined to jointly choose a sub-optimal coefficient, which

Figure 7.2: Block diagram of the AbS-SAC encoder.

provides the smallest error among the other tested coefficients, for the downmix signal, as well as a sub-optimal coefficient for each residual signal. In performing this task, the tree of OTT modules take the decoded spatial parameters, given by the dequantiser, as inputs. However, the process of selecting the sub-optimal spatial parameters is not performed at this stage.

Note that this task has to be carefully completed by considering the Huffman encoding process. As explained previously (see section IV.1.3), there is a case where all spectral coefficients within a group have zero values. For such a case there is no need to transmit the magnitude of the spectral coefficients. Modifying one or more spectral coefficients within that group causes the Huffman encoding process to be associated with another codebook. As a result, the spectral coefficients within the group need to be transmitted, which causes an increase in the transmitted bitrate. However, the error reduction achieved by modifying those spectral coefficients may not provide a worthy advantage, due to an increase in the transmitted bitrate.

For this reason, the search for the sub-optimal spectral coefficients, particularly downmix and residual signals, is not performed if the maximum absolute value of a group is zero. The process of choosing the sub-optimal spectral coefficients is then followed by the selection of the sub-optimal spatial parameters. For each parameter band, predetermined values of CLDs and ICCs are combined to form all possible combinations. A set of CLDs and ICCs are then selected as the sub-optimal spatial parameters. At this stage, the downmix and the residual signal optimisation is not performed.

The whole sequential process of searching for sub-optimal signals and parameters is repeated as an iteration process. For the second and subsequent iterations, the chosen sub-optimal signals and parameters should be used as inputs of sequential process rather than the ones from the spectral decoder and the spatial dequan-

tiser. The iteration process is terminated when error reduction below a given threshold has been reached. The quality of the reconstructed audio is expected to improve with every iteration, however, the amount of improvement may reduce. Even if further iterations are executed the resultant values may not converge to the optimal values of both the spectral coefficients and the spatial parameters. Thus, the goal of this simplified AbS algorithm is not to provide the optimal or near-optimal signals and parameters. Instead, it is intended to provide a solution for impractical implementation of the full search AbS procedure while minimise the signal distortion.

7.3.4 Complexity of Simplified AbS algorithm

The proposed simplified AbS algorithm is scalable, and its algorithmic complexity depends mainly on the number of loop procedures that have to be performed in finding the sub-optimal spectral coefficients and spatial parameters, as well as the number of R-OTT modules whose signals and parameters are optimised. In order to reduce the complexity of the algorithm, the number of loop procedures and the involved R-OTT modules can be decreased. Based on equations from (4.15) to (4.17), for each spectral coefficient an R-OTT module performs a number of operations, N_{ott} , consisting of 264 additions/multiplications. Hence, the number of operations, N_{op} , for each index of spectral coefficient required by the simplified AbS algorithm can be determined as

$$N_{op} = N_{ott} \times N_{loop} \quad (7.11)$$

where N_{ott} being the number of operations performed by an R-OTT module, and N_{loop} being the number of loop procedures need to be performed. As an illustration, the number of loop procedures, N_{loop} , that have to be executed for encoding 5-channel audio signals is given in Table 7.1.

Table 7.1: Complexity of simplified AbS algorithm

v	Size of Predetermined Vector	Number of Loops
0	1	0
1	3	243
2	5	3125
3	7	16807

w	Size of Predetermined Vector	Number of Loops
0	1	0
1	3	6561
2	5	390625

7.4 Results

The conditions of the experiment were setup the same as those explained in Chapter 5. However, the low correlated audio signal was excluded from the test material, as there is no significantly different performance provided by those both high and low correlated audio signals.

7.4.1 Performance for Different Levels of Complexity

The performance of the simplified AbS algorithm is assessed by investigating the SNR of the reconstructed audio signals against the complexity of the algorithm. The simplified AbS algorithm is tested for various computational complexity by assigning $v = 0, 1, 2, 3$ and $w = 0, 1, 2$. Assigning $v = 0$ and $w = 0$ means that the simplified AbS algorithm is not performed. Moreover, $v = 0$ and $w \neq 0$ means that the simplified AbS algorithm chooses the sub-optimal spatial parameter but does not find the sub-optimal spectral coefficients of the downmix and residual

Table 7.2: Average segSNRs (in decibels) of AbS-SAC

Bitrate/channel		$v = 0$	$v = 1$	$v = 2$	$v = 3$
42.8 kb/s	$w = 0$	21.82	24.31	24.45	24.50
	$w = 1$	22.38	25.12	25.16	25.18
	$w = 2$	22.41	25.19	25.19	25.22
65.4 kb/s	$w = 0$	27.82	30.41	30.45	30.46
	$w = 1$	28.33	30.77	30.79	30.79
	$w = 2$	28.34	30.79	30.81	30.82
83.2 kb/s	$w = 0$	30.66	33.51	33.54	33.72
	$w = 1$	31.12	33.72	33.74	33.74
	$w = 2$	31.13	33.73	33.75	33.75

signals. On the other hand, defining $v \neq 0$ and $w = 0$ instructs the algorithm to select the sub-optimal spectral coefficients without selecting sub-optimal spatial parameters. The algorithm is terminated when the segSNR improvement is less than 10^{-4} . The average segSNRs achieved by the AbS-SAC, operating at three different bitrates: 42.8 kb/s, 65.4 kb/s, and 83.2 kb/s per audio channel, are given in Table 7.2.

As it can be seen, the encoder employing the simplified AbS algorithm (i.e. $v \neq 0$ and $w \neq 0$) outperforms the encoder with parameters $v = 0$ and $w = 0$. These results demonstrates that the simplified AbS algorithm can improve the segSNR, depending on the values of the parameters v and w . The encoder performing the simplified AbS algorithm to find either the sub-optimal spatial parameters only (i.e. $v = 0$ and $w \neq 0$) or the sub-optimal downmix and residual signals only (i.e. $v \neq 0$ and $w = 0$) can improve the average segSNR. It seems that the encoder performing the algorithm for selecting the sub-optimal downmix and residual signals only obtains better segSNR than the encoder selecting the sub-optimal

spatial parameters. However, using both the sub-optimal signals and the sub-optimal spatial parameters in the iterative algorithm provides the largest segSNR. This suggests that the best configuration to perform the proposed simplified AbS algorithm, in terms of high segSNR improvement with the smallest number of iteration loops, is achieved by setting $v = 1$ and $w = 1$. However, to lower the complexity to a more reasonable level, one can set $v = 1$ and $w = 0$ and still achieve considerable SNR improvement.

To have a closer look into the performance of the algorithm when both the optimum signals and the optimum spatial parameters are searched (i.e. $v \neq 0$ and $w \neq 0$), the results can be observed as follows. At an operating bitrate of 42.8 kb/s per audio channel, the encoder employing simplified AbS algorithm with $v = 1$ and $w = 1$ provides the average segSNR of 25.12 dB, therefore giving 3.3 dB of SNR improvement, in comparison with the encoder not performing the simplified AbS algorithm. This configuration (i.e. $v = 1$ and $w = 1$) requires the lowest possible complexity to perform. It can be seen that at 65.4 and 83.2 kb/s per audio channel the AbS-SAC using the same complexity configuration achieves almost the same SNR improvement as the coder operating at 42.8 kb/s per audio channel.

As expected, the encoder performing the most complex algorithm, where $v = 3$ and $w = 2$, reaches the highest SNR improvement. For instance, as many as 3.40 dB of SNR improvement is measured on the encoder operating at 42.8 kb/s per audio channel. However, not much difference of segSNRs is measured when different configurations are performed. The average segSNR grows only slightly even though the number of possible combinations significantly increases. This trend also appears on the AbS-SAC operating at both 65.4 and 83.2 kb/s per audio channel. This is a clear indication that increasing the searching ranges, which results in increasing the number of iteration to be completed, for sub-optimal spectral coefficients and spatial parameters can improve the average segSNR.

Figure 7.3: Performance of the proposed AbS approach using simplified AbS algorithm.

However, for instance in the case of increasing the parameter $v = 2$ to $v = 3$, the SNR improvement is only 0.03 dB, even though the number of loops to be completed increases from 3125 to 16807 as can be seen in Table 7.1. It suggests that the best configuration to perform the proposed simplified AbS algorithm, in terms of high SNR improvement and minimal complexity, is by choosing $v = 1$ and $w = 1$.

7.4.2 Performance Comparison at Various Bitrates

This section presents the experiments conducted to evaluate the overall performance of the proposed AbS-SAC for various bitrates. The simplified AbS algorithm is performed using the lowest complexity, where $v = 1$ and $w = 1$. The conventional tree of R-OTT modules, as well as the closed-loop R-OTT performed in both the subband and the frequency domain are included for benchmarking.

The results are given in Fig. 7.3. The figure shows that the segSNRs achieved by the encoder performing the closed-loop R-OTT module are considerably higher than those reached by the coders applying the conventional R-OTT method for both subband and frequency domains. On the other hand, the coders performing spatial analysis in the frequency domain clearly obtain better performance than those extracting the spatial parameters in the subband domain, except when a very low bitrate implementation is used (i.e. bitrates around 40 kb/s per audio channel and below), where the coders employing the subband domain parameter extraction tend to be better than the coders applying the frequency domain parameterisation.

It also shows that the AbS-SAC scheme outperforms the frequency domain tree of R-OTT modules at all bitrate implementations. The AbS-SAC operating at 40 kb/s per audio channel reaches approximately 3 dB of segSNR improvement. On the other hand, about 5 dB of segSNR improvement is achieved by the coder operating at 160 kb/s per audio channel. It confirms the results detailed in the previous section that performing an analysis by synthesis procedure in the encoder to choose the sub-optimal spectral coefficients and spatial parameters can improve the average segSNR.

Compared to the subband domain open-loop R-OTT as applied in the MPS, the AbS-SAC significantly improves the average segSNR. It can be seen in Fig. 7.4 that approximately 3 dB of segSNR improvement is achieved when both audio coders operate at 40 kb/s per audio channel. The improvement becomes significantly greater in the high bitrate implementation cases, where the AbS-SAC operates at 160 kb/s per audio channel and achieves 17 dB of SNR improvement.

Figure 7.4: Performance of the AbS-SAC for various bitrates in comparison with AAC and MPS.

7.4.3 SNR Comparison with AAC Multichannel

This section presents the experiments conducted to evaluate the performance of the proposed AbS-SAC compared with AAC multichannel. The sub-optimum algorithm is performed using the lowest complexity where $v = 1$ and $w = 1$. The conventional open-loop R-OTT as well as the closed-loop R-OTT performed in both subband and frequency domains are included for benchmarking. The results are given in Fig. 7.4. Please note that to get a fair comparison all codecs, including the AAC, used in the experiments for this section, are highly optimised to provide better SNR. At an operating bitrate of 160 kb/s per audio channel, the AbS-SAC achieves approximately 3 dB of SNR improvement, in comparison with AAC.

7.5 Conclusions

A simplified AbS algorithm has been proposed in this chapter. The simplest case of implementing an AbS loop for an R-OTT module was illustrated in the beginning of this chapter and also described the complexity of the algorithm in light of the condition of optimal signals and parameters desired. Three strategies that have been exploited for simplifying the algorithm are also given in this chapter. Subsequently, the result section presented that significant improvement of performance was achieved by implementing this simplified algorithm.

Chapter 8

Audio Codec Evaluation

In this chapter the overall performance of the AbS-SAC audio codec is evaluated using various critical audio materials. The implementation of the closed-loop R-OTT module is assessed by measuring segSNR and comparing it with the conventional open-loop system. Furthermore, the performance of the closed-loop R-OTT module in the frequency domain is also presented including the evaluation of the disadvantage of the QMF filterbank and the benefit of the MDCT transform in terms of applying the closed-loop R-OTT module. Subjective and objective perceptual tests using ITU-R recommendations are also discussed.

8.1 Experimental Setup

In order to evaluate the proposed system, a number of experiments designed to assess the encoding of 5 and 10 audio channels were conducted. The audio excerpts, sampled at 48 KHz, listed in Table 8.1 were prepared for the experiments. They were selectively chosen from a broad range of long sequence 5.1 audio signals ranging from speeches, pop and classical music, as well as specific sounds

Table 8.1: List of Audio Excerpts for Experiments

Excerpt Name	Description
Applause	Clapping hands of hundreds of people
Drum	Drum and male vocal with guitar as background
Laughter	Sound of hundreds of people laughing
Talk	Male and female speech with music background
Vivaldi	Classical music with vocal

such as clapping hands [136, 137]. For each audio sequence, a limited 12 s audio excerpt was selected based on the possibility of more transient events. All of the 10-channel audio signals were produced by upmixing the 5-channel signals using a simple amplitude panning technique.

The tree scheme of R-OTT modules for downmixing 5 channels into a mono downmix, as given in Fig. 8.1 was used in the experiments. However, the Low Frequency Enhancement (LFE) channel was excluded for simplicity. Each channel of audio signals was segmented into 2048 time domain samples with 50% overlap. In calculating CLDs and ICCs 20 parameter bands were used. The 20 parameter bands for the MDCT-based R-OTT were determined by mapping the 20 parameter bands of the QMF to the 49 scale factor bands of the spectral coefficients. The downmix signal was encoded by AAC. The AAC multichannel codec, implemented as FAAC 1.28 and FAAD2 2.7 [138], was used for benchmarking to demonstrate the usefulness of the proposed AbS-SAC approach, even though it is not the best implementation of the AAC standard.

Figure 8.1: A tree scheme of R-OTT modules used in the experiments

8.2 Evaluation of the Closed-loop R-OTT Module

The evaluation of the closed-loop R-OTT module is aimed at demonstrating that the closed-loop R-OTT algorithm improves the segSNR of the reconstructed audio signals. Table 8.2 shows the results of the experiment comparing segSNRs of multichannel audio encoders employing open-loop and closed-loop R-OTT modules in both filterbank-based (FB-Based) and MDCT-based scenarios. All encoders operate at a bitrate of 160 kb/s per audio channel. Here all encoders are optimised to provide maximum segSNR performance.

The results show that the closed-loop R-OTT algorithm improves the average segSNR of all tested audio excerpts. It clearly indicates that the closed-loop R-OTT algorithm is capable of minimising signal distortion, resulting in segSNR improvement. Furthermore, the MDCT-based closed-loop R-OTT module outperforms, in terms of segSNR, the FB-based closed-loop R-OTT module for all tested audio excerpts other than Applause. It is an indication that the closed-loop R-OTT module generally performs better in the MDCT domain, for high bitrate implementation, as compared with the QMF domain.

Table 8.2: SegSNRs (in decibels) of the Closed-Loop R-OTT (CL R-OTT) Compared to the Open-Loop R-OTT (OL R-OTT)

Audio Excerpt	MPS	FB-Based CL R-OTT	MDCT-Based CL R-OTT
Applause	24.42	30.53	30.04
Drum	23.46	32.82	39.72
Laughter	23.92	32.87	37.05
Talk	24.34	31.13	37.82
Vivaldi	21.91	27.35	44.35

Compared to the other audio excerpts, the Applause audio excerpt provides a lower segSNR when encoded by the proposed MDCT-based closed-loop R-OTT module. This is due to the characteristic of the Applause signal which has many transient events [139]. For an audio frame with transient events, it is less effective to perform encoding in MDCT-based (frequency domain) which has a lower time resolution than the QMF. However, it can be seen that for the Applause audio excerpt both QMF and MDCT schemes are still competitive.

8.3 Analysis of Signal Distortion in the QMF Filterbank

In order to show both the disadvantages of the QMF filterbank and the benefit of the MDCT transform when applying the closed-loop R-OTT module, experiments using various signal transformation schemes were conducted. Four schemes use the QMF filterbank for signal transformation while the two other schemes use the MDCT. In the first scheme the audio signals are transformed from time to subband domain on the encoder side before being transmitted to the decoder.

On the decoder side the subband signals are transformed back to the time domain. This scheme is denoted as QMF Subband Tx. The second scheme, named as QMF Time Tx, is performed by transforming the audio signals from time to subband domain on the encoder side. Furthermore, without performing the spatial analysis block the audio signals are transformed back to the time domain and then transmitted to the decoder. The reverse process is performed on the decoder side.

The third scheme is called QMF Cont R-OTT. It uses the MPS system without performing any coding or quantisation process. This means that all signals and parameters are transmitted in their original form. The fourth, called QMF CL R-OTT, is a scheme that applies the closed-loop R-OTT module in the MPS system. The spatial parameters are quantised and then transmitted to the decoder. However, the original form of the downmix and residual signals are transmitted (here no actual transmission is applied and the original signals are copied to the decoder side) instead of the encoded form.

The fifth and the sixth schemes employ the MDCT transform instead of the QMF. In the fifth scheme, the time domain audio signals are transformed into spectral coefficients and then transmitted to the decoder. On the decoder side, the spectral coefficients are transformed back to time domain audio signals. In the sixth scheme the spectral coefficients are used as inputs for the closed-loop R-OTT module. The downmix and the residual signals are not encoded. Instead, they are transmitted in the form of spectral coefficients. However, the spatial parameters are sent as quantised parameters.

The results are given in Table 8.3. It can be seen that there is a segSNR degradation, for each audio excerpt, when using the QMF Time Tx compared to the QMF Subband Tx scheme. It suggests that both the transformation of the audio signals from subband domain to time domain on the encoder side and the

transformation of the audio signals from time domain to subband domain on the decoder side introduce signal error. Furthermore, the results also show that the QMF CL R-OTT has lower segSNRs than the QMF Cont R-OTT. It suggests that the closed-loop R-OTT module applied in the QMF domain cannot reduce the error introduced by the quantisation of the spatial parameters. As explained in the previous section, different lengths of delay introduced by the analysis and synthesis of filterbanks may prevent the proposed closed-loop R-OTT algorithm to perform properly in minimising signal distortion.

Table 8.3: SegSNRs (dB) of Various Signal Transformation Schemes Using QMF and MDCT

Scheme No	(1)	(2)	(3)
Audio Excerpt	QMF Subband Tx	QMF Time Tx	QMF Cont R-OTT
Applause	59.41	53.39	44.69
Drum	61.75	55.73	44.89
Laughter	59.51	53.50	40.26
Talk	60.34	54.12	43.02
Vivaldi	59.09	53.07	35.91
Scheme No	(4)	(5)	(6)
Audio Excerpt	QMF CL R-OTT	MDCT Spectral Tx	MDCT CL R-OTT
Applause	43.26	285.01	285.00
Drum	36.38	284.98	284.97
Laughter	37.78	285.01	285.00
Talk	34.04	285.01	285.01
Vivaldi	27.90	285.00	284.99

In contrast, it is shown that the MDCT CL R-OTT scheme can achieve transparent quality, equivalent to what can be achieved by the MDCT Spectral Tx,

even though the quantised spatial parameters are transmitted to the decoder. It clearly indicates that the proposed CL R-OTT module is capable of minimising the error introduced by the quantisation process of the spatial parameters. It suggests that the proposed CL R-OTT module performs better in the MDCT domain compared with the QMF domain.

8.4 Subjective Evaluation

The proposed AbS-SAC approach for encoding 5-channel audio signals has been evaluated with a subjective assessment of small impairments in the audio system as recommended in the ITU-R BS.1116-1 [54]. The 'double-blind triple stimulus with a hidden reference' method is used where a subject listens to three audio sequences (stimuli). The first sequence is a known reference. The second and the third are a hidden reference and a decoded audio sequence (object) which are randomly presented. The subject is asked to give score to the second and the third audio sequences compared to the first audio sequence (known reference).

The Subjective Difference Grade (SDG) has five grades: 0 (imperceptible), -1 (perceptible but not annoying), -2 (slightly annoying), -3 (annoying), and -4 (very annoying). A total of 20 listeners participated in this listening test which is considered to be a sufficient number of listener [54]. However, less number of listeners, such as 12 [106], 8 [37], were also acceptable in various subjective tests. As specified in [54] the expertise of the listeners are evaluated by averaging their SDG scores over all audio excerpts. Based on this average SDG score, a post-screening method was applied. Three listeners with an average SDG score greater than zero are assumed to be unable to correctly distinguish between the hidden reference and the tested audio object. Thus, the data from those 3 listeners was discarded. Only the SDG scores from the other 17 listeners were used for the

results. Three codecs were taken under test: AbS-SAC, AAC, and HE-AAC. Both the AbS-SAC and AAC codecs operate at similar bitrates of 51.2 kb/s per audio channel, equal to the overall 5-channel bitrates of 256 kb/s. On the other hand, the HE-AAC codec operates at its typical bitrate of 32 kb/s per audio channel, which is equal to 160 kb/s for all 5 audio channels.

Fig. 8.2(a) presents the average SDG score of each tested audio codec averaged over all audio excerpts. The error bars show the 95% confidence intervals of the mean scores. The results show that the proposed AbS-SAC approach achieves the highest SDG score. Furthermore, Fig. 8.2(b) shows the SDG score of every tested audio excerpt averaged over all tested audio codecs, where the Applause audio excerpt has the lowest SDG score. As expected, it suggests that the Applause audio excerpt is the most critical item among the tested audio excerpts due to the signal characteristic that has more transient events [139]. For this Applause audio excerpt the proposed AbS-SAC approach also achieves the highest SDG score as shown in Fig. 8.2(c).

8.5 Objective Evaluation

The goal of the experiments presented in this section is to objectively assess the perceptual quality of the audio reconstructed from the AbS-SAC technique for various operating bitrates in comparison with the tested AAC codec technique. Considering the complexity of the AbS-SAC encoder, for encoding 5 audio channels, the simplified AbS algorithm is assigned with $v = 1$ and $w = 1$ while for encoding 10 audio channels parameters are set to $v = 1$ and $w = 0$. The experiments using 10-channel audio signals are intended to show that, for larger channels, the proposed AbS-SAC approach can improve the performance for the given bitrate per audio channel. The Objective Difference Grade (ODG), having

Figure 8.2: The results of the subjective test: (a) the SDG scores of the tested audio codecs averaged over all audio excerpts (b) the SDG scores of all audio excerpts averaged over the tested audio codecs (c) the SDG scores of the tested audio codec for Applause audio excerpt.

five grades that are similar to SDG, as recommended in the ITU-R BS.1387-1 [56] is used as a metric for the Perceptual Evaluation of Audio Quality (PEAQ). The PEAQ software developed by McGill University [140] is used for calculating the ODG. The ODG was measured for each channel of audio signal. The average values of the ODG scores are then presented as the final results.

The results of the experiments for encoding 5-channel and 10-channel audio signals are given in Fig. 8.3. For simplicity, the results of AAC 10-channel are not shown, as they are almost identical to those achieved on AAC 5-channel. The overall ODG, averaged over all audio excerpts as shown in the lowest right plot, shows that the AbS-SAC applied to both 5 and 10 channels significantly outperforms, in terms of PEAQ, the tested AAC encoder for all operating bitrates from 40 to 96 kb/s per audio channel. However, the performance increase is greater when encoding 10 channels. It can be seen that an improvement of more than 2 points of ODG grade is achieved on the Applause audio excerpt. Moreover, at least 1 point of ODG improvement is achieved on every tested audio excerpt. The results indicate that the proposed AbS-SAC technique significantly improves encoding performance for a wide range of tested audio materials.

In order to assess the performance of the simplified AbS algorithm with regards to complexity, the bottom-left graph in Fig. 8.3 shows the ODGs of two variants of the proposed AbS-SAC technique, with two extremely different complexity scales, for encoding 10 channels of a Vivaldi audio excerpt. The first one is the AbS-SAC codec using the simplified AbS algorithm where $v = 1$ and $w = 0$ and all spectral coefficients at every OTT module are optimised, and the other does not use the simplified AbS algorithm. The results clearly demonstrate that without the simplified AbS algorithm the proposed codec is still able to achieve significant quality improvement while the simplified AbS algorithm improves the performance further.

Figure 8.3: Objective difference grade (ODG) of the AbS-SAC for various bitrates in comparison with the tested AAC. The ODG scores of the tested AAC encoding 10-channel audio signals are not plotted as they are similar to the ODG scores of the tested AAC encoding 5-channel audio signals.

8.6 Conclusions

This chapter has presented the evaluation of the proposed AbS-SAC audio codec. The performance of the AbS-SAC audio codec is compared to AAC multichannel at bitrates ranging from 40 to 96 kb/s per audio channel. SegSNR measurements, as well as Objective Difference Grades (ODGs), were used as the metric for evaluation. In addition, several audio materials were used for perceptual evaluation. The results have demonstrated that a significant improvement is achieved.

Chapter 9

Conclusions

This chapter presents the summary of this thesis. The aim and objectives of the work are restated. Furthermore, the technical achievement of this work is briefly presented. At the end of the chapter future work is suggested.

9.1 Research Aim and Objectives

This work has aimed to develop a novel technique for spatial audio coding to achieve a higher quality of the reconstructed audio signals. In details this aim can be broken down into some objectives as follows:

- To develop a novel analysis through the synthesis of spatial audio coding, named Analysis by Synthesis Spatial Audio Coding (AbS-SAC) framework, based on an AbS concept. The AbS-SAC framework was designed to be appropriate for different types of spatial analysis and synthesis. Any technique of spatial audio coding is assumed to be applicable within this framework as long as the downmix signal and spatial parameters can be made variable.

Furthermore, it was also expected to be suitable for different number of input and output channels of audio signals.

- To develop the MPEG Surround (MPS) within the AbS-SAC framework. Based on the proposed framework the Reverse One-To-Two (R-OTT) module that is used in the MPS has been developed to work in the form of a closed-loop system. Furthermore, it was focused to investigate the performance of the adapted R-OTT module compared with the conventional open-loop R-OTT module.
- To investigate the possibility of performing an AbS optimisation procedure to find the optimal downmix and residual signals, as well as the optimal spatial parameters. Moreover, the requirement for memory and processing power was also investigated. A simple algorithm with a reasonable complexity level was expected to be successfully implemented.
- To evaluate the proposed AbS-SAC technique and to compare the performance with the existing modern audio coding techniques. Objective and subjective tests to evaluate the proposed technique based on ITU-R recommendation were desired to be sufficient for perceptual evaluation.

9.2 Research Achievements

A comprehensive literature review has been completed on the area of spatial audio coding technique to have a complete understanding on the state-of-the-art approaches. An analysis by synthesis technique is then proposed as a novel approach for spatial audio coding. A number of experiments have demonstrated that the proposed approach can achieve a significant higher performance, in terms of subjective and objective tests, compared with the conventional open-loop approach. This section summarises the proposed technique in detail.

9.2.1 A closed-loop R-OTT module

Chapter 5 has presented a new AbS-SAC framework, where the AbS concept is applied to optimise the downmix and residual signals and also the spatial parameters. It has been presented that the closed-loop R-OTT module was developed, based on the proposed AbS-SAC framework. The downmix and residual signals in the closed-loop R-OTT module were recalculated based on the quantised spatial parameters. Using this method the error introduced by the quantisation of the spatial parameters can be minimised.

Various experiments using 2-channel and 5-channel audio signals as inputs have been conducted to show that the average segmental Signal-to-Noise-Ratio (segSNR) improves significantly. Moreover, the ability of the proposed closed-loop R-OTT module to minimise errors has been thoroughly investigated. It has been shown that there was an indication that the errors contributed by the spatial parameter quantisers were not reduced as much as expected. It suggested that there is an opportunity to minimise the signal distortion further.

9.2.2 The MDCT-based Implementation of the R-OTT Module

Based on the investigation in Chapter 5, the closed-loop R-OTT module was proposed to be applied in the frequency domain. In the beginning part of Chapter 6, the disadvantages of the QMF filterbank were discussed. It was presented that the fractional sample delay of 0.015625 samples that were introduced in the synthesis filterbank caused a less than perfect comparison between the synthesised and the target signals. Thus, it is necessary to use another time-frequency transformation tools to properly operate the closed-loop R-OTT module. In this work the MDCT transform was proposed for some reasons including its ability to

provide a perfect signal reconstruction. Additionally, applying the MDCT-based closed-loop R-OTT module can also simplify the structure of the AbS-SAC encoder.

In order to achieve a better understanding of the implementation of the R-OTT module in the frequency domain, a number of experiments were conducted which focused on investigating the performance of both the open-loop and closed-loop R-OTT modules using various spatial quantisers at a range of operating bitrates between 50 and 150 kb/s per audio channel. The results of the experiments indicated that an appropriate spatial quantiser should be employed in the R-OTT module. Moreover, the results of the experiments also suggested that the MPS's spatial quantisers are the best in terms of coding efficiency. The results also indicated that the MDCT-based closed-loop R-OTT module is capable of minimising the error introduced by the spatial parameter quantisers. Furthermore, the results indicated that the errors are uniformly distributed among the audio channels.

9.2.3 Simplified AbS Optimisation

In Chapter 7 a simplified AbS algorithm, able to find sub-optimal signals and parameters as a solution for the impractical implementation of the full search AbS optimisation procedure, was presented. This simplified algorithm was proposed, as the full search AbS algorithm seems to be an impossible implementation, due to the need for performing huge loops for optimisation. It was previously discussed that the full search AbS procedure for encoding the 5-channel audio signals needed to perform a total of 2.7251×10^{26} loops for finding the optimal parameters for each index of the spectral coefficient. Furthermore, the simplified AbS algorithm offers various complexity levels that can suit different hardware implementation.

The experiments have demonstrated that the proposed simplified AbS algorithm is capable of minimising signal distortion, leading to a higher segSNR compared with the open-loop and the closed-loop R-OTT modules. It is shown that up to 5 dB of segSNR improvement, compared with the upper bound of segSNR that can be achieved by the open-loop R-OTT module, was achieved in the experiments at operating bitrates between 40 and 160 kb/s per audio channel. Moreover, comparison with the AAC multichannel also showed that a significant improvement of segSNR was achieved.

9.2.4 Audio Codec Evaluation

Chapter 8 presented the results of a number of experiments using objective and subjective test methodologies. The evaluation of the proposed AbS-SAC technique was performed using a number of critical audio materials. The objective test, based on the ITU-R BS.1387-1, showed that up to a 2-point of ODG score improvement was achieved in comparison with the AAC multichannel. For each the audio items, it is shown that the improvement of up to 1 ODG score was achieved. Moreover, the results also indicated that, for a larger number of audio channels, the AbS-SAC technique can achieve a higher improvement of ODG score.

The subjective test, performed at a bitrate of 51.2 kb/s per audio channel, showed that the proposed AbS-SAC method achieved the highest SDG score among the tested audio codecs, which included the AAC and the HE-AAC multichannel. For the most critical audio item, the Applause audio excerpt, the AbS-SAC technique also has the highest SDG score. In addition to objective and subjective perceptual evaluation, the segSNR measurement was also performed in the experiments to investigate the performance of the proposed AbS-SAC technique for encoding those critical audio excerpts. The results showed that significant improvement,

compared with the conventional open-loop MPS, was achieved for all tested audio materials.

9.3 Future Work

The R-OTT module has been developed in the form of a closed-loop system. For encoding multichannel audio signals, several R-OTT modules must be structured in a tree scheme. This structure can be very complicated when encoding a large number of audio channels, such as 32 and 64. An investigation to find a more efficient method for downmixing audio signals, as well as a novel spatial parameter, will reduce the complexity of the coding structure. The proposal to implement the frequency domain R-OTT module has a limitation in that the operating bitrates should be higher than 40 kb/s per audio channel. Below this rate the proposed approach will not provide a significant improvement. Methods for implementing the proposed AbS-SAC system at lower bitrates will be interesting to research. In addition, applying the closed-loop R-OTT module in the subband domain of QMF filterbank is also possible to improve the quality of the reconstructed audio signal. This can be carried out by applying a fractional delay filter to delay the original audio signals and make both the synthesised and the original signals perfectly synchronised.

Appendix A

List of Publications

1. I. Elfitri, B. Gunel, and A. Kondo, *Multichannel Audio Coding Based on Analysis By Synthesis*, Proceedings of the IEEE, vol. 99, no. 4, pp. 657-670, April 2011
2. I. Elfitri, A. Kondo, B. Gunel, *Spatial Audio Coding*, UK Patent Application, Application No: GB1020087.1, Date of Filing: 26 November 2010, Publication No: GB2485979, Date of Publication: 6 June 2012
3. I. Elfitri, X. Shi, J. Calic, and A. Kondo, *Analysis By Synthesis Spatial Audio Coding*, IEEE Transaction on Speech, Audio, and Language Processing, Submitted to on 4 November 2011, First revision: 2 June 2012, Second revision: 4 September 2012

Appendix B

List of Abbreviations

3-D	three dimensional
AAC	Advanced Audio Coding
AbS	Analysis by Synthesis
AbS-SAC	Analysis by Synthesis Spatial Audio Coding
ADPCM	Adaptive Differential Pulse Code Modulation
BC	Backward Compatible
BCC	Binaural Cue Coding
BMLD	Binaural Masking Level Difference
CD	Compact Disk
CELP	Code-Excited Linear Prediction
CL	Closed-loop
CLD	Channel Level Difference
QMF	Complex-Modulated Quadrature Mirror Filterbank
CPC	Channel Prediction Coefficient
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DirAC	Directional Audio Coding

DTS	Digital Theater Systems
ETSI	European Telecommunication Standards Institute
FAAC	Freeware Advanced Audio Coder
FAAD	Freeware Advanced Audio Decoder
HD	Harmonic Distortion
HE-AAC	High Efficiency Advanced Audio Coding
HRTF	Head-Related Transfer Function
ICC	Inter Channel Coherence
ICLD	Inter Channel Level Difference
ICTD	Inter Channel Time Difference
ILD	Inter-aural Level Difference
IMDCT	Inverse Modified Discrete Cosine Transform
IOC	Inter Object Cross coherence
ITD	Inter-aural Time Difference
ITU	International Telecommunication Union
ITU-R	International Telecommunication Union - Recommendation
KLT	Karhunen Loeve Transform
LC	Low Complexity
LFE	Low Frequency Enhancement
LPC	Linear Predictive Coding
MCPS	Mega Cycles Per Second
MDCT	Modified Discrete Cosine Transform
MP3	MPEG-1 Layer 3
MPEG	Moving Picture Expert Group
MPS	MPEG Surround
MSE	Mean Squared-Error
MUSHRA	Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems

M/S	mid/side
NRG	Object Energies
ODG	Objective Difference Grade
OL	Open-loop
OLD	Object Level Difference
OTT	One-To-Two
PCM	Pulse Code Modulation
PEAQ	Perceptual Evaluation of Audio Quality
PS	Parametric Stereo
QMF	Quadrature Mirror Filterbank
R-OTT	Reverse One-To-Two
R-TTT	Reverse Three-To-Two
SAC	Spatial Audio Coding
SAOC	Spatial Audio Object Coding
SASC	Spatial Audio Scene Coding
SBR	Spectral band Replication
SDG	Subjective Difference Grade
segSNR	segmental Signal to Noise Ratio
SMR	Signal to Mask Ratio
SNR	Signal-to-Noise Ratio
SSR	Scalable Sampling Rate
TNS	Temporal Noise Shaping
TTT	Three-To-Two
USAC	Unified Speech and Audio Coding
WFS	Wave Field Synthesis
VBAP	Vector Based Amplitude Panning

Bibliography

- [1] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*. New York, USA: Springer, 2002.
- [2] M. Bosi, "High-quality multichannel audio coding: trends and challenges," *J. Audio Eng. Soc.*, vol. 48 No. 6, pp. 588–595, 2000.
- [3] J. Herre, "Audio coding: An all-round entertainment technology," Presented at the 22th Int. Conf.: virtual, synthetic, and entertainment audio, Espoo, Finland, June 2002.
- [4] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, April 2000.
- [5] M. F. Davis, "The AC-3 multichannel coder," Presented at the 95th Convention of the Audio Engineering Society, New York, October 1993.
- [6] C. C. Todd, G. A. Davidson, M. F. Davis, L. D. Fielder, B. D. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," Presented at the 96th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 1994.
- [7] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Ad-

-
- vanced Audio Coding,” *J. Audio Eng. Soc.*, vol. 45 No. 10, pp. 789–814, 1997.
- [8] M. Wolters, K. Kjorling, D. Homm, and H. Purnhagen, “A closer look into MPEG-4 high efficiency AAC,” Presented at the 115th Convention of the Audio Engineering Society, New York, USA, October 2003.
- [9] J. Herre and M. Dietz, “MPEG-4 high-efficiency AAC coding,” *IEEE Signal Proc. Mag.*, vol. 25, no. 3, pp. 137–142, 2008.
- [10] J. Herre, C. Faller, S. Disch, C. Ertel, J. Hilpert, A. Hoelzer, K. Linzmeier, C. Spenger, and P. Kroon, “Spatial audio coding: Next-generation efficient and compatible coding of multi-channel audio,” Presented at the 117th Convention of the Audio Engineering Society, San Fransisco, CA, USA, Oct. 2004.
- [11] J. Herre, “From joint stereo to spatial audio coding - recent progress and standardization,” in *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFx’04)*, Naples, Italy, October 2004.
- [12] J. Herre and S. Disch, “New concepts in parametric coding of spatial audio: From SAC to SAOC,” in *Proc. IEEE Int. Conf. on Multimedia and Expo*, San Fransisco, CA, USA, Oct. 2007.
- [13] F. Baumgarte and C. Faller, “Why binaural cue coding is better than intensity stereo coding,” Presented at the 112th Convention of the Audio Engineering Society, Munich, Germany, May 2002.
- [14] —, “Binaural cue coding-part I : Psychoacoustic fundamentals and design principles,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

-
- [15] C. Faller and F. Baumgarte, "Binaural cue coding-Part II : Schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.
- [16] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: efficient and compatible coding of multi-channel audio," Presented at the 116th Convention of the Audio Engineering Society, Berlin, Germany, May 2004.
- [17] B. Grill, O. Hellmuth, J. Hilpert, J. Herre, and J. Plogsties, "Closing the gap between the multichannel and the stereo audio world: Recent mp3 surround extensions," Presented at the 120th Convention of the Audio Engineering Society, Paris, France, May 2006.
- [18] H. Moon, "A low-complexity design for an mp3 multichannel audio decoding system," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 20, no. 1, pp. 314–321, January 2012.
- [19] J. Blauert, *Spatial Hearing, The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 2001.
- [20] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 1305–1322, 2005.
- [21] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," Presented at the 116th Convention of the Audio Engineering Society, Berlin, Germany, May 2004.
- [22] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," Presented at the 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, Mar. 2003.

-
- [23] S. Quackenbush and J. Herre, "MPEG Surround," *IEEE Multimedia*, vol. 12, no. 4, pp. 18–23, 2005.
- [24] J. Herre *et al.*, "MPEG Surround - The ISO/MPEG standard for efficient and compatible multichannel audio coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, 2008.
- [25] J. Hilpert and S. Disch, "The MPEG Surround audio coding standard [Standards in a nutshell]," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 148–152, Jan. 2009.
- [26] M. M. Goodwin and J.-M. Jot, "Spatial audio scene coding," Presented at the 125th Convention of the Audio Engineering Society, San Fransisco, CA, USA, October 2008.
- [27] J.-M. Jot, J. Merimaa, M. M. Goodwin, A. Krishnaswamy, and J. Laroche, "Spatial audio scene coding in a universal two-channel 3-D stereo format," Presented at the 123th Convention of the Audio Engineering Society, New York, NY, USA, October 2007.
- [28] M. M. Goodwin and J.-M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in *Proc. 123rd AES Convention*, New York, USA, Oct. 2007.
- [29] V. Pulkki and C. Faller, "Directional audio coding: Filter bank and STFT-based design," Presented at the 120th Convention of the Audio Engineering Society, Paris, France, May 2006.
- [30] V. Pulkki, "Directional audio coding in spatial sound reproduction and stereo upmixing," in *Proc. Audio Engineering Society 28th Intl. Conf.*, Pitea, Sweden, Jul. 2006.
- [31] —, "Spatial sound reproduction with directional audio coding," *Journal of Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

-
- [32] J. Vilkamo, T. Lokki, and V. Pulkki, "Directional Audio Coding: Virtual microphone-based synthesis and subjective evaluation," *Journal Audio Eng. Soc.*, vol. 57, no. 9, pp. 709–724, September 2009.
- [33] R. S-Amling, F. Kuech, M. Kallinger, G. D. Galdo, J. Ahonen, and V. Pulkki, "Planar microphone array processing for the analysis and reproduction of spatial audio using directional audio coding," Presented at the 124th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, May 2008.
- [34] J. Engdegard *et al.*, "Spatial audio object coding (SAOC)-The upcoming MPEG standard on parametric object based audio coding," Presented at the 124th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, May 2008.
- [35] J. Herre, C. Falch, D. Mahne, G. del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," Presented at the 128th Convention of the Audio Engineering Society, London, UK, May 2010.
- [36] J. Herre and L. Terentiv, "Parametric coding of audio objects: Technology, performance, and opportunities," Presented at the 42nd Int. Conference: Semantic Audio, Ilmenau, Germany, July 2011.
- [37] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. V. de Par, "Background, concepts, and architecture for the recent MPEG Surround standard on multichannel audio compression," *J. Audio Eng. Soc.*, vol. 55, pp. 331–351, 2007.
- [38] H. Nyquist, "Certain factors affecting telegraph speed," *Trans. of the American Institute of Electrical Engineers*, vol. XLIII, no. 2, pp. 412–422, January 1924.

-
- [39] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IEEE*, vol. 86, no. 2, pp. 447–457, February 1998.
- [40] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th ed. Pearson Prentice Hall, 2007.
- [41] G. J. Miao, *Signal Processing in Digital Communications*, 1st ed. Artech House, Inc, 2007.
- [42] W. B. Kleijn and K. K. Paliwal, *Speech Coding and Synthesis*. Amsterdam/The Netherlands: Elsevier, 1995, ch. An Introduction to Speech Coding, pp. 1–47.
- [43] C. Lamblin, "Recent audio/speech coding developments in ITU-T and future trends," presented at the 16th European Signal Processing Conference (EUSIPCO-2008), 2008.
- [44] A. H. Reeves, "The past, present and future of pcm," *IEEE Spectrum*, pp. 58–63, May 1965.
- [45] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2 No. 2, pp. 60–72, 1995.
- [46] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Tampa, Florida, USA, March 1985.
- [47] J.-M. Valin, T. B. T. C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10-ms delay," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 18, no. 1, pp. 58–67, January 2010.
- [48] Opus, "Opus interactive audio codec," URL <http://www.opus-codec.org/>, accessed 1 February 2013.

-
- [49] FLAC, “Free lossless audio codec,” URL <http://xiph.org/flac/documentation-format-overview.html>, accessed 1 February 2013.
- [50] K. Brandenburg, G. Stoll, Y. Dehery, J. Johnston, L. Kerkhof, and E. Schroder, “ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio,” *J. Audio Eng. Soc.*, vol. 42 No. 10, pp. 780–792, 1994.
- [51] K. Brandenburg, “MP-3 and AAC explained,” Presented at AES 17th Int. Conf. on High Quality Audio Coding, Signa, Italy, September 1999.
- [52] H. G. Musmann, “Genesis of the MP3 audio coding standard,” *IEEE Transactions on Consumer Electronics*, vol. 52, pp. 1043–1049, 2006.
- [53] A. T. C. of Coding of Audio Signals, “Perceptual audio coders: What to listen for,” CD-ROM, 2002.
- [54] ITU-R, “Method for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems,” Recommendation ITU-R BS.1116-1, 1997.
- [55] —, “Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems (MUSHRA),” Recommendation ITU-R BS.1534, 2001.
- [56] —, “Method for Objective Measurements of Perceived Audio Quality,” Recommendation ITU-R BS.1387-1, 2001.
- [57] R. Kilgore, M. Chignell, and P. Smith, “Spatialized audioconferencing: what are the benefits?” in *Proc. Conference of the Centre for Advanced Studies on Collaborative Research*, Toronto, Canada, May 2003.

-
- [58] M. J. Evans, A. I. Tew, and J. A. S. Angus, "Perceived performance of loudspeaker-spatialized speech for teleconferencing," *J. Audio Eng. Soc.*, vol. 48, no. 9, pp. 771–785, Sep. 2000.
- [59] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and B-format microphone array for directional audio coding," in *Proc. AES 30th International Conference*, Saariselka, Finland, Mar. 2007.
- [60] E. Cheng, B. Cheng, C. Ritz, and I. Burnett, "Spatialized teleconferencing: Recording and squeezed rendering of multiple distributed sites," in *Proc. Australian Telecom. Net. and Appl. Conf.*, Adelaide, Australia, Dec. 2008.
- [61] S. Quackenbush and R. Lefebvre, "Performance of MPEG unified speech and audio coding," Presented at the 131st Convention of the Audio Engineering Society, New York, USA, October 2011.
- [62] ISO/IEC, "Information Technology - MPEG Audio Technologies, Part 3: Unified Speech and Audio Coding," ISO/IEC 23003-3/FDIS, International Standards Organization, Geneva, Switzerland, 2012.
- [63] S. Ragot *et al.*, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and voice over IP," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Honolulu, Hawaii, USA, Apr. 2007.
- [64] B. Geiser *et al.*, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, no. 8, pp. 2496–2509, November 2007.
- [65] Y. Hiwasaki *et al.*, "G.711.1: A wideband extension to ITU-T G.711," in *Proc. 16th European Signal Processing Conference (EUSIPCO-2008)*, Lausanne, Switzerland, Aug. 2008.

-
- [66] T. Vaillancourt *et al.*, "ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels," in *Proc. 16th European Signal Processing Conference (EUSIPCO-2008)*, Lausanne, Switzerland, Aug. 2008.
- [67] M. Jelinek, T. Vaillancourt, and J. Gibbs, "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 124–130, October 2009.
- [68] R. V. Cox, D. C. Neto, C. Lamblin, and M. H. Sherif, "ITU-T coders for wideband, superwideband, and fullband speech communication," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 106–109, October 2009.
- [69] B. Geiser *et al.*, "Candidate proposal for ITU-T super-wideband speech and audio coding," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009.
- [70] T. Yamanashi, M. Oshikiri, and H. Ehara, "Low bit-rate high-quality audio encoding and low-complexity bandwidth extension technologies for ITU-TG.718/G.718-SWB," in *Proc. IEEE Intl. Conf. on Communications*, Kyoto, Japan, June 2011.
- [71] L. Miao *et al.*, "G.711.1 Annex D and G.722 Annex B - New ITU-T super-wideband codecs," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Prague, Czech Republic, May 2011.
- [72] V. Eksler and M. Jelinek, "Coding of unquantized spectrum sub-bands in superwideband audio codecs," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Prague, Czech Republic, May 2011.
- [73] B. Kovesi, S. Ragot, C. Lamblin, L. Miao, Z. Liu, and C. Hu, "Re-engineering ITU-T G.722: Low delay and complexity superwideband coding

- at 64 kbit/s with G.722 bitstream watermarking,” in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Prague, Czech Republic, May 2011.
- [74] C. Quinquis and P. Usai, “Fullband conversational codec: What testing methodology ?” in *Proc. 16th European Signal Processing Conference (EUSIPCO-2008)*, Lausanne, Switzerland, Aug. 2008.
- [75] A. Taleb and S. Karapetkov, “G.719: The first ITU-T standard for high-quality conversational fullband audio coding,” *IEEE Communications Magazine*, vol. 47, no. 10, pp. 124–130, October 2009.
- [76] Q. 10/16 ITU-T. (2008) Speech and audio coding and related software tools. Internet draft. [Online]. Available: <http://www.itu.int/ITU-T/studygroups/com16/sg16-q10.html>
- [77] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, 2nd ed. John Wiley Ltd, 2004.
- [78] P. Kroon and W. B. Kleijn, *Speech Coding and Synthesis*. Amsterdam/The Netherlands: Elsevier, 1995, ch. Linear-prediction based analysis-by-synthesis coding, pp. 79–120.
- [79] P. Eisert, “Model-based camera calibration using analysis by synthesis techniques,” in *Proc. of vision, modeling, and visualization*, Erlangen, Germany, November 2002.
- [80] J. H. Chang and N. S. Kim, “A new structural approach in system identification with generalized analysis-by-synthesis for robust speech coding,” *IEEE Trans. on Audio, speech, and Lang. Processing*, vol. 14, no. 3, pp. 747–751, 2006.
- [81] C. G. Bell *et al.*, “Reduction of speech spectra by analysis-by-synthesis techniques,” *J. Acoust. Soc. Am.*, vol. 33, no. 12, pp. 1725–1736, Dec. 1961.

-
- [82] B. S. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Paris, France, May 1982.
- [83] ISO/IEC, "Information Technology - Coding of audio-visual objects, Part 3: Audio," ISO/IEC 14496-3:2009(E), International Standards Organization, Geneva, Switzerland, 2009.
- [84] W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of high frequency tones," *J. Acoust. Soc. Am.*, vol. 29, pp. 988–991, 1957.
- [85] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853, 1999.
- [86] F. Rumsey, *Spatial Audio*, 2nd ed. Oxford, England: Focal Press, 2001.
- [87] P. Mackensen, "Auditive localization. head movements, an additional cue in localization," Ph.D. dissertation, Technischen Universitat Berlin, Berlin, Germany, Jan. 2004.
- [88] U. Zolzer, *DAFX - Digital Audio Effects*, 1st ed. Chichester, England: John Wiley Ltd, 2002.
- [89] C. I. Cheng and G. H. Wakefield, "Introduction to head-related transfer functions HRTF's: Representations of HRTF's in time, frequency, and space," *J. Audio Eng. Soc.*, vol. 49, no. 4, pp. 231–249, 2001.
- [90] ITU-R, "Multi-channel Stereophonic sound system with or without accompanying picture," Recommendation ITU-R BS.775, 1993.
- [91] K. Hamasaki, T. Nishiguchi, R. Okumura, and Y. Nakayama, "Wide listening area with exceptional spatial sound quality of a 22.2 multichannel

-
- sound system,” Presented at the 122nd Convention of the Audio Engineering Society, Vienna, Austria, May 2007.
- [92] S. Kim, Y. Lee, and V. Pulkki, “New 10.2-channel vertical surround system (10.2-vss); comparison study of perceived audio quality in various multi-channel sound systems with height loudspeakers,” Presented at the 129th Convention of the Audio Engineering Society, San Francisco, USA, November 2010.
- [93] E. B. T. Chen, “The native b-format microphone: Part i,” Presented at the 119th Convention of the Audio Engineering Society, New York, USA, October 2005.
- [94] —, “The native b-format microphone: Part ii,” Presented at the 120th Convention of the Audio Engineering Society, Paris, France, May 2006.
- [95] (2009) Internet draft. [Online]. Available: <http://www.holophone.com/>
- [96] (2009) Benefits of a soundfield system. Internet draft. [Online]. Available: <http://www.soundfield.com/soundfield/soundfield.php>
- [97] V. Pulkki, “Virtual sound source positioning using vector based amplitude panning,” *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, Jun. 1997.
- [98] V. Pulkki, M. Karjalainen, and V. Vesa, “Localization, coloration and enhancement of amplitude-panned virtual sources,” in *Proc. AES 16th International Conference*, Rovaniemi, Finland, Apr. 1999.
- [99] V. Pulkki, “Compensating displacement of amplitude panned virtual sources,” in *Proc. AES 22th Int. Conf. on Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, Jun. 2002.

-
- [100] V. Pulkki and T. Hirvonen, "Localization of virtual sources in multichannel audio reproduction," *IEEE Transaction on Speech and Audio Processing*, vol. 13 no. 1, pp. 105–119, 2005.
 - [101] V. Pulkki and M. Karjalainen, "Multichannel audio rendering using amplitude panning," *IEEE Signal Processing Mag.*, vol. 25, no. 3, pp. 118–122, May 2008.
 - [102] J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjorling, J. Plogsties, and J. Koppens, "Multi-channel goes mobile: MPEG Surround binaural rendering," Presented at the AES 29th Int. Conference, Seoul, Korea, September 2006.
 - [103] G. Theile, "Wave filed synthesis-a promising spatial audio rendering concept," in *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFX'04)*, Naples, Italy, Oct. 2004.
 - [104] Y. You, Z. Fejzo, S. Smyth, and P. Smith, "Dts multi-channel 96 khz audio compression," Presented at the AES 17th Int. Conf.: High Quality Audio Coding, Florence, Italy, August 1999.
 - [105] Z. Fejzo, D. Yee, K. McDowell, and L. Kramer, "Dts-hd: Technical overview of lossless mode of operation," Presented at the 118th Convention of the Audio Engineering Society, Barcelona, Spain, May 2005.
 - [106] J. Breebaart, J. Herre, C. Faller, J. Roden, F. Myburg, S. Disch, H. Purnhagen, H. G. M. Neusinger, K. Kjorling, and W. Oomen, "MPEG spatial audio coding/ MPEG surround: Overview and current status," Presented at the 119th Convention of the Audio Engineering Society, New York, USA, October 2005.
 - [107] B. Cheng, C. Ritz, and I. Burnett, *Advances in Multimedia Information Processing 2006*. Berlin, Heidelberg: Springer, 2006, ch. Squeezing the

-
- Auditory Space: A New Approach to Multichannel Audio Coding, pp. 572–581.
- [108] —, “Principles and analysis of the squeezing approach to low bit rate spatial audio coding,” in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Honolulu, Hawaii, USA, Apr. 2007.
- [109] —, *Advances in Multimedia Information Processing-PCM 2007*. Berlin, Heidelberg: Springer, 2007, ch. Encoding Independent Sources in Spatially Squeezed Surround Audio Coding, pp. 804–813.
- [110] —, “A spatial squeezing approach to Ambisonic audio compression,” in *Proc. IEEE Intl. Conf. Acoust. Speech, Signal Process.*, Las Vegas, Nevada, USA, Apr. 2008.
- [111] —, “Psychoacoustic-based quantization of spatial audio cues,” *Electronic Letters*, vol. 44 No. 18, pp. 1098–1099, 2008.
- [112] D. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, “An interchannel redundancy removal approach for high-quality multichannel audio compression,” in *Proc. 109th AES Convention*, Los Angeles, USA, Sep. 2000.
- [113] —, “High-fidelity multichannel audio coding with karhunen-loeve transform,” *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 4, pp. 365–379, Jul. 2003.
- [114] M. Briand, D. Virette, and N. Martin, “Parametric representation of multichannel audio based on principal component analysis,” in *Proc. 120th AES Convention*, Paris, France, May 2006.
- [115] —, “Parametric coding of stereo audio based on principal component analysis,” in *Proc. 9th Int. Conf. on Dig. Audio Effects (DAFx06)*, Montreal, Canada, Sep. 2006.
-

-
- [116] Y. Wang, M. Vilermo, M. Vaananen, and L. Yaroslavsky, "A multichannel audio coding algorithm for inter-channel redundancy removal," in *Proc. 110th AES Convention*, Amsterdam, The Netherlands, May 2001.
 - [117] J. Watkinson, *The MPEG Handbook*. Burlington, MA: Focal Press, 2004.
 - [118] G. Hotho, L. F. Villemoes, and J. Breebaart, "A backward-compatible multichannel audio codec," *IEEE Trans. Audio, Speech Language Process.*, vol. 16, no. 1, pp. 83–93, Jan. 2008.
 - [119] J. Roden, J. Breebart, J. Hilpert, H. Purnhagen, E. Schuijers, J. Koppens, K. Linzmeier, and A. Holzer, "A study of the MPEG Surround quality versus bit-rate curve," Presented at the 123th Convention of the Audio Engineering Society, New York, USA, Oct. 2007.
 - [120] A. Mason, D. Marston, F. Kozamernik, and G. Stoll, "EBU test of multichannel audio codecs," Presented at the 122nd Convention of the Audio Engineering Society, Vienna, Austria, May 2007.
 - [121] D. Marston, F. Kozamernik, G. Stoll, and G. Spikofski, "Further EBU test of multichannel audio codecs," Presented at the 126th Convention of the Audio Engineering Society, Munich, Germany, May 2009.
 - [122] M. Nema and A. Malot, "Comparison of multichannel audio decoders for use in handheld devices," Presented at the 128th Convention of the Audio Engineering Society, London, UK, May 2010.
 - [123] J. Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding : A tutorial introduction," Presented at the AES 17th Int. Conference on High Quality Audio Coding, Florence, Italy, September 1999.
 - [124] ISO/IEC, "Information Technology - Generic coding of moving pictures and associated audio information, Part 7: Advanced Audio Coding," ISO/IEC

-
- 13818-7:2006(E), International Standards Organization, Geneva, Switzerland, 2006.
- [125] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–55, January 1940.
- [126] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. of Am.*, vol. 33, no. 2, p. 248, February 1961.
- [127] J. Engdegard, H. Purnhagen, J. Roden, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," Presented at the 116th Convention of the Audio Engineering Society, Berlin, Germany, May 2004.
- [128] J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjorling, E. Schuijers, J. Hilpert, and F. Myburg, "The reference model architecture for MPEG spatial audio coding," Presented at the 118th Convention of the Audio Engineering Society, Barcelona, Spain, May 2005.
- [129] D. P. Chen, H. F. Hsiao, H. W. Hsu, and C. M. Liu, "Gram-schmidt-based downmixer and decorrelator in the MPEG surround coding," Presented at the 128th Convention of the Audio Engineering Society, London, UK, May 2010.
- [130] ISO/IEC, "Information Technology - MPEG Audio Technologies, Part 1: MPEG Surround," ISO/IEC 23003-1:2007(E), International Standards Organization, Geneva, Switzerland, 2007.
- [131] I. Elfitri, B. Gunel, and A. M. Kondo, "Multichannel audio coding based on analysis by synthesis," *Proc. IEEE*, vol. 99, no. 4, pp. 657–670, April 2011.
- [132] —, "Spatial audio coding," UK Patent Publication No. GB2485979, 2012.

-
- [133] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay," *IEEE Signal Proc. Magazine*, pp. 30–60, January 1996.
- [134] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: Exact results and applications to coding artifacts reduction," *IEEE Trans. on Speech and Audio Proc.*, vol. 12, no. 3, pp. 302–312, May 2004.
- [135] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 16, no. 8, pp. 1361–1372, November 2008.
- [136] (2012) The Nordic Sound. [Online]. Available: <http://www.2l.no/hires/index.html>
- [137] (2012) B-Format Audio. [Online]. Available: <http://www.soundfield.com/downloads/b-format.php>
- [138] (2012) AudioCoding.com. [Online]. Available: www.audiocoding.com
- [139] G. Hotho, S. van de Par, and J. Breebaart, "Multichannel coding of applause signal," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, January 2008.
- [140] P. Kabal, "TSP Lab - Audio file programs and routines - AFsp," URL <http://www-mmsp.ece.mcgill.ca/documents/Software/>, McGill University, Montreal, Quebec, Canada, 2003.