Implementation and optimization of parametric stereo encoding in enhanced aacPlus encoder

Samsudin

2007

Samsudin. (2007). Implementation and optimization of parametric stereo encoding in enhanced aacPlus encoder. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/3521>

<https://doi.org/10.32657/10356/3521>

Nanyang Technological University

Downloaded on 26 Apr 2023 17:16:04 SGT

Implementation and Optimisation of Parametric Stereo Encoding in Enhanced aacPlus Encoder

Samsudin

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Master of Engineering

2007

Acknowledgements

First and foremost, I would like to thank God, my parents, my brother and sister for their abundant blessings and guidance in my life and what I have achieved so far. I would like to extend my gratitude to NTU School of Electrical and Electronic Engineering and STMicroelectronics Asia Pacific, Pte. Ltd. for giving me the opportunity and financial support to work on this project.

These two years of research and works had been demanding and tough. I would like to thank these following peoples; without their sincere help I would have never come to writing these paragraphs.

I would like to express my deepest gratitude to my supervisors Prof. Ng Boon Poh and Dr. Farook Sattar for their continuous support, both mentally and academically. Their confidence in me and their input to the project have provided a great motivation to complete this research.

I would also like to thank my supervisors at STMicroelectronics, Ms. Evelyn Kurniawati and Ms. Sapna George, for giving valuable inputs and providing the necessary resources to carry out the project. I truly enjoy our weekly discussions which have generated many fruitful ideas.

Finally, I thank all my friends who have been there for me, keeping me sane during those tough and challenging times, and the staffs of ISRL 3 Laboratory who have provided a great research facilities and environment.

I dedicate this thesis to all of you.

Table of Contents

ACKNOWLEDGEMENTS.....	i
TABLE OF CONTENTS.....	ii
ABSTRACT.....	vi
LIST OF FIGURES.....	viii
LIST OF TABLES.....	xi
LIST OF PSEUDOCODES.....	xii
LIST OF TERMS AND ABBREVIATIONS.....	xiii
CHAPTER 1 : INTRODUCTION	1
1.1 Overview of Lossy Audio Compression.....	2
1.1.1 Lossy Audio Coding Schemes	2
1.1.2 Perceptual and Parametric Audio Coding Paradigms	4
1.1.3 Recent Development in Low Bitrate Audio Coding	7
1.2 Motivations.....	10
1.3 Objectives and Scope of Research	12
1.4 Contributions of Thesis.....	13
1.5 Outline of Thesis.....	14
CHAPTER 2 : OVERVIEW OF PARAMETRIC STEREO CODING.....	16
2.1 Psychoacoustic Background	18
2.2 MPEG-4 Parametric Stereo Encoding.....	20

2.2.1	Hybrid Analysis Filtering	21
2.2.2	Spatial Parameter Extraction	23
2.2.3	Quantization and Bitstream Formatting	28
2.2.4	Stereo to Mono Downmixing	28
2.3	MPEG-4 Parametric Stereo Decoding	28
2.3.1	Analysis Filtering and Decorrelation	29
2.3.2	Stereo Synthesis	29

CHAPTER 3 : SOFTWARE IMPLEMENTATION OF MPEG-4 PARAMETRIC STEREO ENCODER 33

3.1	Overview of Enhanced aacPlus Encoder	33
3.2	C-based MPEG-4 Parametric Stereo Encoder Implementation	36
3.2.1	Hybrid Analysis Filtering	37
3.2.2	Spatial Parameter Extraction	38
3.2.3	Stereo to Mono Downmixing	40
3.2.4	Bitstream Formatting	41
3.2.5	Encoding Optimizations	42
3.3	Matlab-based MPEG-4 Parametric Stereo Model	48
3.3.1	Data Structure	49
3.3.2	Main modules	50

CHAPTER 4 : OPTIMIZED DOWNMIXING SCHEME FOR PARAMETRIC STEREO ENCODER 53

4.1	Stereo to Mono Compatibility Problem	53
4.2	Proposed Optimized Downmixing Scheme	56
4.3	Results and Discussion	57
4.3.1	Objective Audio Quality	59
4.3.2	Power Preservation	60

4.4	Downmixing Scheme Simplification.....	65
------------	--	-----------

CHAPTER 5 : UNIFIED TRANSIENT DETECTOR FOR ENHANCED AACPLUS ENCODER.....	66
---	-----------

5.1	Overview of Transient Detection in Music Signals	67
------------	---	-----------

5.2	Enhanced aacPlus Transient Detection Schemes	70
------------	---	-----------

5.3	Proposed Unified Transient Detector	73
------------	--	-----------

5.4	Results and Discussion.....	76
------------	------------------------------------	-----------

5.4.1	Test Items and Detection Signal	76
-------	---------------------------------	----

5.4.2	Transient Information Matching Accuracy	78
-------	---	----

5.4.3	Conformance to the 3GPP SBR Transient Detection Conformance Testing	78
-------	---	----

5.5	Overall Assessment of the Implementation and Optimization.....	81
------------	---	-----------

5.5.1	Complexity Analysis	81
-------	---------------------	----

5.5.2	Subjective Listening Test Result	82
-------	----------------------------------	----

CHAPTER 6 : OBJECTIVE EVALUATION METHOD FOR SPATIAL IMAGE DISTORTION IN STEREO AUDIO PROCESSING	86
--	-----------

6.1	Stereo Audio Model	87
------------	---------------------------------	-----------

6.2	Proposed Methodology.....	88
------------	----------------------------------	-----------

6.2.1	Spatial Cue Calculation	88
-------	-------------------------	----

6.2.2	Spatial Cue Selection	91
-------	-----------------------	----

6.2.3	Fixing ICC Threshold for Spatial Cue Selection: Recommendation	92
-------	--	----

6.2.4	Statistical Analysis Method	95
-------	-----------------------------	----

6.3	Evaluation of the Proposed Method.....	96
------------	---	-----------

6.3.1	Spatial Image Narrowing	96
-------	-------------------------	----

6.3.2 Spatial Image Instability	97
---------------------------------	----

CHAPTER 7 : SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	
.....	101
7.1 Summary and Conclusions.....	101
7.2 Recommendation for Future Research.....	103

Abstract

Digital broadcasting, internet streaming, and 3rd Generation (3G) mobile technology are some of the few emerging applications which require low-bitrate audio transmission over limited bandwidth. The state-of-the-art low-bitrate audio coder is *enhanced aacPlus*, which is a combination of *Parametric Stereo* (PS), *Spectral Band Replication* (SBR) and *Advanced Audio Coding* (AAC). PS as the newest addition to the coder makes it possible to encode the audio at a bitrate of as low as 24 kbps with acceptable audio quality.

The idea behind PS is to code stereo audio as a *monaural downmix* signal and a small amount of *spatial parameters* which describe its spatial image. The monaural signal can then be encoded by any generic audio coder while the spatial parameters are embedded into the resulting mono audio bitstream. At the decoder, the decoded monaural signal is expanded back into the stereo audio using the transmitted parameters. The details of PS encoding as well as a general overview of the decoding process are presented in this thesis.

A baseline version of PS encoder has been provided in the enhanced aacPlus encoder implementation from 3rd Generation Partnership Project (3GPP). This reference encoder only supports a small subset of the PS configurations as defined in the MPEG-4 PS standard. An implementation of PS encoder which supports the full MPEG-4 PS configurations is presented. Along with the implementation, two optimizations are proposed.

As the signal content of the original stereo audio is solely represented in the monaural signal, the downmixing process has to preserve the stereo signal components as much as possible. However, when the stereo-to-mono downmixing is performed, signal coloration and unwanted attenuation might

occur due to the phase difference of the stereo signals. As a first optimization, a subband-domain downmixing scheme that minimizes this problem is proposed. The objective audio quality evaluation result confirms that the proposed downmixing scheme gives a better quality as compared to the reference downmixing scheme. This is shown by the improvements of the objective audio quality by up to 1.5 *objective difference grade* (ODG) points.

In enhanced aacPlus encoder, the time resolution of SBR and AAC encoding are adaptive. Whenever there is a transient detected in the current frame, these encoders increase their time resolution accordingly. For this purpose, SBR and AAC have their own transient detectors. However, both detectors basically perform detection on the same signal. A unified, low-complexity transient detector which targets the reduction of the encoding complexity is presented as the second optimization. The detection is performed as a side product of PS spatial parameter calculation and the transient information is subsequently passed to SBR and AAC encoders. When replacing both transient detectors, a reduction in the encoder complexity of up to 8% can be achieved, while still giving at least 80% matching to the transient information originally output by both detectors.

A subjective listening test to evaluate both optimizations reveals that the optimized encoder is able to perform as well as the reference encoder with a total saving of 4% of the computational complexity.

In addition to the optimizations, a concept of an objective method to evaluate spatial image distortion due to audio processing is proposed. The method is useful to assess how certain processing affects the spatial image of the processed audio. It is offered as an alternative to the time-consuming subjective listening test. Two cases of spatial image distortion are analyzed: spatial image *narrowing* and *instability*. The evaluation of the proposed method reveals that the output metrics defined are able to approximate the simulated spatial distortion.

List of Figures

Figure 1.1: General framework of MPEG-4 audio coding.	4
Figure 1.2: Perceptual audio coding scheme.	6
Figure 1.3: A general model of parametric audio coding scheme.	7
Figure 1.4: Subjective listening test result comparing enhanced aacPlus, aacPlus and AAC in its low complexity (LC) profile.	10
Figure 2.1: General model of audio encoder-decoder with a combination of perceptual audio coder and parametric stereo coder.	17
Figure 2.2: Informative PS encoding process as suggested by MPEG-4 PS standard.	20
Figure 2.3: Hybrid analysis filtering structure for PS encoder.	21
Figure 2.4: Illustration of the PS sub-framing.	25
Figure 2.5: MPEG-4 Parametric Stereo decoding process.	29
Figure 2.6: Frequency and phase response of IIR all-pass decorrelation filter for 34 stereo bands configuration.	30
Figure 3.1: General structure of enhanced aacPlus encoder.	34
Figure 3.2: Top – down enhanced aacPlus encoder software module tree.	35
Figure 3.3: Implementation structure of PS encoding in eaacPlus encoder.	36
Figure 3.4: Hybrid analysis filtering scheme for 34 stereo bands configuration in PS encoder.	37
Figure 3.5: Bitstream structure of enhanced aacPlus codec.	41
Figure 3.6: Direct phase quantization.	43
Figure 3.7: Comparison of the conventional and direct quantization method.	45
Figure 3.8: Testing of OPD calculation optimization.	47
Figure 3.9: Matlab-based parametric stereo encoder – decoder model.	49
Figure 3.10: Modules in the Matlab-based PS model.	49

Figure 4.1: Normalized power of the monaural signal generated with the time-domain averaging and subband-domain equalization schemes.	55
Figure 4.2: Structure of parametric stereo encoder with the proposed downmixing scheme.	57
Figure 4.3: Normalized power of the monaural signal generated with the proposed downmixing scheme.	58
Figure 4.4: Objective audio quality test results of stereo audio decoded from monaural audio generated by subband-domain equalization and proposed downmixing schemes.	60
Figure 4.5: Inter-channel phase difference histogram of audio segment from <i>canyon</i> and <i>castanet</i> .	62
Figure 4.6: Downmixing power preservation for <i>canyon</i> .	63
Figure 4.7: Downmixing power preservation for <i>castanet</i> .	64
Figure 5.1: Ideal case of a time-domain transient signal and its common representation by onset, attack and decay of transient, and general transient detection scheme.	68
Figure 5.2: AAC transient detection in enhanced aacPlus encoder.	72
Figure 5.3: Percentage of the complexity of the unified transient detector with respect to the total complexity of the SBR and AAC transient detectors.	75
Figure 5.4: A transient audio segment from <i>castanet</i> .	77
Figure 5.5: Transient information of <i>castanet</i> audio segment translated and passed from the unified transient detector to the SBR module, as compared to the original SBR transient detector output.	79
Figure 5.6: Transient information of <i>castanet</i> audio segment translated and passed from the unified transient detector to the AAC module, as compared to the original AAC transient detector output.	80
Figure 5.7: Matching accuracy of the transient information detected and passed by the unified transient detector to the transient information from the original SBR and AAC detectors.	80
Figure 5.8: Enhanced aacPlus encoder complexity distribution.	83

Figure 5.9: Subjective listening result to compare optimized and reference encoder.	85
Figure 6.1: Model of objective evaluation of stereo audio spatial image distortion.	88
Figure 6.2: Percentage of the selected directional source power as a function of the ICC threshold at different DSAR.	94
Figure 6.3: Percentage of selected ICLD values that falls within ± 1 dB from the original source ICLDs.	95
Figure 6.4: Spatial distortion evaluation result for spatial image narrowing.	97
Figure 6.5: Spatial distortion analysis output for unstable spatial image with varying instability rate.	99
Figure 6.6: Spatial distortion analysis output for unstable spatial image with varying degree of instability.	100

List of Tables

Table 2.1: MPEG-4 PS stereo band configuration and the associated number of frequency channels of the low frequency filtering.	24
Table 2.2: Mixing procedures defined in MPEG-4 Parametric Stereo standard.	32
Table 3.1: Phase quantization grid defined in MPEG-4 Parametric Stereo standard.	43
Table 4.1: Test stereo audio signals used to evaluate the performance of the proposed downmixing scheme.	59
Table 5.1: Complexity of the SBR transient detector.	70
Table 5.2: Complexity of the AAC transient detector.	73
Table 5.3: Complexity counts of the proposed unified transient detector.	75
Table 5.4: Test items used to evaluate the performance of the proposed unified transient detector.	76

List of Pseudocodes

Pseudocode 3.1: PS spatial parameter extraction.	39
Pseudocode 3.2: Direct phase quantization when the complex sample is located between quantization level 0 and 1.	44
Pseudocode 3.3: Direct phase quantization when the complex sample is located between quantization level 1 and 2.	44
Pseudocode 5.1: SBR transient detection in enhanced aacPlus encoder.	71
Pseudocode 6.1: Spatial cue calculation for the spatial image distortion evaluation method.	90

List of Terms and Abbreviations

3GPP	3rd Generation Partnership Project
AAC	Advanced Audio Coding
aacPlus	a coding scheme combining SBR and AAC
CD	compact disc
DSAR	Direct Source to Ambience Ratio
eaacPlus	Enhanced aacPlus, a coding scheme combining PS, SBR and AAC
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
hybrid subband	subband signal output from the hybrid analysis filter, which is a QMF filterbank followed by low-frequency filterbanks to increase the frequency resolution of the lower QMF subbands
IC	Inter-aural Coherence
ICC	Inter-channel Coherence
ICLD	Inter-Channel Level Difference
IID	Inter-channel Intensity Difference
IIR	Infinite Impulse Response
ILD	Inter-aural Level Difference
IPD	Inter-channel Phase Difference
ITD	Inter-aural Time Difference
kbps	kilobits per second
MPEG	Moving Picture Experts Group
MPEG-4	Multimedia standard from MPEG, standardized in 1999

ODG	Objective Difference Grade
OPD	Overall Phase Difference
PEAQ	Perceptual Evaluation of Audio Quality
PS	Parametric Stereo
QMF	Quadrature Mirror Filter
SBR	Spectral Band Replication
stereo band	non-uniform frequency group of hybrid subband to approximate ERB bandwidth, the PS spatial parameters are calculated at each of the stereo band
wMOPS	weighted million operations per second

Chapter 1

Introduction

The introduction of the *compact disc* (CD) in the early 1980s brought a shift in the trend of sound representation from analog to digital. Conventional CD stores audio with a sample rate of 44.1 kHz and resolution of 16 bits per sample. This brings up to 1,500 kbps of stereo audio data rate. With the development of the Internet and wireless multimedia technology, digital music distribution, and portable music player, to name a few, a need for more compact audio representation arose. The main problems of such applications are the limitation of bandwidth and storage space. These problems have led many researchers to develop audio compression schemes which are able to represent digital audio at a much lower bit-rate while retaining the original audio quality (*'transparent quality'*).

Basically, there are two classes of audio compression: *lossless* and *lossy* audio compression. Lossless audio compression schemes preserve the audio signal sample-by-sample such that the decompressed audio is identical to the original one. However, the exact reconstruction property comes at the cost of lower compression efficiency. Lossy audio compression schemes are able to achieve greater compression efficiency by relaxing the constraint of perfect waveform reconstruction. Instead of reconstructing the exact original audio waveform, these schemes recreate audio signal that is *perceptually* similar to the

original one. Lossy audio compression schemes are able to operate on an extremely wide range of bitrate and applications.

This chapter gives a brief overview of the various lossy audio compression schemes and standards. Two lossy audio coding paradigms: *perceptual* and *parametric* coding will be reviewed. Some recent developments in the field of low bitrate audio coding which motivates the works presented in this thesis will be summarized. Finally, the objectives and contributions of this thesis will be presented.

1.1 Overview of Lossy Audio Compression

Over the decades, many lossy audio compression schemes have been introduced. *Moving Picture Experts Group* (MPEG) has been playing an important role in the standardization of audio compression schemes. The first three standards: MPEG-1, MPEG-2 and MPEG-4 define state-of-the-art high quality compression of digital audio. Apart from MPEG, there are several proprietary audio compression schemes which have made their way to fame in the commercial audio coding market.

1.1.1 Lossy Audio Coding Schemes

MPEG coding schemes *MPEG-1* was standardized in 1992, reaching transparent quality at a stereo audio bitrate of as low as 192 kbps [1]. MPEG-1 layer 3 which is more popularly known as *MP3* eventually becomes the most successful and widely used coding scheme even up to the present days. MPEG-2 was standardized in 1994, extending MPEG-1 towards backward-compatible multi-channel audio coding at low bitrate. Beside that, a non backward-compatible coding scheme called *Advanced Audio Coding* (AAC) was introduced.

AAC gives a similar quality to MP3 at 128 kbps stereo hence it is about 30% more efficient.

MPEG-4 was standardized in 1999, supporting practically any application scenario from extremely low bitrate to high quality multi-channel audio applications. It provides a complete toolbox for audio and speech coding of natural and synthetic content and bitrates ranging from 2 to 64 kbps [2][3]. The general framework for MPEG-4 Audio is shown in Figure 1.1.

MPEG-4 *Natural Audio Coding* tools contain a set of different coders for different classes of signal and bitrates:

- *General Audio (G/A)* coder, built around MPEG-2 AAC and TwinVQ (vector quantization),
- *Code Excited Linear Prediction (CELP)* speech coder for narrow-band and wide-band speech,
- *Parametric coder* for speech: Harmonic Vector Excitation Coding (HVXC), and audio: Harmonic and Individual Lines and Noise (HILN), Sinusoidal Coding (SSC),
- Combined *scalable coder*.

MPEG-4 *Synthetic Audio Coding* tools define methods for synthetic (artificial) sound materials:

- *Structured Audio (SA)*, which convert a structured representation into a synthetic sound signal,
- *Text To Speech (TTS)*, which allows generation of synthetic speech from a text or a text with prosodic parameters.

The choice of which coding tool to use depends on the nature of the application. Some factors are the bandwidth constraints, nature of the signal, audio quality, interactivity and flexibility of the multimedia content.

Figure 1.1: General framework of MPEG-4 audio coding with a wide range of choices of bitrates for various applications.

Other coding schemes A few other well-known proprietary audio coding schemes basically utilize waveform coding paradigm. The differences amongst them generally lie in the usage of the filter banks, psychoacoustic models, quantization and coding methods, and the approach to multi-channel coding. Some of the popular schemes are: Dolby's *AC-2* and *AC-3*, Sony's *Adaptive Transform Acoustic Coding* (ATRAC) and Microsoft's *Windows Media Audio* (WMA).

1.1.2 Perceptual and Parametric Audio Coding Paradigms

Looking at the development of MPEG audio standard, two lossy audio coding paradigms can be observed: *perceptual coding* (MPEG-1, MPEG-2) and *parametric coding* (MPEG-4 HILN and SSC). MPEG-4 as the latest MPEG standard in audio

compression is moving towards parametric coding for high-quality, low bitrates compression.

Perceptual audio coding

The principle behind perceptual audio coding scheme is to compress an audio signal by removing the *statistical redundancy* and *perceptual irrelevancy* in the signal [4][5]. Generally, adjacent audio samples contain a certain degree of correlation. Statistical redundancy is exploited by decorrelating the signal using e.g. prediction or transform followed by a quantizer. The quantized data might still contain considerable redundancy which can be removed by using run length or entropy coding such as Huffman and arithmetic coding [4]. Figure 1.2(a) shows a generic model of a perceptual audio encoder which is a basic block in many perceptual coders such as MP3 and AAC. The statistical redundancy removal is depicted in the upper branch.

The exploitation of the perceptual irrelevancy is possible due to the limitation of human auditory system, which has been characterized from psychoacoustic studies. The most important concept is the masking phenomena, whereby the presence of a sound masks the audibility of another sound [5]. The masked sound is irrelevant and hence it does not need to be coded. From another point of view, quantization noise can be introduced up to a certain level where it is still inaudible (masked), hence allowing adaptive bit allocation to quantize the data. The lower branch of Figure 1.2(a) shows the perceptual analysis which delivers masking threshold to regulate the bit allocation while Figure 1.2(b) illustrates the concept of masking.

Parametric audio coding

As compared to perceptual coding, parametric audio coding scheme decomposes the audio signal in a more adaptive and flexible way. In this scheme, the audio signal is decomposed into several objects whereby each object is a source model which is described by a set of parameters. Generally,

the approach taken is *analysis-by-synthesis* where each model is analyzed, synthesized and subtracted from the original signal. Other models are then applied to the residual signal. Some source models used for parametric coder are: harmonic and individual lines, sinusoidal, transient, and noise [6][7][8].

The general model for parametric coding is shown in Figure 1.3. Parametric coder employs perceptual model for quantization of model parameter, as well as to select only those perceptually relevant components or parameters to be transmitted, leading to further bitrate reduction.

(a)

(b)

Figure 1.2: Perceptual audio coding scheme: (a) a general model of perceptual audio coding scheme, (b) an illustration of the masking phenomena with a tone (masker) masking another nearby tone (masked sound) which has a power below the masking threshold.

Figure 1.3: A general model of parametric audio coding scheme.

1.1.3 Recent Development in Low Bitrate Audio Coding

Digital broadcasting, internet streaming, and 3rd Generation (3G) mobile technology are some of the few emerging applications which require audio transmission over limited bandwidth. Some typical mobile application scenarios are news and music listening, commercial advertisements, interactive gaming, and music download. It has been recognized that for many audio services, especially when accompanied by video, the audio data rate will need to be pushed to as low as possible, including the sub- 32 kbps range. For example 3rd Generation Partnership Project (3GPP) which is the body standardizing Global System for Mobile (GSM) and 3G mobile applications defines two bitrate range for its audio codec standard: *low-rate range* up to 24 kbps and *high-rate range* higher than 24 kbps [9][10]. These applications require efficient audio coding schemes that provide low bitrate while still maintaining the quality and processing complexity.

State-of-the-art perceptual coding schemes such as MP3 and AAC provide transparent quality (meaning, the decoded audio is perceptually indistinguishable from the original audio) at 192 and 128 kbps stereo, respectively. As the bitrate is lowered, less bits are available to code the full audio bandwidth or to keep the quantization noise below the masking threshold [1]. These factors contribute to the restriction of audio bandwidth and degradation of the audio quality. As the bitrate is lowered even further, the audio quality deteriorates to the point where the degradation is unacceptable. Hence, these perceptual coders are not suitable for very low bitrate applications.

As a solution, MPEG-4 audio standard goes towards the parametric coding schemes. Two most recent additions to the standard are MPEG-4 *Spectral Band Replication* (SBR) [11] and *Sinusoidal Coding* (SSC) [12], which are essentially parametric audio coding tools aiming for low bitrate coding.

SBR is a bandwidth parameterization tool used in combination with the AAC perceptual audio codec. This combination is known as *High Efficiency* (HE-) AAC profile in MPEG-4, or simply *aacPlus*. In SBR coding scheme, wideband audio signal is represented by its lower audio spectrum and a small amount of data which describes the high frequency portion of the spectrum [11][13]. The decoder uses these data to reconstruct the missing high frequency portion during the decoding process. The data rate required for the bandwidth parameter is much less than it is required to code the full audio bandwidth using conventional AAC encoder. Hence, it can be used to lower the audio bitrate or to improve the audio quality for the same bitrate as compared to non-SBR encoding.

SSC is a wideband, high-quality parametric audio coder that models the audio signal as sinusoidal, transient, and noise components at very low bitrates. More importantly, SSC provides *Parametric Stereo* (PS) as a tool to parameterize the spatial image of the stereo audio. Only the monaural downmix signal and a small amount of spatial parameters need to be coded. The PS decoder will then

use the spatial parameters to reconstruct the stereo audio from the decoded monaural signal. Similar to SBR, PS can be used to lower the encoding bitrate or to improve the audio quality for the same bitrate since more bits are available to code just the monaural audio.

Recently, 3GPP adopts *aacPlus* in combination with PS in its high-quality audio codec standard [14]. The codec is referred to as HE-AAC version 2 or *enhanced aacPlus* (*eaacPlus*). 3GPP conducted an extensive double-blind listening test using the *Multi Stimulus test with Hidden Reference and Anchors* (MUSHRA) method which were designed to give a reliable and repeatable measure of the audio quality of intermediate-quality signal. In MUSHRA listening test, the assessors are required to rate the quality of the audio produced by the various codecs under test by assigning quality scores which typically range from 0 to 100. The listening test showed that *eaacPlus* provides excellent quality stereo audio at a bitrate of as low as 24 kbps [15], as compared to AAC and *aacPlus* coding schemes.

The listening test result is shown in Figure 1.4. It can also be observed that *eaacPlus* at 24 kbps gives an equal quality as *aacPlus* at 32 kbps stereo. Interestingly, MPEG-4 verification test [16] revealed that *aacPlus* at 32 kbps stereo achieves a perceptual quality that is better than AAC at 48 kbps stereo and similar or slightly worse than AAC at 64 kbps stereo. Hence it can be concluded that *eaacPlus* gives a similar audio quality, but with more than twice the coding efficiency as compared to the conventional AAC.

It can be seen that at very low bitrate, the combination of perceptual and parametric audio coding is superior to the conventional perceptual coding. For such bitrates where transparent perceptual coding is virtually impossible, *eaacPlus* proves to be a powerful audio coding scheme that provides a solution. PS as one of the enabling technology in *eaacPlus* offers a substantial coding gain as compared to basic AAC and *aacPlus* coding schemes.

Figure 1.4: Subjective listening test result comparing enhanced aacPlus, aacPlus and AAC in its low complexity (LC) profile. The plot shown is the MUSHRA score relative to a 7 kHz stereo anchor.

1.2 Motivations

EaacPlus is considered as the state-of-the-art low-bitrate audio coding scheme. It is standardized in both 3GPP audio codec standard and MPEG-4 audio standard. While implementation of the AAC and aacPlus encoder has been available much earlier, PS encoder has just been recently integrated into aacPlus and this combination is made available as a reference eaacPlus software implementation by 3GPP. The reference eaacPlus encoder includes a baseline version of the PS encoder. This version only supports a small subset of the PS encoding configurations as defined in MPEG-4 PS standard. Some important features that are not implemented are phase parameter calculation and improved spatial parameter update rate.

For audio coder developer, audio quality is always a main concern. In PS coding, as the content of the stereo audio is solely represented by the monaural downmix, the quality of the audio is greatly affected by how much the downmixing process is able to preserve the overall signal content. In practice, many stereo recording techniques result in frequency-dependent out-of-phase signal components. Downmixing these recordings by simple averaging may result in signal coloration due to the phase cancellation of the stereo signal. This problem provides a motivation to work on a more complex downmixing scheme that is able to minimize the phase cancellation.

In eaacPlus encoding, a significant amount of computational resources are dedicated to perform transient detection in AAC and SBR encoders. Transient detection is important to adaptively control the time resolution of the encoding block, such that unwanted audio artifacts can be avoided. The main principle of a transient detector is to detect a sudden rise of energy level in the signal.

AAC encoder has been available as early as 1999 and it has a transient detection module that regulates the switching from a long to a short encoding block when transient is detected. When SBR was standardized in 2001, it was subsequently integrated with the AAC encoder to build aacPlus. SBR itself has a transient detection module that regulates a variable framing scheme which is dependant on the location of the transient. Finally, PS encoder is integrated with the AAC and SBR encoder to build the eaacPlus encoder.

Although it does not have a transient detector, PS encoder calculates the stereo signal energy during the spatial parameter extraction. The calculation of signal energy is an important step in transient detection. This provides a motivation to propose a low-complexity transient detector that takes advantage of this calculation. Furthermore, it is possible to reduce the overall complexity of the eaacPlus encoder by replacing both SBR and AAC transient detectors with this single transient detector.

As an important issue in an audio encoder development, one might be

interested in the methods to measure or quantify the resulting audio quality. Conventionally, it can be evaluated subjectively by performing a listening test to compare the original and compressed audio, or objectively by using a model of human auditory system to approximate the subjective perception of the original and compressed audio. Objective audio quality evaluation methods consider the stereo audio channels independently. However, in the context of PS encoding, the evaluation of the stereo audio quality requires a consideration of the audio spatial image, hence the inter-channel relationships.

Subjective listening tests that focus on the spatial image quality can be designed and performed. But it is time-consuming to perform the test repeatedly during the development stage of the encoder, for example to compare the spatial image distortion as a result of varying certain PS encoding configuration. To the best knowledge of the author, an objective method to assess the stereo audio spatial image quality is not yet available. This has posed a problem for PS encoder development, and provided a motivation to look into an objective evaluation method of the audio spatial image.

1.3 Objectives and Scope of Research

The main objective of this thesis is to implement and optimize the PS encoding module in the `eaacPlus` encoder. The implementation takes into account the support of full MPEG-4 PS configurations as defined in [12]. The deliverable is a C-based floating-point software that can be used as a fully-functional encoder or further optimized for fixed-point platforms.

Motivated by the problem related to the downmixing and the possibility of complexity reduction, the optimizations target two areas:

- a stereo-to-mono downmixing scheme that is able to minimize the signal cancellation in the case of out-of-phase signal components, and
- a low-complexity transient detector for eaacPlus encoder that takes advantage of the signal energy calculation performed during PS spatial parameter extraction, and replaces both AAC and SBR transient detectors.

Due to the difficulties faced in analyzing the spatial image quality, it is necessary to come out with an objective evaluation method to evaluate spatial image distortion. The model would generally apply not only to the PS and spatial audio coding, but also to general audio processing. However, a thorough evaluation method would require an intensive research of up to psychoacoustic level. Hence the presentation of the method will be limited to a conceptual stage with a preliminary method evaluation.

1.4 Contributions of Thesis

The research, implementation and optimization works in this thesis have contributed to three conference publications and one submitted journal letter. In details, the contributions are presented as follows.

- Overview, theoretical backgrounds and MPEG-4 standard on Parametric Stereo coding is presented. The implementation of the encoder is elaborated and the optimization at software coding level is presented. The optimization targets the calculation of the phase parameter such that it can be calculated efficiently using only simple arithmetic. It is also shown how to calculate the *overall phase difference* (OPD) parameter prior to performing downmixing, hence a saving in memory requirement.

- A Matlab model of PS encoder – decoder is implemented. The model serves as a research platform for PS coding development. Possible optimizations can be implemented, tested and justified easily on the model before the actual software implementation.
- After identifying the possible problem of signal cancellation that arises during the stereo-to-mono downmixing, an enhanced downmixing scheme is proposed. The proposed subband-domain downmixing scheme is able to minimize the signal cancellation due to the phase differences of the stereo signals.
- The encoding complexity of eaacPlus encoder can be substantially lowered by removing the redundancy of the transient detector modules in the encoder. A unified low-complexity transient detector which operates in the PS encoder is proposed. The proposed detector is implemented to replace two redundant transient detectors in eaacPlus encoder. In overall, the optimization results in 4% reduction of eaacPlus encoder complexity.
- An objective evaluation of spatial image distortion is proposed as a concept. The verification of the proposal shows that the method is able to identify spatial narrowing and instability in the test audio. From this point, several recommendations are proposed to further refine the evaluation method such that it can be practically applied in future audio coding development work.

1.5 Outline of Thesis

This thesis is organized into seven chapters. In Chapter 2, overview and theoretical details on parametric stereo coding are presented. This chapter is followed by the implementation of the MPEG-4 compliant PS encoder and

Matlab-based PS model in Chapter 3. In addition, the efficient calculation of the phase parameter is elaborated.

The two chapters following Chapter 3 discuss the algorithm-level optimizations for the PS encoder. Chapter 4 starts by identifying the signal cancellation problem that arises with simple downmixing schemes. The proposed enhanced downmixing scheme and the impact on the quality improvement are subsequently presented. In Chapter 5, the proposed low-complexity transient detector is presented. The performance of the proposed transient detector is compared to both original AAC and SBR here. In addition, a complexity analysis and subjective listening test result of the overall optimization is presented.

Finally in Chapter 6 the concept of an objective evaluation of spatial image distortion is presented. The design and verification of the method, as well as the response of the model to the simulated spatial distortion is presented in details. Chapter 7 gives concluding remarks and recommendation for future research.

Chapter 2

Overview of Parametric Stereo Coding

Conventional perceptual audio coder encodes two stereo audio channels separately, consuming approximately half of the total encoding bitrate for each channel. One of the principles of the encoding is to remove redundancies between adjacent (or a group of) audio samples of each channel independently. However, redundancies in audio signal are not only present between adjacent samples; they can also be present across channels in the case of stereo or multi-channel audio. These redundancies have been exploited previously in audio coder by means of: *mid-side* (M/S) coding [17], *intensity stereo* (IS) coding [18], and *adaptive inter-channel prediction* [19]. And just recently, as a predecessor to the parametric stereo, *Binaural Cue Coding* (BCC) was introduced [20][21].

The idea behind PS is to code a stereo audio as a *monaural downmix* signal and a small amount of *spatial parameters* to describe the spatial image of the original audio. The monaural downmix can be encoded by any conventional perceptual audio coder while the spatial parameters are embedded into the mono audio bitstream. At the decoder, the decoded mono audio is expanded back into stereo audio using the transmitted parameters.

Conventionally, given an encoding bitrate of (x) kbps, each channel is encoded with a bitrate of approximately $(x/2)$ kbps. With the PS scheme, the total bitrate required to code the stereo audio is $(\lceil x/2 \rceil + s)$ kbps, where s is the

spatial parameters bitrate. The spatial parameters typically take about 1 – 3 kbps for the most basic PS configuration, and up to 8 kbps for high quality configuration [22]. Since the spatial parameter bitrate is relatively low, PS enables a higher coding efficiency. From another point of view, with the same encoding bitrate of (x) kbps, $(x - s)$ kbps is available to code the audio signal. This means that at low bitrate more bits are available to code just a single monaural downmix channel hence it suffers less bandwidth limitation and quality degradation. The result is a higher audio quality as compared to $(x/2)$ kbps for each individual stereo channel.

Figure 2.1 shows a generalized PS encoder and decoder as combined with a perceptual audio coder. It can be seen that the PS encoder and decoder are essentially a pre- and post-processing to the perceptual coder. Due to this nature, PS is independent of the perceptual audio coding scheme hence it can be combined with any mono audio coder. PS bitstream can be embedded in the ancillary or extension part of the mono audio bitstream for backward compatibility.

Figure 2.1: General model of audio encoder – decoder with a combination of perceptual audio coder and parametric stereo coder.

In this chapter, the psychoacoustic principles behind the PS coding scheme are elaborated. Following that, a detailed description of MPEG-4 PS standard and processing is presented. It must be noted that MPEG-4 PS standard defines the standard PS *bitstream structure* and *normative decoding* process. In addition, it provides an *informative encoding* scheme. There are many ways to implement and optimize the encoder, as long as the resulting bitstream format complies to the standard. This gives rooms to further enhance the audio quality by optimization of the encoding process.

2.1 Psychoacoustic Background

Parametric stereo coding was proposed based on the concept of exploiting the limitations of human auditory spatial perception. The works on PS embarked from a few concepts and findings from research in the area of psychoacoustic. The supporting theories and their implications to PS coding are presented in detail in [22] and summarized here.

Duplex Theory *Duplex Theory* by Lord Rayleigh (1907) states that sound source localization is facilitated by the *interaural intensity differences* at high frequencies and the *interaural time differences* at low frequencies [23]. These so-called binaural cues give cue to the location of the sound source. This implies that spatial image can be parameterized as a set of binaural cue parameters.

Limitation of auditory spatial resolution Human auditory system has a limited spatial resolution. As the binaural cues are synthesized through headphone, a listener perceives the sound to be located inside the head at the line connecting the left and right ear. To move the position of the sound, the

cues have to be changed by a certain amount. This implies that it is possible to quantize the binaural cue parameters.

Frequency-dependent spatial cues Binaural cues are rendered in a set of non-linear frequency bands, with finer resolution at lower frequencies and increasing bandwidth at higher frequencies. The bandwidth of the bands follows the *equivalent rectangular bandwidth* (ERB) [27] which models the bandwidth of the auditory filters. This implies a frequency-dependent binaural cue parameters analysis.

Binaural sluggishness The phenomena of *binaural sluggishness* reveals the limitation of auditory system in tracking the changes of binaural cues, with a time constant between 30 and 100 ms. This implies a limited temporal resolution, allowing limited binaural cue parameters update rate.

Perception of spatial diffuseness Despite binaural sluggishness phenomena, a fast change in binaural cues leads to the perception of *spatial diffuseness* which relates to the spatial width of the audio. And commonly, audio recording contains a certain degree of *spaciousness* or *compactness* introduced by room echo and reverberation during recording which gives impression of a wide or narrow audio image. It has been demonstrated that spatial diffuseness mostly depends on the *coherence* (cross-correlation) of the two stereo channels. This implies that additional parameter is needed to describe diffuseness or compactness of the original audio.

Based on these psychoacoustic findings, PS scheme describes spatial image of audio by three *spatial parameters*. The parameters are *inter-channel intensity difference* (IID), *inter-channel phase difference* (IPD) and *inter-channel coherence* (ICC). IID and IPD describe the spatial location of the source, while ICC aims to

describe the spatial width of the original audio. The parameters are extracted in a set of non-uniform frequency bands where each band has its own set of parameters. The temporal resolution is in the order of tens of milliseconds, associated to the time constant of binaural sluggishness.

2.2 MPEG-4 Parametric Stereo Encoding

The diagram of the PS encoder as suggested in the MPEG-4 PS standard [12] is shown in Figure 2.2. The encoding starts by transforming a block of stereo audio samples $l(n)$ and $r(n)$ into time – frequency subband signals $l(k,n)$ and $r(k,n)$. A set of spatial parameters is calculated and quantized. The quantized parameters are then assembled as PS bitstream. To obtain the monaural signal $m(n)$, a stereo-to-mono downmixing is performed on the time-domain stereo samples.

Figure 2.2: Informative PS encoding process as suggested by MPEG-4 PS standard.

2.2.1 Hybrid Analysis Filtering

The first stage of PS encoding is to transform the input signal into the subband domain such that the parameters can be analyzed in a set of non-uniform frequency bands which resembles the frequency decomposition in human auditory system. Initially, PS was developed using *fast fourier transform* (FFT-) based frequency analysis. However, it was shown that *quadrature mirror filter* (QMF-) based subband filtering can be efficiently implemented with lower complexity and it results in better PS audio quality performance [16]. Furthermore for mobile devices, it is desirable to minimize the complexity of the processing (eg. to save battery power). Hence, MPEG-4 adopted the QMF-based PS in its standard.

The QMF filtering is followed by *low-frequency filtering* to increase the frequency resolution of the lower subbands. This combined filtering scheme and the resulting output subbands are referred to as *hybrid analysis filtering* and *hybrid subbands*, respectively. The hybrid filter structure for the left stereo audio channel is illustrated in Figure 2.3. An explanation of the hybrid analysis filtering follows.

Figure 2.3: Hybrid analysis filtering structure for PS encoder, which is a combination of complex-modulated QMF filterbank and low-frequency filters for increased frequency resolution.

QMF Analysis Filtering

A PS frame consists of 2048 non-overlapping time-domain samples per channel. The samples are fed into a 64-channel *complex-exponential modulated* QMF filterbank. The impulse response $h_k(n)$ of the filterbank is defined as [24]

$$h_k[n] = p_o[n] \exp\left\{j \frac{\pi}{4} (2k+1)(2n-1)\right\} \quad (2.1)$$

where $p_o[n]$ is the prototype filter, k is the QMF subband channel index, $n = 0, \dots, N_q - 1$ is the time index where $N_q = 640$ is the length of the prototype filter, and j denotes the complex operator $\sqrt{-1}$.

The output of the filterbank are subsequently downsampled by a factor of 64, resulting in a time - frequency signals of 32×64 complex-subband samples. Due to the complex representation of the filterbank output, the subbands are effectively oversampled by a factor of two. This reduces aliasing and allows flexible signal modification such as phase calculation and modification [25].

Low-frequency Filtering

To account for the finer frequency resolution of human auditory system at low frequencies, the first few lowest QMF subbands are further filtered by means of *oddly-modulated M^{th} band low-frequency* filterbanks [26] to increase the frequency resolution. There are 4 prototype low-frequency filters defined in the standard: 12-, 8-, 4-, and 2- *sub-subband* channels filters.

Two type of filters: *Type A* and *Type B* are introduced and the impulse responses $G_{k,m}[q]$ of sub-subband channel m of QMF subband channel k are defined as [12]

$$G_{k,m}^{\text{TypeA}}[q] = g_k[q] \exp\left\{j \frac{2\pi}{M_k} \left(m + \frac{1}{2}\right)(q-6)\right\} \quad (2.2)$$

$$G_{k,m}^{\text{TypeB}}[q] = g_k[q] \cos\left\{\frac{2\pi}{M_k} m(q-6)\right\} \quad (2.3)$$

where q is the time index, $g_k(q)$ is the prototype filter and M_k is the number of low-frequency filters at the corresponding QMF frequency channel ($m = 0, \dots, M_k - 1$). The order of these filters is 12, hence the low frequency filtering results in a delay of six QMF subband samples. To compensate for this additional delay, the remaining QMF subband which are not low-frequency filtered are delayed by six subband samples. In this thesis the label k and n are subsequently used to refer to the frequency and time index of the hybrid subband, respectively.

2.2.2 Spatial Parameter Extraction

The first step in the spatial parameter extraction is to group the hybrid subbands non-uniformly in frequency such that the bandwidth bw approximates the ERB according to [27]

$$bw = 24.7(0.00437f_c + 1) \quad (2.4)$$

where f_c is the center frequency of the corresponding hybrid subband group, hereby referred to as *stereo band* and indexed by b . MPEG-4 PS standard defines three configurations: 10, 20 and 34 stereo-band configurations which correspond to the number of the non-uniform frequency groups. Each configuration differs in their low-frequency filtering as summarized in Table 2.1.

Table 2.1: MPEG-4 PS stereo-band configuration and the associated number of frequency channels of the low frequency filtering.

QMF subband	34 stereo-band configuration		10, 20 stereo-band configuration	
	M_k	Type	M_k	Type
0	12	A	8 (grouped into 6)	A
1	8		2	B
2	4		2	
3	4		-	
4	4		-	-
5 – 63	-	-	-	-
Total =	91 hybrid bands		71 hybrid bands	

10, 20 Stereo-band Configurations For this configuration, the low-frequency filtering is applied to the three lowest QMF subbands, resulting in 8, 2, and 2 sub-subbands respectively. The first eight sub-subbands are further grouped into six frequency channels. When combined with the remaining QMF subbands, a total of 71 hybrid subbands are available: $l(k,n)$, $r(k,n)$ where $0 \leq k < 70$ and $0 \leq n < 31$. The hybrid subbands are then grouped non-uniformly into 20 stereo bands ($0 \leq b < 20$). For 10 stereo bands configuration, the hybrid filtering and frequency grouping follows the 20 stereo bands grouping, however the calculated parameter during parameter extraction is mapped from 20 to 10 parameters.

34 Stereo-band Configuration For this configuration, the lower frequency filtering is applied to the first five QMF subbands, resulting in 12, 8, 4, 4, and 4 sub-subbands respectively. A total of 91 hybrid subbands ($0 \leq k < 91$) are subsequently grouped into 34 stereo bands ($0 \leq b < 34$).

Following the frequency grouping, three main spatial parameters are extracted at each stereo band. The standard allows the parameters to be updated up to four times per frame, and the set of parameters related to each of

the update position is referred to as '*envelope*' in the standard. This can be viewed as a sub-framing of the spatial parameter calculation. The envelope is assigned to the last subband timeslot of the corresponding parameter sub-frame, which is referred to as the '*border position*'. The border positions can be distributed uniformly or variably within the 32 stereo band timeslot, hence enabling a flexible parameter update time resolution.

In the case of uniform sub-framing, 3 sub-frame configurations are allowed: 1, 2, and 4 sub-frames per frame. In the case of variable sub-framing, 4 sub-frame configurations are allowed: 1, 2, 3, and 4 sub-frames per frame. The sub-framing concept is illustrated in Figure 2.4 for a 34 stereo-band configuration. Here, the frame is divided into 4 sub-frames with a variable border positioning, whereby the border positions are indicated by the shaded slots. The spatial parameters as derived from psychoacoustic findings are described and calculated as follows.

Figure 2.4: Illustration of the PS sub-framing, which divides a frame into 4 sub-frames with variable envelope positioning.

Inter-channel Intensity Difference (IID) This parameter describes the signal level difference between the left and right channel and it is calculated as

$$IID_e(b) = 10 \log_{10} \frac{\sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} |l(k,n)|^2}{\sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} |r(k,n)|^2} \quad (2.5)$$

The summation over k corresponds to the non-uniform grouping of the hybrid subbands into stereo bands, where k_b denotes the hybrid subband boundary of stereo band b . The summation over n determines the spatial parameter update rate where n_e denotes the border position of parameter envelope e .

Inter-channel Coherence (ICC) This parameter describes the coherence between the two audio channels, which is defined as the normalized cross-correlation. It is related directly to spatial width of the original audio and calculated as

$$ICC_e(b) = \frac{\left| \sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} l(k,n) \cdot r^*(k,n) \right|}{\sqrt{\sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} |l(k,n)|^2 \cdot \sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} |r(k,n)|^2}} \quad (2.6)$$

where ' $*$ ' denotes complex conjugation.

Inter-channel Phase Difference (IPD) This parameter describes the phase difference between the left and right channel, and it is only calculated for frequency band of up to about 2 kHz ($b = 5, 11, 17$ for 10, 20, and 34 stereo bands configuration respectively). For frequencies above 2 kHz, psychoacoustic

research reveals that human auditory system is insensitive to inter-aural fine structure phase difference [22][23]. IPD is calculated as

$$IPD[b] = \angle \left(\sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} l(k, n) \cdot r^*(k, n) \right) \quad (2.7)$$

where $\angle(\sigma)$ denotes the angle of the complex-sample σ .

Overall Phase Difference (OPD) An additional parameter is calculated to describe the relative phase distribution between the mono and the left channel. It is transmitted together with IPD such that the decoder is able to distribute the IPD correctly between the left and right channel during stereo reconstruction. The decoder applies a phase shift equal to the OPD to reconstruct the phase of the left channel from the decoded mono signal and a phase shift equal to the OPD minus the IPD to reconstruct the phase of the right channel from the decoded mono signal. OPD is calculated as

$$OPD[b] = \angle \left(\sum_{n=n_{e-1}+1}^{n_e} \sum_{k=k_b}^{k_{b+1}-1} l(k, n) \cdot m^*(k, n) \right) \quad (2.8)$$

The standard allows the encoder to enable or disable the phase parameter (IPD and OPD) transmission in the bitstream for bit saving.

When more than one envelope per frame is calculated, the outer summation boundary for Equations (2.5) to (2.8) is changed accordingly over the corresponding sub-frame boundaries. Three additional information needs to be provided in the bitstream to inform the decoder: the number of envelopes, a bit indicating whether the envelope is distributed uniformly or variably within the frame, and in the case of variable sub-framing, a matrix which contains the border positions.

2.2.3 Quantization and Bitstream Formatting

After the parameter extraction, each parameter value is quantized and differentially coded to increase the coding efficiency. Subsequently, Huffman coding is applied. The standard defines quantization grids and Huffman tables for each of the spatial parameter. The entropy coded parameters and PS configuration are then assembled into the mono audio bitstream.

2.2.4 Stereo to Mono Downmixing

The standard suggests a time-domain averaging to mix the stereo signal into monaural signal according to

$$m(n) = \frac{l(n) + r(n)}{2} \quad (2.9)$$

where $m(n)$ denotes the time-domain monaural signal samples. This time domain monaural signal is subsequently passed on to a generic perceptual audio coder.

2.3 MPEG-4 Parametric Stereo Decoding

The PS decoding process is shown in Figure 2.5. The decoding process is beyond the scope of this thesis. However, it is summarized in this section to give a more complete picture on the PS coding scheme.

Figure 2.5: MPEG-4 Parametric Stereo decoding process.

2.3.1 Analysis Filtering and Decorrelation

The decoded mono audio $m'(n)$ is first filtered by the same hybrid analysis filtering scheme as described for the encoder. The mono subband signals $m'(k,n)$ are then decorrelated to produce the side signals $d(k,n)$. These side signals aim to introduce certain coherency to the reconstructed stereo audio to approximate the spatial width of the original audio. The decorrelation is carried out by applying *infinite impulse response (IIR) all-pass reverberator* filter to the mono audio. To reduce complexity, the IIR all-pass filter is only applied to frequency bands up to 8.625 kHz [28]. Simple constant delay reverberation is applied to the remaining upper frequency bands. The IIR all-pass reverberator is a cascade of a fractional delay line and three all-pass filter links. The frequency and unwrapped phase response of the IIR decorrelation filter is shown in Figure 2.6.

2.3.2 Stereo Synthesis

After PS bitstream decoding, the decoded parameters are used to reconstruct the stereo audio. Each parameter envelope is defined at its border position as obtained from the transmitted envelope position. To prevent blocking artifacts,

the parameter values for the remaining slot within the corresponding parameter sub-frame are linearly interpolated from the previous envelope.

Figure 2.6: Frequency and phase response of IIR all-pass decorrelation filter for 34 stereo bands configuration.

The IID is first transformed into scale factor $c(b)$ according to

$$c(b) = 10^{\frac{IID(b)}{20}} \quad (2.10)$$

The stereo subband signals are reconstructed as

$$\begin{bmatrix} l'(k, n) \\ r'(k, n) \end{bmatrix} = \begin{bmatrix} H_{11}(k, n) & H_{21}(k, n) \\ H_{12}(k, n) & H_{22}(k, n) \end{bmatrix} \begin{bmatrix} m'(k, n) \\ d(k, n) \end{bmatrix} \quad (2.11)$$

where the matrix elements $H_{xy}(k,n)$ are calculated from the spatial parameter. H_{11} and H_{12} are applied to the mono downmix subband to reconstruct the level difference of the original left and right channel. H_{21} and H_{22} are applied to the decorrelated signal to control the amount of coherency introduced to approximate the spatial width of original audio.

MPEG-4 PS standard defines two different mixing procedures to calculate the intermediate mixing matrix elements $h_{xy}(k,n)$: mixing procedure R_a and R_b . These procedures are described in Table 2.2. If phase parameter is disabled, the mixing matrix elements $H_{xy}(k,n) = h_{xy}(b(k))$ where $b(k)$ is the inverse mapping from the stereo band index to the hybrid subband index. If phase parameter is enabled, $H_{xy}(k,n)$ is calculated by applying phase rotation to $h_{xy}(b(k))$ according to

$$H_{11}(k,n) = h_{11}(b(k)).\exp(j\varphi_1(b(k))) \quad (2.12)$$

$$H_{12}(k,n) = h_{12}(b(k)).\exp(j\varphi_2(b(k))) \quad (2.13)$$

$$H_{21}(k,n) = h_{21}(b(k)).\exp(j\varphi_1(b(k))) \quad (2.14)$$

$$H_{22}(k,n) = h_{22}(b(k)).\exp(j\varphi_2(b(k))) \quad (2.15)$$

where $\varphi_1(b) = \varphi_{OPD}(b)$ and $\varphi_2(b) = \varphi_{OPD}(b) - \varphi_{IPD}(b)$ are the phase distribution calculated from IPD and OPD parameters. Finally, hybrid synthesis filtering is applied to the reconstructed subbands to obtain the stereo-reconstructed audio $l'(n)$ and $r'(n)$.

Table 2.2: Mixing procedures defined in MPEG-4 Parametric Stereo standard.

Mixing Procedure R _a	Mixing Procedure R _b
$c_1(b) = \frac{\sqrt{2}}{\sqrt{1+c^2(b)}} \quad c_2(b) = \frac{\sqrt{2}c(b)}{\sqrt{1+c^2(b)}}$ $\alpha(b) = \frac{1}{2} \arccos(ICC(b))$ $\beta(b) = \alpha(b) \frac{c_1(b) - c_2(b)}{\sqrt{2}}$	$\alpha(b) = \frac{1}{2} \arctan\left(\frac{2c(b)ICC(b)}{c^2(b)}\right)$ $\mu(b) = 1 + \frac{4ICC^2(b) - 4}{(c(b) + c^{-1}(b))^2}$ $\gamma(b) = \arctan\left(\sqrt{\frac{1 - \sqrt{\mu(b)}}{1 + \sqrt{\mu(b)}}}\right)$
$h_{11}(b) = \cos(\alpha(b) + \beta(b))c_2(b)$ $h_{12}(b) = \cos(\beta(b) - \alpha(b))c_1(b)$ $h_{21}(b) = \sin(\alpha(b) + \beta(b))c_2(b)$ $h_{22}(b) = \sin(\beta(b) - \alpha(b))c_1(b)$	$h_{11}(b) = \sqrt{2} \cos(\alpha(b)) \cos(\gamma(b))$ $h_{12}(b) = \sqrt{2} \sin(\alpha(b)) \cos(\gamma(b))$ $h_{21}(b) = -\sqrt{2} \sin(\alpha(b)) \sin(\gamma(b))$ $h_{22}(b) = \sqrt{2} \cos(\alpha(b)) \sin(\gamma(b))$

Chapter 3

Software Implementation of MPEG-4 Parametric Stereo Encoder

As the main objective of this thesis, PS encoder complying to the MPEG-4 PS standard is implemented as a fully-functioning software. Two implementations are presented: C-based PS encoder as part of the eaacPlus encoder, and Matlab-based PS encoder – decoder model that serve as testing and research platform. This chapter presents in detail the encoder implementation, starting by introducing the eaacPlus encoder and moving on to the practical aspect of the PS encoder development.

3.1 Overview of Enhanced aacPlus Encoder

Enhanced aacPlus is a coding scheme which combines perceptual and parametric coding concept. It is a combination of AAC [29], SBR, and PS. AAC is considered as the state-of-the-art perceptual audio coder, while SBR and PS are audio coding tools which parameterize the contents of the audio signal.

Figure 3.1 shows the general structure of enhanced aacPlus encoder. The original stereo audio is first coded by the PS encoder, which parameterizes the spatial image of the audio signals and subsequently performs downmixing into a monaural signal. The parameter extraction and downmixing is performed in the complex-subband domain. The monaural subband signal is passed to the SBR encoder, which parameterizes the upper bandwidth portion of the monaural signal based on the lower counterpart. The band-limited monaural subband signal is then synthesized back into the time domain and passed to the AAC encoder to be perceptually coded.

Figure 3.1: General structure of enhanced aacPlus encoder.

3GPP implemented a baseline version of PS in its eaacPlus encoder, which has a very limited functionality as compared to the full functionality as defined in the MPEG-4 PS standard. This baseline version only supports 10 and 20 stereo-band configurations, does not enable phase parameter, and it does not support sub-framing for increased parameter update rate.

Chapter 3. Software Implementation of MPEG-4 PS Encoder

Figure 3.2: Top - down enhanced aacPlus encoder software module tree.

Figure 3.2 shows the top – down eaacPlus encoder software processing flow. At the top of the module tree is the *main* function of the software. It performs initialization and configures the SBR and AAC encoder according to the user input: input audio file, output bitstream file, encoding bitrate, and encoding mode (mono or stereo). The supported input audio file is *PCM wave* (.wav) file and the output bitstream file is in the 3GPP (.3gp) format. The frame looping consists of reading a frame of audio samples (2048 samples per channel) followed by SBR, PS and AAC encoding, and finally writing the frame bitstream into the output file. In the module tree, the solid boxes denote modules that are related to the developed PS encoder which are explained in the following section.

3.2 C-based MPEG-4 Parametric Stereo Encoder Implementation

The implementation structure of the MPEG-4 PS encoder is shown in Figure 3.3. As opposed to the suggested structure in Figure 2.2, the downmixing is performed in the subband domain instead of time domain.

Figure 3.3: Implementation structure of PS encoding in eaacPlus encoder.

3.2.1 Hybrid Analysis Filtering

As PS and SBR works in the same complex QMF subband domain, this combination gives advantage in terms of processing complexity. The analysis filtering is carried out prior to PS parameter extraction and after the mono downmixing, the mono subband signals are directly used by the SBR encoder without additional synthesis – analysis filtering step. Following QMF analysis filtering, low-frequency filtering is performed on the few lowest subbands to increase the frequency resolution. Figure 3.4 illustrates the hybrid filtering process for 34 stereo-band configuration graphically.

Figure 3.4: Hybrid analysis filtering scheme for 34 stereo-band configuration in PS encoder.

The computation-intensive complex modulation in the filterbank is decomposed into fundamental *discrete cosine transform* (DCT) and *discrete sine transform* (DST) where fast algorithms are available [30]. Type A low-frequency filters are implemented similar to the implementation in FAAD2 decoder [31], whereby the same decomposition into fundamental DCT is used. Type B low-frequency filter, which is a real cosine-modulated filter, is implemented simply by convolution.

3.2.2 Spatial Parameter Extraction

The parameter extraction is performed at each sub-frame and stereo band. The number of sub-frames in the corresponding frame is indicated by the variable *num_env*. The sub-frame border positions are calculated as:

$$border_position(e) = \left\lfloor \frac{32 \cdot (e+1)}{num_env} \right\rfloor - 1, \text{ where } e = 0, \dots, num_env - 1 \quad (3.1)$$

These parameters are calculated according to Pseudocode 3.1. Here, *number_of_stereo_bands* refers to 10, 20, or 34 stereo bands depending on the stereo bands configuration, and *number_of_phase_bands* refers to the number of lower stereo bands where the phase parameters are calculated. The function *quantise_phase(a, b)* performs direct quantization of the phase parameter which is explained in detail in Section 3.2.5.

Chapter 3. Software Implementation of MPEG-4 PS Encoder

```

env = 0;
for (n = 0; n < 32; n++) {
    for (b = 0; b < number_of_stereo_bands; b++) {
        
$$e_l(b) = \sum_{k=k_l(b)}^{k_h(b)} |l(k, n)|^2 \quad (3.2)$$


$$e_r(b) = \sum_{k=k_l(b)}^{k_h(b)} |r(k, n)|^2 \quad (3.3)$$


$$e_R(b) = \sum_{k=k_l(b)}^{k_h(b)} l(k, n)r^*(k, n) \quad (3.4)$$

    }

    if ((n == border_position(env) ) & (env < num_env)){
        for (b = 0; b < number_of_stereo_bands; b++) {
            
$$IID(env, b) = 10 \log_{10} \frac{e_l(b)}{e_r(b)} \quad (3.5)$$

            
$$ICC(env, b) = \frac{|e_R(b)|}{\sqrt{e_l(b)e_r(b)}} \quad (3.6)$$


            if (b < number_of_phase_bands) {
                IPD(env, b) =
                    quantise_phase(re(e_R(b)), im(e_R(b)))
                OPD(env, b) =
                    quantise_phase(e_l(b)+re(e_R(b)), im(e_R(b)))
            }
            reset e_l, e_r, e_R to 0
            env++
        }
    }
}

```

Pseudocode 3.1: PS spatial parameter extraction.

3.2.3 Stereo to Mono Downmixing

MPEG-4 PS standard suggests a simple averaging in the time domain to mix the stereo audio signal into a monaural signal. However, this method might result in signal cancellation or coloration, for example when performing the downmixing on stereo audio which are not mono-compatible or stereo audio with anti-phase signal components. In the implementation, the downmixing process is performed in hybrid subband domain according to

$$m(k, n) = \frac{l(k, n) + r(k, n)}{2} \cdot \gamma(k, n) \quad (3.7)$$

where $m(k, n)$ is the monaural hybrid subband sample and $\gamma(k, n)$ is the stereo scale factor to ensure overall power preservation, defined as:

$$\gamma(k, n) = \sqrt{\frac{|l(k, n)|^2 + |r(k, n)|^2}{0.5 \cdot |l(k, n) + r(k, n)|^2}} \quad (3.8)$$

To comply with the PS decoding process [12], the stereo scale factor is defined such that the power of the mono signal is half the total power of the stereo signals. It is limited to 6 dB ($\gamma(k, n) = 2$) to prevent artifacts resulting from a large gain when the attenuation of the power of the sum signal is significant [32].

With this subband domain downmixing scheme, signal coloration is minimized. However, it can still possibly happen due to the heavy phase cancellation when the stereo signals are very much anti-phase. Further optimization of the downmixing scheme will be discussed in detail in Chapter 4.

3.2.4 Bitstream Formatting

In enhanced aacPlus, the bitstream scheme follows the syntax defined by MPEG-4 for the audio bitstream. SBR bitstream is put at the extension part of the AAC data. PS bitstream is appended to the extension part of SBR bitstream. In this way compatibility is ensured, whereby decoders that do not support SBR or PS will not read the extension data and only decodes the monaural AAC bitstream. The bitstream structure is illustrated in Figure 3.5.

Figure 3.5: Bitstream structure of enhanced aacPlus codec.

The bitstream formatting starts with quantization of the spatial parameters. MPEG-4 PS standard defines two quantization grids for IID (15-level for coarse and 31-level for fine quantization grid), and 8-level quantization grid for the ICC and phase parameters (IPD and OPD). Differential coding is then applied to the quantized data: over time (differential from the previous envelope) and over frequency (differential from the previous stereo band of the same

envelope). Whichever results in less number of bits will be chosen and subsequently Huffman-coded. The Huffman codebook for time and frequency differential coding are provided in the standard. A flag is then transmitted to inform the decoder whether the differential coding is performed over the time or frequency direction.

The Huffman-coded data is subsequently assembled to the PS bitstream. The bitstream syntax is defined in the PS standard [12]. It contains header, data, and extension bitstream. The header consists of the PS configuration for the corresponding frame: number of stereo bands, number of sub-frames and sub-frame distribution (uniform/variable). The data part consists of the IID and ICC data. The phase parameters are placed at the extension part of the PS bitstream. The extension can be enabled or disabled by setting the *enable_ipdopd* flag.

3.2.5 Encoding Optimizations

Direct Phase Quantization

Phase of a complex number can be calculated by taking the inverse tangent of the imaginary part divided by the real part. In PS encoder, IPD and OPD are the phase difference between left – right and left – mono channels respectively, and they can be calculated by inverse tangent of the complex subband sample values as well. However, in this implementation, the phase calculation and quantization are combined into one step: *direct quantization*, taking advantage of the trigonometry properties and the quantization grid. The direct quantization method requires only simple arithmetic and logic comparison which is advantageous for fixed-point implementation. This is implemented in the function *quantise_phase(real, imaginary)*.

PS standard defines an 8-level phase quantization grid which is at an increment of $\pi/4$ radian as presented in Table 3.1. The direct quantization is illustrated by looking at the quantization in the first quadrant in Cartesian

coordinate system. The first quadrant consists of quantization level 0, $\pi/4$, and $\pi/2$ radian. For a uniform quantization, any angle closest to a level will be quantized to that level.

Table 3.1: Phase quantization grid defined in MPEG-4 Parametric Stereo standard.

Index	0	1	2	3	4	5	6	7
Representation level	0	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3\pi}{4}$	π	$\frac{5\pi}{4}$	$\frac{3\pi}{2}$	$\frac{7\pi}{4}$

(a)

(b)

Figure 3.6: Direct phase quantization for: (a) a complex-sample located between quantization level 0 and 1; (b) a complex-sample located between quantization level 1 and 2.

For any complex sample $x(re, im)$ in the first quadrant between quantization level 0 and 1 (see Figure 3.6(a)), the quantization is carried out following Pseudocode 3.2. If the complex sample $x(re, im)$ lies in the first quadrant between quantization level 1 and 2 (see Figure 3.6(b)), the quantization is carried out following Pseudocode 3.3. Similarly for any complex number located in another quadrant, first the point is rotated such that it is located in the first quadrant and quantized accordingly. Subsequently, a quantization level offset corresponding to the quadrant is added to the basic quantized level to obtain the final quantized phase.

```

if im ≤ re * tan( $\pi/8$ )
    quantization_level = 0
else
    quantization_level = 1

```

Pseudocode 3.2: Direct phase quantization when the complex sample is located between quantization level 0 and 1.

```

if re ≥ im * tan( $\pi/8$ )
    quantization_level = 1
else
    quantization_level = 2

```

Pseudocode 3.3: Direct phase quantization when the complex sample is located between quantization level 1 and 2.

The direct quantization method is compared to a conventional phase calculation and quantization method, where the phase is calculated by taking an inverse tangent and subsequently quantized by assigning the closest quantization level

to the calculated angle. A set of 41 complex sample points are generated with a phase variation of 0 to 2π rad, as shown in Figure 3.7(a). The quantized phase with the conventional and direct quantization methods are shown in Figure 3.7(b). It can be seen that both methods give identical quantization result.

Figure 3.7: Comparison of the conventional and direct quantization method: (a) complex sample points ranging from 0 to 2π radians, (b) quantized level of the corresponding sample points.

OPD Calculation

In software implementation (especially for firmware implementation), memory space is often limited. It is desirable to efficiently design the memory distribution for various buffers in the software system. For PS encoder a small amount of memory can be saved by performing *in-place* downmixing process,

which is explained as follows. It can be seen that after downmixing, basically the original left- and right-channel hybrid subbands are not required anymore and only the monaural subbands are subsequently utilized by SBR and AAC encoder. Instead of allocating a separate buffer, the monaural hybrid subbands can be overwritten to the left-channel hybrid subbands buffer.

OPD is defined as the phase difference between left channel and the mono downmixed channel. It means that to calculate OPD, the downmixing has to be carried out prior to parameter extraction. In this case, an additional buffer is required to store the downmixed subbands as the left-channel hybrid subband values are still needed for the parameter extraction and hence can not be overwritten.

In this implementation, the OPD calculation is carried out prior to downmixing, saving the memory space required for the additional monaural hybrid subband buffer. This is possible by looking at Equation (2.8) and (3.7). Taking the left and right complex-subband sample as $l = re_L + j \cdot im_L$ and $r = re_R + j \cdot im_R$ respectively, the monaural subband samples according to Equation (3.7) can be rewritten as

$$m = \frac{(re_L + re_R) + j \cdot (im_L + im_R)}{2} \quad (3.9)$$

where the stereo scale factor, frequency and time index has been omitted for clarity. Substituting Equation (3.9) to Equation (2.8) and again omitting the frequency and time index, OPD can be calculated as follows:

$$\begin{aligned} OPD &= \angle(l(k, n)m^*(k, n)) \\ &= \angle\left(\frac{1}{2}(re_L + j \cdot im_L)\{(re_L + re_R) - j \cdot (im_L + im_R)\}\right) \\ &= \angle\left(\frac{1}{2}\{(re_L^2 + im_L^2 + re_L re_R + im_L im_R) + j \cdot (im_L re_R - re_L im_R)\}\right) \end{aligned}$$

$$= \angle \left(\frac{1}{2} \{ (e_l + \text{real}(e_R)) \} + j \cdot \text{imag}(e_R) \right) \quad (3.10)$$

where e_l and e_R are calculated in Equation (3.2) and (3.4), $\text{real}(a)$ denotes real part of a , and $\text{imag}(a)$ denotes imaginary part of a . It can be observed that the OPD value can be calculated without having to perform the downmixing beforehand.

Figure 3.8: Testing of OPD calculation optimization: (a) inter-channel phase difference profile of the test signal, (b) OPD calculated from left - mono subband signal, (c) OPD calculated from left - right subband signal.

Figure 3.8(a) shows the inter-channel phase difference profile of a 800 Hz stereo sinusoidal signal, which varies from 0 to 2π . Figure 3.8(b) and (c) shows the unquantized OPD of the signal calculated by using the left – mono signal (Equation 2.8) and left – right signal (Equation 3.10), respectively. It can be seen that the optimized OPD calculation gives the same calculated values as the normal OPD calculation.

3.3 Matlab-based MPEG-4 Parametric Stereo Model

In addition to the C-based PS encoder development, a Matlab model of PS encoder – decoder was developed. The purpose of this model is to serve as a testing and research platform to support the encoder development. Using the Matlab model, new ideas or additional processing can be implemented quickly and the results and data can be analyzed more easily.

As the C-based *eaacPlus* is a combination of PS, SBR and AAC encoder, it is difficult to analyze the audio quality related to the PS processing. The reason is because the decoded audio would contain artifacts and distortion not only from the PS, but also the SBR and AAC processing. The implemented Matlab model consists of solely PS processing without the presence of the SBR and mono audio coder. Hence, the artifacts and quality resulting from PS can be justified for research and optimization purposes.

The structure of the Matlab PS model is shown in Figure 3.9. The model supports all encoding configurations described in the standard. For ease of usage, testing and data analysis, the model is built as a collection of modules as shown in Figure 3.10. Following the figure, an overview of the the main modules are provided. The notations being used for the explanation are standard Matlab notations.

Figure 3.9: Matlab-based parametric stereo encoder – decoder model. Here, the monaural hybrid subband signals are directly passed to the decoder for stereo synthesis.

Figure 3.10: Modules in the Matlab-based PS model.

3.3.1 Data Structure

A collection of global data structure is defined in an m-file *structure.m*. The global data structure contains global variables to store data that can be accessed

by all of the modules. The main global variables are listed and explained as follows.

- PS_ENC: a structure that holds the encoding configurations and extracted spatial parameters.
- QMF_ANALYSIS_LEFT, QMF_ANALYSIS_RIGHT: structures that hold the stereo QMF subband signals and the filter delay lines.
- HYBRID_ANALYSIS_LEFT, HYBRID_ANALYSIS_RIGHT: structures that hold the low-frequency filtered subband signals and the low-frequency filter delay lines.
- QMF_MIX, HYBRID_MIX: global variables that hold the monaural hybrid subband signals.
- QMF_DECODED_LEFT, QMF_DECODED_RIGHT: structures that hold the stereo QMF subband signals which are reconstructed by the PS decoder.

3.3.2 Main modules

ps_model

This module is the main encoding – decoding function of the model. The syntax of the function is

$$[audio_out] = ps_model(input_vector, audio_in)$$

where *input_vector* is a vector that contains the encoding configuration, *audio_in* and *audio_out* are the input and output stereo signal in *PCM wave* (.wav) format, respectively. The function performs the top level operations in the order given as follows:

- reading of input stereo audio and encoding configurations,
- QMF analysis filtering,
- encoder – decoder framing and looping,
- QMF synthesis filtering,
- writing to output stereo audio file.

The encoding configurations are saved into the configuration buffer in the global structure *PS_ENC*.

qmf_analysis_filtering

This is the QMF analysis filtering module, which decomposes the input stereo signal into 64 complex-subband signals. The syntax of the function is

$$qmf_analysis_filtering(timeInL, timeInR, coeff)$$

where *timeInL* and *timeInR* are the left and right channel time-domain stereo signals respectively, and *coeff* is the prototype filter coefficients. In this Matlab-implementation, the output of the analysis filtering is stored in the variables *QMF_ANALYSIS_LEFT.subband()* and *QMF_ANALYSIS_RIGHT.subband()* such that these subband signals can be easily accessed by the encoding module.

EncodePsFrame

This is the main encoding routine for a frame of input stereo signals. This function does not have any input and output arguments. The function access the global variables *QMF_ANALYSIS_LEFT.subband()* and *QMF_ANALYSIS_RIGHT.subband()* to retrieve the subband signals and perform the low-frequency filtering. Subsequently the PS encoding is performed based on the encoding configuration available in *PS_ENC*. The extracted spatial parameters are quantized and saved in the parameter data buffer in *PS_ENC*. Additionally, downmixing is performed on the stereo hybrid

subband signals. The monaural hybrid subband signals are stored in QMF_MIX and HYBRID_MIX.

SynthesizeStereo

This function also does not have input and output arguments. It retrieves the spatial parameter from PS_ENC and the monaural hybrid subband signals from QMF_MIX and HYBRID_MIX. Signal decorrelations are performed, followed by the stereo reconstruction. The synthesized stereo hybrid subband signals are subsequently stored in the signal buffer in QMF_DECODED_LEFT and QMF_DECODED_RIGHT.

qmf_synthesis_filtering

After stereo synthesis, the hybrid subband signals are reconstructed back into time-domain stereo signals. The syntax of this function is

$$[synthL, synthR] = qmf_synthesis_filtering_dec(coeff)$$

where *synthL* and *synthR* are the time-domain stereo signal, and *coeff* is the prototype filter which is identical to the one in *qmf_analysis_filtering*. The module retrieves the synthesized stereo hybrid subband signals from QMF_DECODED_LEFT and QMF_DECODED_RIGHT.

For a more detail look into the model and source code, a complete listing of the source code and data structure is provided in [33].

Chapter 4

Optimized Downmixing Scheme for Parametric Stereo Encoder

In PS coding paradigm, the signal content of the original stereo audio is solely represented in the monaural signal. Consequently, the downmixing process has to preserve all of the signal components. However, when the stereo-to-mono downmixing is performed, signal coloration and unwanted attenuation might occur due to phase difference of the stereo signals.

In this chapter, the problems related to downmix and stereo-to-mono compatibility is reviewed. A subband-domain downmixing scheme that addresses this problem is proposed as an optimization to the PS encoder. Following that, the implementation and evaluation result of the proposed scheme are presented.

4.1 Stereo to Mono Compatibility Problem

Signal coloration can occur during a stereo to mono downmixing due to the stereo signal components that are out of phase. An example is a stereo recording produced by spaced microphone technique [34] which is still in

widespread use nowadays. Using this technique, two microphones are placed with a large spacing of a few feet from each other. For any given point of sound origin between the two microphones, there is a path length difference from the sound source to each microphone. At some frequency, this difference will correspond to half an acoustic wavelength such that the signal of the two microphones will be fully out of phase. When downmixing is performed, phase cancellation occurs at various frequencies causing signal coloration.

Another example is a stereo audio that is created by downmixing from surround audio material using *Dolby Motion Picture Matrix* encoder [35]. The stereo downmix contains the surround signal part which has been phase-shifted to create a 180 degree phase difference between the signal component feeding the left and right stereo audio channels. Obviously when the stereo audio is further downmixed into monaural signal, that portion of surround signal is fully canceled and lost.

In Chapter 2 and 3, two downmixing schemes have been presented: time-domain averaging scheme and subband-domain averaging with stereo power equalization (subsequently referred to as '*subband-domain equalization*' scheme). Time-domain averaging scheme simply averages the time domain samples. Subband-domain equalization scheme averages the subband sample pairs and performs power equalization such that the power of the monaural signal is proportional to the total power of the original stereo signal.

Figure 4.1 shows the performance of the time-domain averaging and subband-domain equalization downmixing schemes, for stereo signals with inter-channel phase difference range of 0 to -2π and inter-channel level difference range of -70 to 70 dB. The test signals are narrowband 800 Hz stereo sinusoidal signals. The *normalized power* is calculated as the ratio of the monaural-downmixed signal power to the total original stereo signal power.

(a)

(b)

Figure 4.1: Normalized power of the monaural signal (with respect to the total power of the original 800 Hz sinusoidal stereo signal) for a range of IPD and IID. The monaural signals are generated with: (a) time-domain averaging, and (b) subband-domain equalization schemes.

It can be seen that time-domain averaging method is very prone to phase cancellation. Subband-domain equalization method gives substantial improvement, however the total power is not preserved for all cases due to strong phase cancellation when the signals are very much out-of-phase.

4.2 Proposed Optimized Downmixing Scheme

Downmixing is carried out on the pair of hybrid subband signals $l(k,n)$ and $r(k,n)$. The proposed downmixing scheme is highlighted by the shaded box in Figure 4.2. At each hybrid subband k , signal phase alignment is first carried out by performing phase shifting to the right hybrid subband samples according to

$$r'(k,n) = r(k,n) \exp(jIPD_k) \quad (4.1)$$

where $r'(k,n)$ is the phase-shifted right subband samples.

The amount of the phase shift IPD_k is equal to the phase difference between the left and right signal of the corresponding hybrid subband. This phase alignment minimizes phase cancellation when the signals are mixed subsequently. Note that as the subband samples are in complex representation, phase difference calculation as well as phase shifting can be performed efficiently.

After the phase alignment, the monaural subband samples are obtained by summing the stereo subband samples according to

$$m(k,n) = (l(k,n) + r'(k,n)) \cdot g(k,n) \quad (4.2)$$

where $g(k,n)$ is the signal dependent gain factor calculated as

$$g(k, n) = \sqrt{\frac{|l(k, n)|^2 + |r'(k, n)|^2}{0.5 \cdot |l(k, n) + r'(k, n)|^2}} \quad (4.3)$$

to perform power equalization such that the power of the monaural subband signal is half the total power of the original stereo subband signals.

Figure 4.2: Structure of parametric stereo encoder with the proposed downmixing scheme (shaded box).

4.3 Results and Discussion

Figure 4.3 shows the performance of the proposed downmixing scheme for the same stereo sinusoidal test signals as described in Section 4.1. The normalized power is calculated as previously defined in Section 4.1. It can be seen that the

proposed scheme eliminates the phase cancellation hence preserving overall power for all range of IID and IPD.

Figure 4.3: Normalized power of the monaural signal (with respect to the total power of the original 800 Hz sinusoidal stereo signal) for a range of IID and IPD. The monaural signal is generated with the proposed downmixing scheme.

In Section 4.3.1 and 4.3.2, the performance of the proposed downmixing scheme is compared to the subband-domain equalization scheme as implemented in the reference 3GPP eacPlus encoder. For this purpose, both downmixing schemes are implemented in the Matlab-based codec. The reason for using Matlab-based codec as opposed to C-based eacPlus is to eliminate the possible contribution of audio distortion by the SBR and AAC encoder. This way, the performance of the downmixing schemes can be justified independently i.e. the results can be interpreted as the contribution from the downmixing schemes in PS encoding.

4.3.1 Objective Audio Quality

Several types of stereo audio recordings as summarized in Table 4.1 are tested. Each test item is encoded by the PS encoder and downmixed using the proposed and subband-domain equalization schemes. The monaural audio obtained from both downmixing schemes are subsequently passed to the PS decoder. The objective quality of the decoded stereo audio with respect to the original stereo audio are then evaluated objectively using Basic version of *Perceptual Evaluation of Audio Quality* (PEAQ) method defined by *International Telecommunications Union* (ITU) [36].

PEAQ uses psychoacoustic principles to model human perception on the audio quality. It is a method commonly used by developers of audio coding schemes to objectively evaluate the performance of the algorithm. In the context of PS coding, this evaluation method has also been similarly used to evaluate the performance of PS encoding optimizations [37]. The output of PEAQ algorithm is *objective difference grade* (ODG) score which is a score ranging from 0.0 (transparent quality or imperceptible difference to the original audio) to -4.0 (highly perceptible and annoying distortion).

Table 4.1: Test stereo audio signals used to evaluate the performance of the proposed downmixing scheme.

Type of test audio	Test audio label
Speech in reverberant room	<i>spferev</i>
Solo instruments	<i>castanet</i> <i>glocken</i>
Commercial audio recording	<i>corrs</i> <i>funky</i> <i>rock</i>
Audio with complex spatial image	<i>canyon</i> <i>coleman</i> <i>applause</i>

The ODG scores of the decoded stereo audio from both downmixing schemes are plotted side by side in Figure 4.4. It can be observed that the objective quality of the stereo audio synthesized from the monaural downmix signal by both schemes are similar (showing a slight improvement for the proposed downmixing scheme), except for a particular test item labeled as *canyon*. For this test item a significant improvement in quality is obtained.

Figure 4.4: Objective audio quality test results of stereo audio decoded from monaural audio generated by subband-domain equalization and proposed downmixing schemes. Less negative ODG score implies better quality.

4.3.2 Power Preservation

Figure 4.5(a) shows the inter-channel phase difference histogram of 55 frames of audio from *canyon*, where each color spectrum bar represents the hybrid subband ($k = 0, 1, \dots, 70$). It can be seen that this audio segment contains considerable anti-phase signal components. Figure 4.6(a) shows the downmix

signal power ratio of the same audio segment for the subband-domain equalization downmixing scheme. The loss of the signal components due to phase cancellation is significant, resulting in heavy coloration of the signal. The loss of signal component in this case is also clearly identifiable when listening to the decoded audio clip and comparing it with the original audio clip. The proposed downmixing scheme significantly minimizes the loss as shown in Figure 4.6(b), hence the improvement in the audio quality. These audio clips can be downloaded from http://www.orbitfiles.com/download/thesis_ps.

As a comparison, Figure 4.5(b) shows the inter-channel phase difference histogram of 41 frames audio segment from *castanet*. It can be seen that the signal components are strongly in-phase. Hence phase cancellation is already minimal even with the subband-domain equalization downmixing scheme (Figure 4.7(a)). This explains the similar objective audio quality for this test item. The proposed downmixing scheme results in a slightly better signal preservation (Figure 4.7(b)). However, the important point here is that the scheme does not introduce degradation in audio quality for typical stereo audio with dominantly in-phase signal components.

Figure 4.5: Inter-channel phase difference histogram of audio segment from: (a) *canyon*, (b) *castanet*.

Figure 4.6: Downmixing power preservation for *canyon*: (a) monaural signal power ratio from subband-domain equalization scheme, (b) monaural signal power ratio from proposed downmixing scheme.

Figure 4.7: Downmixing power preservation for *castanet*: (a) monaural signal power ratio from subband-domain equalization scheme, (b) monaural signal power ratio from proposed downmixing scheme.

4.4 Downmixing Scheme Simplification

A simplification to the proposed downmixing scheme is possible, by taking advantage of the spatial parameter calculation. Instead of calculating the inter-channel phase difference at each subband k and then performing the phase shifting accordingly, the angle for phase shifting can be calculated only once at each stereo band. A phase shift of the same amount of angle is then performed for all of the hybrid subband samples within the stereo band group.

Extra computation can be kept low by taking intermediate values from ICC parameter calculation as performed following Equation (3.4) to calculate the stereo band phase shift angle. Informal listening test on the simplified phase alignment reveals that it does not significantly degrade the perceptual quality of the decoded stereo audio. It still provides signal component preservation for *canyon* test signal.

Chapter 5

Unified Transient Detector for Enhanced aacPlus Encoder

In SBR and AAC encoding, the time and frequency resolutions of the encoding are adaptive. For SBR, higher time resolution and lower frequency resolution are chosen for transient or percussive-like signal, and vice versa for stationary-like signal [13]. For this purpose, SBR encoder has a transient detection module that outputs transient information in the form of transient *flag* and *position*. Transient flag indicates whether the current frame contains transient. If transient is present, transient position indicates the position of the transient in terms of subband sample index. This information is fed into a *frame generator* module, where the time and frequency resolutions for the corresponding frame are determined [38].

For AAC, blocksize switching is performed when transient is detected in the frame [29][39]. By using shorter transform block for transient portion of the audio signal, *pre-echo* artifact [40] is minimized. Hence, AAC employs a transient detector which also outputs transient flag and position to regulate the block switching.

Looking at the structure of eacPlus encoder as shown in Figure 3.1, PS, SBR and AAC encoding are basically performed on the same signal. The differences are that PS works on the original stereo signal, SBR on the downmixed monaural subband-domain signal, and AAC on the band-limited

monaural time-domain signal. Hence, both transient detectors employed in each of the SBR and AAC encoding modules are basically also performing detection on the same signal.

In this chapter a unified, low-complexity transient detector which can replace both the SBR and AAC transient detectors is presented. The proposed transient detection is integrated into the PS spatial parameter extraction module. It is not a stand-alone transient detection module, i.e. the detection is performed as a side product of PS spatial parameter calculation. The so-called *unified transient detector* utilizes subband-based energy calculations performed during PS spatial parameter extraction to obtain the detection signal and subsequently determine the presence and location of the transient. Hence, it requires few additional computations. By performing the detection in PS encoder and subsequently passing the transient information to SBR and AAC encoder, the computational requirement of eacPlus encoding can be reduced.

5.1 Overview of Transient Detection in Music Signals

In the context of this thesis, transient can be defined as a short interval where the signal level (or power) jumps significantly with respect to the signal level of the previous neighboring samples or block of samples. As presented in details in [41], the steps involved can be summarized as follows. Figure 5.1(a) illustrates the ideal case of a single transient event which is characterized by its *onset* (time when the transient begins), *attack* (the sudden jump in energy), and *decay* (period when the energy level decays to steady state). A flowchart of the general transient detection scheme is shown in Figure 5.1(b).

Figure 5.1: (a) Ideal case of a time-domain transient signal and its common representation by onset, attack and decay of transient, (b) general transient detection scheme.

Pre-processing First an optional *pre-processing* of the signal is performed to enhance or attenuate certain features that are related to the transient. Such pre-processing can be performed in the time domain (eg. filtering) or frequency domain (eg. multiple subband decomposition). For example, as the sudden jump of energy level creates a discontinuity of the signal, the portion of the signal where the transient occurs can be enhanced by performing high-pass filtering as a pre-processing.

Reduction A *reduction* is performed by subsampling the pre-processed signal into *detection function* following certain criteria (eg. envelope of the signal). Based on the definition of transient which is related to a sudden jump of energy level in the signal, the algorithms generally use energy-based measure to produce the detection signal. For example, it can be done by grouping the samples into blocks and calculating the energy of the block.

The evolution of energy level from block to block is then observed as the detection function. In the case of a subband-domain detector, the time-domain signal is decomposed into frequency subbands and the detection function is obtained at each of the subband. The detection signal features from the different subbands are combined to characterize the transient.

Peak-picking Finally, *peak-picking* is performed on the detection signal to localize the transient event and extract the transient information. Common method is by thresholding, whereby the detection signal is compared to a fix or adaptive threshold. For example, transient can be characterized by a sudden jump of energy of the current block with respect to the previous blocks. When the jump of energy exceeds certain threshold, a transient is declared.

5.2 Enhanced aacPlus Transient Detection Schemes

In 3GPP eacPlus encoder, two transient detectors are employed: subband-domain detector for SBR encoder and time-domain detector for AAC encoder. The transient detection algorithms and complexity in terms of number of additions and multiplications are elaborated as follows.

SBR transient detector

The pseudo-code of the detection algorithm is shown in Pseudocode 5.1. The detection is performed on subband domain on each of the monaural QMF subbands. The algorithm detects sudden energy jump by comparing the energy gradient of neighboring energy blocks L and R against an adaptive threshold $t(i)$ which is calculated from the standard deviation $temp$ of the subband sample energies in the corresponding energy block and subband. The corresponding algorithm complexity is presented in Table 5.1.

Table 5.1: Complexity of the SBR transient detector.

Calculation	Multiplication	Addition
Mean energy	64	64×47
Standard deviation	$(64 \times 48) + 64$	$64 \times 2 \times 48$
Threshold calculation	64×3	64
Energy comparison	0	$64 \times 32 \times 6$
Thresholding	32	0
Total	3,424	21,504

```

for (i = 0; i < 64; i++)
    m =  $\frac{1}{48} \sum_{n=16}^{63} |X(i, n)|^2$ 
    temp =  $\sqrt{\frac{1}{47} \sum_{n=16}^{63} [m - |X(i, n)|^2]^2}$ 
    t(i) = MAX{128000, 0.66 · t(i) + 0.34 · temp}
for (n = 16; n < 48; n++)
    for (d = 1; d < 4; d++)
        L =  $\frac{1}{2} \left\{ \left| X(i, 2 \cdot \text{INT}\left\{\frac{n-d}{2}\right\}) \right|^2 + \left| X(i, 2 \cdot \text{INT}\left\{\frac{n-d}{2}\right\} + 1) \right|^2 \right\}$ 
        R =  $\frac{1}{2} \left\{ \left| X(i, 2 \cdot \text{INT}\left\{\frac{n+d}{2}\right\}) \right|^2 + \left| X(i, 2 \cdot \text{INT}\left\{\frac{n+d}{2}\right\} + 1) \right|^2 \right\}$ 
        if (R - L > t(i))
            a(n) = a(n) +  $\frac{R - L - t(i)}{t(i)}$ 

for (n = 8; n < 40; n++)
    if  $\left( a(n) < \frac{9}{10} a(n-1) \text{ AND } a(n-1) > 203.125 \right)$ 
        tranPos =  $\text{INT}\left(\frac{n-8}{2}\right)$ 
        tranFlag = 1
        break
    else
        tranPos = 0
        tranFlag = 0

```

Pseudocode 5.1: SBR transient detection in enhanced aacPlus encoder.**AAC transient detector**

Figure 5.2 shows the algorithm of the AAC transient detector. The detection is performed on a frame of 1024 monaural time-domain samples, which is

obtained by synthesis filtering of the band-limited (with half sampling rate) monaural subband signal from SBR encoding.

Figure 5.2: AAC transient detection in enhanced eacPlus encoder.

The algorithm first pre-processes the signal by performing a high-pass filtering whereby the filter transfer function is defined in z-domain as

$$H(z) = \frac{0.7548 \cdot (z - 1)}{z - 0.5095} \quad (5.1)$$

Reduction is performed by grouping the filtered signal $filtered(n)$ into eight energy blocks $en(block)$ according to

$$en(block) = \sum_{n=block \times 128}^{\{(block+1) \times 128\}-1} filtered^2(n) \quad , 0 \leq block < 8 \quad (5.2)$$

Finally, the energy is compared to the sliding average energy $avg_en(block)$ which is calculated recursively according to

$$avg_en(block) = \{\alpha \times en(block)\} + \{(1 - \alpha) \times avg_en(block - 1)\} \quad (5.3)$$

Whenever the energy of the current energy block exceeds the average energy by a certain fixed ratio *attack_ratio*, a transient is declared and the current block index *block* is defined as the transient position. The corresponding complexity of the algorithm is presented in Table 5.2.

Table 5.2: Complexity of the AAC transient detector.

Calculation	Multiplication	Addition
IIR Filtering	3×1024	2×1024
Energy Block Grouping	1024	8×127
Sliding Averaging	2×8	8
Energy comparison	8	
Total	4,120	3,072

5.3 Proposed Unified Transient Detector

Babu *et al.* [42] proposed a subband-domain transient detection where the detection is performed on a set of uniform frequency bands. Transient measures are estimated based on the signal energy at each band and the measures are combined to reach the transient decision. The scheme has been tested across wide range of audio signals and shown to perform robustly.

Parametric stereo encoder works on subband domain whereby the input signal is decomposed into many frequency subbands. As elaborated in Chapter 3, during the spatial parameter extraction the energy of the signals at each subband are calculated. Hence, it is possible to detect transient during PS parameter extraction with a low additional computations, because the energy value is readily available to be used for this purpose. The proposed unified transient detector is described as follows.

First the average monaural energy at stereo band b , $m_{transient}(b,n)$, is calculated by averaging the left and right stereo band energy. Note that the energy of the left and right stereo band have been pre-calculated during the spatial parameter extraction according to Equation (3.2) and (3.3). Hence the values can be directly used for the averaging. The current average power, $avg(b,n)$, is updated recursively according to

$$avg(b,n) = \alpha \times m_{transient}(b,n) + (1 - \alpha) \times avg(b,n-1) \quad (5.4)$$

where α is the recursive update coefficient. The detection signal at each band, $D(b,n)$, is then calculated as

$$D(b,n) = \frac{m_{transient}(b,n)}{avg(b,n)} \quad (5.5)$$

Whenever $D(b,n)$ exceeds a pre-defined threshold, a transient is declared for subband b . As a final decision, if transient is declared at more than a few stereo bands, the transient flag at the corresponding frame is set to 1 and the transient position is set as the current sample index n .

The transient information is subsequently passed to SBR and AAC encoder by translating the transient position index n into the corresponding SBR position index and AAC block index. The index translation between PS, SBR and AAC is performed according to the time index relationship between the modules as described in [38][39][43].

The recursive update coefficient α , the pre-defined threshold value, and the number of stereo bands to reach the final transient decision has been carefully tuned from experiments over a set of percussive signals. Generally transient are more clearly defined in the high frequency region, hence it is possible to use only a few highest stereo bands for the detection.

The complexity of the unified transient detector is presented in Table 5.3, where n_b refers to the number of stereo bands considered for the detection. The percentage of the computational requirement of the unified transient detector with respect to the total computations of both SBR and AAC transient detectors are shown in Figure 5.3, as a function of n_b . In the extreme case of 34- stereo band configuration where all 34 stereo bands are used for detection, it can be seen that the unified transient detector is still able to save up to 40% and 90% of the number of multiplications and additions, respectively.

Table 5.3: Complexity counts of the proposed unified transient detector, n_b is the number of stereo bands considered in the detection.

Operation	Multiplication	Addition
Energy averaging	$n_b \times 32$	$n_b \times 32$
Recursive power averaging	$n_b \times 32 \times 2$	$n_b \times 32$
Thresholding	$n_b \times 32$	0
<i>Total</i>	128 n_b	64 n_b

Figure 5.3: Percentage of the complexity of the unified transient detector with respect to the total complexity of the SBR and AAC transient detectors. The percentage is shown as a function of n_b .

5.4 Results and Discussion

The proposed unified transient detector has been implemented in enhanced aacPlus encoder as a part of PS encoder implementation. It replaces both original detectors in SBR and AAC.

5.4.1 Test Items and Detection Signal

The performance of the transient detector is evaluated using several percussive audio recordings as listed in Table 5.4. In this test, only three highest stereo bands are used for detection ($n_b = 3$). Each of the test items is encoded at a constant bitrate of 24 kbps with the reference and the modified encoders. Figure 5.4(a) shows a segment from *castanet* which contains a few transient events. The detection signal, $D(b,n)$ of the three highest stereo bands of the corresponding segment is shown in Figure 5.4(b).

Table 5.4: Test items used to evaluate the performance of the proposed unified transient detector.

Test item	Description
<i>castanet</i>	solo castanet
<i>glocken</i>	solo glocken
<i>triangle</i>	solo triangle
<i>drum</i>	solo tom tom drum
<i>castguitar</i>	mixture of castanet and guitar
<i>blue</i>	mixture of high-hat, bass guitar and melodic instrument

Figure 5.4: A transient audio segment from *castanet*: (a) time domain signal, (b) detection signal from three highest stereo bands for 20 bands PS configuration.

5.4.2 Transient Information Matching Accuracy

The resulting transient information from *castanet* as translated and passed to SBR is shown in Figure 5.5, and hereby it is compared to the transient information from the original SBR transient detector. Similarly, the transient information translated and passed to AAC is shown in Figure 5.6. In Figure 5.7, the percentage of transient information matching between the unified transient detector and the original SBR and AAC detectors for the test items listed in Table 5.4 are shown. A match is declared if:

- the transient flag is equal for both unified and original detectors, and
- as a reasonable tolerance, the transient position is within one index window from the original transient position.

With a computational reduction of more than 90% in this case ($n_b = 3$), the unified transient detector is able to achieve at least 80% matching detection to both original detectors.

5.4.3 Conformance to the 3GPP SBR Transient Detection Conformance Testing

3GPP defines conformance testing method for eacPlus encoder in [44]. For the SBR portion of the conformance, the implementation of transient detector is one of the conformance criteria. The test of the transient detector is performed by comparing the transient information of the encoder under test to the reference encoder provided by 3GPP. To pass the conformance, the *root mean square* (RMS) of the difference between the transient position vector of the encoder under test and the reference encoder shall not be greater than 0.2.

In this case, the comparison is performed between the transient information generated by the reference SBR transient detector to the one generated by the proposed unified transient detector. Two percussive test items are provided for the transient detector conformance: *hi-hat* (drum set) and *castagnettes* (latin percussion) signals. The proposed unified transient detection passed the SBR transient conformance for both test signals.

Figure 5.5: Transient information of *castanet* audio segment translated and passed from the unified transient detector ('o') to the SBR module, as compared to the original SBR transient detector output ('+').

Chapter 5. Unified Transient Detector for eacPlus Encoder

Figure 5.6: Transient information of *castanet* audio segment translated and passed from the unified transient detector ('o') to the AAC module, as compared to the original AAC transient detector output ('x').

Figure 5.7: Matching accuracy of the transient information detected and passed by the unified transient detector to the transient information from original SBR detector ('+') and AAC detector ('x').

5.5 Overall Assessment of the Implementation and Optimization

The implementation of PS encoder and the optimization in terms of downmixing and transient detection have been presented in the two previous chapters and this chapter. In this section, an overall assessment result of the implementation and optimization is presented. The assessment were carried out by performing complexity analysis and subjective listening test result that benchmarks the overall performance of the implementation with respect to the baseline 3GPP implementation.

5.5.1 Complexity Analysis

Hereby, the complexity performance of the optimized encoder is compared to the reference encoder implementation from 3GPP (*reference encoder*) [14]. To evaluate the complexity of the encoder implementation, the optimizations presented in the earlier chapters of this thesis are implemented into the fixed-point eacPlus encoder provided by 3GPP [45]. The complexity analysis tool has been integrated into the fixed-point code by 3GPP. The analysis tool outputs *weighted million operations per second* (wMOPS) as its complexity metric.

The calculation of the wMOPS can be summarized as follows. A set of fixed-point mathematical, logical, and data memory operators are defined for the fixed-point implementation. Each time a specific operator is called, the counter of the operator is incremented. The total number of times each operator was called during the encoding is then weighted according to its weightage. Finally to obtain the wMOPS figure, all weighted operations performed during the encoding are accumulated. The set of operators and their weightage can be found in the source file *count.h* and *basicop2.c* in the fixed-point source code [46].

Figure 5.8 presents the complexity profile of the reference and optimized eacPlus encoder in terms of the percentage of the total wMOPS. The modules that are of interest are the PS spatial parameter extraction and downmixing (*PS encoder*) and the original AAC and SBR transient detectors (*Transient Detect*). Analysis filtering refers to the hybrid analysis filtering module, while *SBR parameter calc.*, *Tonality Quotas* and *Synthesis downsample* are the modules related to the SBR encoder.

The optimized encoder includes phase parameter calculation, the proposed downmixing scheme and the unified transient detector. It can be seen that the optimizations cost 4% additional complexity to the PS encoder. However, now that the unified transient detector replaces both SBR and AAC transient detectors, 8% of transient detectors complexity is saved. In total, the optimized eacPlus encoder is able to save 4% of complexity with respect to the reference encoder.

5.5.2 Subjective Listening Test Result

A subjective listening test has been performed to benchmark the overall performance of the optimized encoder. The listening test was adapted from the *Comparison Category Rating* (CCR) approach [47] which compares and rates the two audio sequences on a scale of -3.0 to 3.0. The description of the rating scale and the corresponding audio quality is described in Figure 5.9(a).

For the test, the same test items as listed in Table 5.4 are used. Each test item is encoded by the modified and reference encoder at 24 kbps. Both bitstreams are then decoded using *FAAD2* decoder [31]. Subsequently, the decoded audio are labeled as either *A* or *B* in a random manner. Listeners were instructed to listen to the original excerpt, then version *A* and *B* of each test item. The listeners had to compare and rate the quality of *A* with respect to *B* (or vice versa) based on the rating scale.

Figure 5.8: Enhanced eacPlus encoder complexity distribution: (a) reference encoder, (b) optimized encoder.

Eleven listeners participated in the test, with the age ranging from 22 to 45 years old. The test was performed using *Sennheiser HD 590* high-definition hi-fi stereo headphone in an audio listening room. The listeners were allowed to listen to the audio excerpt as many times as they require.

The mean and 95% confidence interval rating from the listening test are shown in Figure 5.9(b). In the result presented, the decoded audio from the optimized and reference encoder are labeled as *A* and *B*, respectively. The result shows that, with less computational complexity from the optimizations, the optimized encoder is able to achieve similar or just a slightly lower decoded audio quality. For all of the test items, the 95% confidence interval range includes the score 0.0 which indicates similar quality. The audio files used in this listening test can be downloaded from http://www.orbitfiles.com/download/thesis_ps.

(a)

(b)

Figure 5.9: Subjective listening result to compare optimized and reference encoder: (a) description of the rating scale and the corresponding audio quality description, A and B refers to the decoded audio from the optimized and reference encoder respectively; (b) mean and 95% confidence interval ratings for the test items. All test items are not more than 10 seconds long.

Chapter 6

Objective Evaluation Method for Spatial Image Distortion in Stereo Audio Processing

Spatial image in stereo audio recordings determines how a listener perceives the direction of various audio sources in the stereo ensemble. During stereo audio playback, auditory spatial image is evoked. As a result, the listener is able to perceive several instruments as coming from various directions.

Processing performed on stereo audio might alter the spatial image of the processed audio, with respect to the original stereo audio. Example of such processing are PS coding presented in this thesis and Binaural Cue Coding (BCC) [20][21]. These schemes basically parameterize the spatial image of the stereo audio by a set of time- and frequency-varying spatial parameters. Distortion of the spatial image, whether perceivable or not, is bound to occur.

During the development of such audio processing algorithm, one might wish to evaluate how the algorithm affects the audio spatial image, and hence the perceived locations of the various instruments in the stereo ensemble. Of course it can always be assessed subjectively by performing listening test. However, aside from being tedious and time consuming, it is difficult to judge the spatial image distortion due to the complex mixture of instruments placed at various positions in stereo recording.

In this chapter, an objective method to evaluate the spatial image distortion is proposed. The proposed method is presented as a concept, and the evaluation of the concept is presented without practical implementation. The method evaluates the spatial image distortion of a processed stereo audio with respect to the original stereo audio. Two common spatial image distortion cases are analyzed: spatial image narrowing and instability. This chapter is organized as follows. First the method of spatial image distortion analysis is elaborated, followed by evaluation results and discussions. Finally, concluding remarks are presented.

6.1 Stereo Audio Model

Given directional sources S_i where i denotes the source index, and decorrelated ambience components A_L and A_R where L and R denotes the left and right audio channels, at time index n , a stereo audio can be represented as a mixture of the directional sources and ambience component by

$$L(n) = \sum_i g_{L,i} S_i(n) + A_L(n), \quad (6.1)$$

$$R(n) = \sum_i g_{R,i} S_i(n) + A_R(n), \quad (6.2)$$

where g is the panning index (weightage) of the corresponding audio source and channel. Here, it is assumed that the stereo audio recording consists solely of panned sources, which is the case for most studio recording and live recording with *coincident microphone* technique [34]. The term *spatial image* is used to refer to the position of the sources in the stereo audio ensemble.

Each audio source occupies certain region in the time-frequency plane and the value of its panning gain determines the position in the stereo ensemble. If the sources are not positioned at the same spatial location, each source can be uniquely identified by its panning gain. Therefore, by identifying the sources and extracting the panning gain, the spatial image of the stereo audio is obtained. To analyze the distortion of spatial image due to audio processing, the values of the panning gain between the original and processed audio may be compared and statistically analyzed.

6.2 Proposed Methodology

The objective spatial image distortion evaluation model diagram is shown in Figure 6.1, followed by detailed explanation of the method.

Figure 6.1: Model of the objective evaluation of stereo audio spatial image distortion.

6.2.1 Spatial Cue Calculation

Many spatial hearing models decompose the stereo signal into time-frequency components to approximate human auditory frequency decomposition

[48][49][50]. Subsequently, signal processing is performed to mimic the processing in the middle and inner ears. The spatial cues, *inter-aural level difference* (ILD) and *inter-aural time difference* (ITD) are subsequently computed, and these values reflect the location information of the audio sources.

As described earlier, each source can be identified by the value of the panning gain which is reflected in the value of the ILD as shown in Equation (6.4). In the proposed method, the analysis is performed directly on the stereo audio signal without considering the effect of middle and inner ears. Hence the value of *inter-channel level difference* (ICLD) is calculated instead of ILD. As a general scheme whereby the time domain stereo signals are decomposed into the left- and right-channel time – frequency components $L_{tf}(u,n)$ and $R_{tf}(u,n)$, ICLD is calculated as

$$ICLD(u,n) = 10 \log_{10} \frac{|L_{tf}(u,n)|^2}{|R_{tf}(u,n)|^2} \quad (6.3)$$

where u denotes the frequency channel index and n denotes the time index. At a time instance n_t where a particular stereo band u_t is occupied by a directional source i , assuming that the stereo audio is modeled as in Equation (6.1) and (6.2) and in the absence of ambience component, ICLD at that time – frequency point can be related to the panning gain of S_i as

$$ICLD(u_t, n_t) \approx 10 \log_{10} \frac{g_{L,i}^2}{g_{R,i}^2} \quad (6.4)$$

In PS coding, approximation to human auditory frequency decomposition is performed by the hybrid analysis filtering followed by stereo band grouping. The resulting stereo band signals $l(b,n)$ and $r(b,n)$ are used to calculate the PS spatial parameter, one of which is the IID. Hence the PS encoding method itself

can be adapted for the spatial cue calculation. Comparing Equation (3.5) and (6.3), and adapting b to refer to the frequency channel index, the spatial cue for the spatial distortion evaluation method is calculated following the PS spatial parameter extraction as

$$ICLD(b, n) = 10 \log_{10} \frac{e_l(b, n)}{e_r(b, n)} \quad (6.5)$$

where $e_l(b, n)$ and $e_r(b, n)$ are calculated according to Equation (3.2) and (3.3) with a slight modification. The pseudocode of the spatial cue calculation is presented in Pseudocode 6.1. Here, the spatial cue is calculated at 34 stereo bands for higher frequency resolution.

```

for (n = 0; n < 32; n++) {
  for (b = 0; b < 34; b++) {
    
$$e_l(b, n) = \sum_{k=k_l(b)}^{k_h(b)} |l(k, n)|^2$$


$$e_r(b, n) = \sum_{k=k_l(b)}^{k_h(b)} |r(k, n)|^2$$


$$e_R(b) = \sum_{k=k_l(b)}^{k_h(b)} l(k, n) r^*(k, n)$$

    }
    
$$ICLD(b, n) = 10 \log_{10} \frac{e_l(b, n)}{e_r(b, n)}$$


$$ICC(b, n) = \frac{|e_R(b, n)|}{\sqrt{e_l(b, n) e_r(b, n)}}$$

  }

```

Pseudocode 6.1: Spatial cue calculation for the spatial image distortion evaluation method.

Experiments in the field of psychoacoustic have shown that the auditory system has a limited temporal resolution in tracking the spatial localization cues. The corresponding time constants vary between 30 and 100 ms as summarized in [24]. Hence, to reduce the dimension of the data, especially for analyzing very long audio recording, ICLD can be calculated with a limited (but sufficient) time resolution by grouping every few stereo band samples. Note that in this model, middle and inner ear processing are not considered as the model is limited to objectively analyze and compare the spatial image of the stereo audio rather than the perception of the spatial image distortion itself.

Having calculated the spatial cue, there are a few things that can be noted. Firstly, not all of the points in the time-frequency plane are occupied by the sources. A lot of the points are occupied by noise or very weak signals that contribute to insignificant ICLD values. Secondly in real practice, the sources overlap in the time-frequency plane. The overlapping of the audio sources may introduce errors in the spatial parameter estimation [51]. Thirdly, the decorrelated ambience component due to natural or artificial reverberation might overlap with the directional sources causing similar parameter estimation errors. If the ICLD from these three cases are considered for the statistical analysis, these values might contribute to inaccuracy of the analysis result. Hence prior to the statistical analysis, spatial cue selection is performed to select the significant cue. The selection process is explained in the following section.

6.2.2 Spatial Cue Selection

Faller and Merimaa [50] proposed a source localization cues selection scheme based on *inter-aural coherence* (IC). These cues are only considered at time instances when only the direct energy of a single source is present. Such instances can be identified based on the IC values. Motivated by the results, the proposed spatial distortion evaluation method calculates ICC following the PS

spatial parameter extraction. The calculation is performed as shown in Pseudocode 6.1 in page 90.

The ICLD selection is subsequently performed by selecting only those time – frequency points where the ICC values are above a fix threshold $icc_threshold$. A time – frequency mask $m(b,n)$ is defined by looking at the ICC values of the original stereo audio $ICC_{original}(b,n)$ according to

$$m(b,n) = \begin{cases} 1 & , \text{ if } ICC_{original}(b,n) \geq icc_threshold \\ 0 & , \text{ otherwise} \end{cases} \quad (6.6)$$

Only the ICLD values of both the original and processed audio at the time – frequency points where $m(b,n) = 1$ are selected. The frequency and time index of the selected cues are denoted as b' and n' respectively. The cue selection process results in the selected original and processed spatial cues $ICLD_{original}(b',n')$ and $ICLD_{processed}(b',n')$.

The rationale of using the ICC of the original stereo audio to select the ICLD is that, it is desirable to obtain the valid spatial image from the original stereo audio and compare how much the same set of selected ICLD has been distorted in the processed stereo audio. The selected cues are subsequently used for the statistical analysis. In the following section, criteria to select $icc_threshold$ are presented as a recommendation. Following that, the statistical analysis method is presented in Section 6.2.4.

6.2.3 Fixing ICC Threshold for Spatial Cue Selection: A Recommendation

To illustrate the criteria for ICC threshold selection, as well as to evaluate the performance of the model, a semi-artificial reference stereo audio with known spatial property is created. The stereo audio is a mixture of two directional

sources: a female English speech S_1 panned to the left side location with ICLD of 12 dB and a solo castanet S_2 panned to the right side location with ICLD of -12 dB. The ICLD values of this reference stereo audio are referred to as *reference ICLD* subsequently.

A weakly correlated ambience component A_L and A_R taken from a binaural recording of a market street noise are added to the mixture. With this arrangement, the level of direct sources with respect to the ambience signal can be varied to give the desired *direct source to ambience ratio* (DSAR) which is defined as the ratio of the mean power of the direct sources ($\sigma_{S_1}^2$ and $\sigma_{S_2}^2$) to the ambience signal ($\sigma_{A_L}^2$ and $\sigma_{A_R}^2$) and formulated as [53]

$$DSAR = 10 \times \log_{10} \frac{\sigma_{S_1}^2 + \sigma_{S_2}^2}{\sigma_{A_L}^2 + \sigma_{A_R}^2} \quad (6.7)$$

Figure 6.2 shows the percentage of the selected directional source power as a function of ICC threshold at different DSAR values, where *Direct mixing* means that the directional sources are mixed without adding the ambience components ($DSAR \approx \infty$ dB). The *normalized selected power* (NSP) percentage is defined as

$$NSP = \frac{\sum_{b'} \sum_{n'} (L_{mixture}^2(b', n') + R_{mixture}^2(b', n'))}{\sum_b \sum_n (S_1^2(b, n) + S_2^2(b, n))} \times 100\% \quad (6.8)$$

where $L_{mixture}(b', n')$ and $R_{mixture}(b', n')$ are the left and right channel stereo band samples of the stereo audio mixture at the time - frequency points where they are selected by ICC thresholding, respectively. $S_1(b, n)$ and $S_2(b, n)$ are the complete time - frequency set of stereo band samples of S_1 and S_2 , respectively.

Figure 6.3 shows the percentage of the selected ICLD values that fall within ± 1 dB from actual S_1 and S_2 ICLD of 12 and -12 dB.

It can be observed from Figure 6.2 that the percentage of selected sources power increases as the ICC threshold decreases. However, from Figure 6.3 the percentage of the selected ICLD values that fall within the correct sources ICLD values decreases. The ICC threshold can be selected by compromising on how much direct sources power is selected and how accurate is the selected ICLD values. For the evaluation of the proposed method presented in the latter part of this chapter, ICC threshold value of 0.98 is selected that still gives at least 40% selected source power and 85% ICLD accuracy for a DSAR of 5 dB.

Figure 6.2: Percentage of selected directional source power as a function of ICC threshold at different DSAR.

Figure 6.3: Percentage of selected ICLD values that falls within ± 1 dB from the original source ICLDs.

6.2.4 Statistical Analysis Method

To obtain the output of the distortion analysis model, first the absolute ICLD error E_{ICLD} of the processed stereo audio is calculated according to

$$E_{ICLD}(b', n') = |ICLD_{processed}(b', n') - ICLD_{original}(b', n')| \quad (6.9)$$

Jiao *et al.* [52] proposed an objective measure of localization accuracy and stability of directional sources by the mean perceived direction and standard deviation of the localization error. Similar in this proposed method, the

objective measure of distortion is defined as distortion mean D_m and distortion standard deviation D_{std} of the selected ICLD error which is calculated as follows:

$$D_m = \frac{\sum_{b'} \sum_{n'} E_{ICLD}(b', n')}{M} \quad (6.10)$$

$$D_{std} = \sqrt{\frac{\sum_{b'} \sum_{n'} (E_{ICLD}(b', n') - D_m)^2}{M}} \quad (6.11)$$

where M is the total number of the selected ICLD points. The output values D_m and D_{std} represent the distortion distance and stability with respect to the original audio spatial image, as will be shown in the next section.

6.3 Evaluation of the Proposed Method

6.3.1 Spatial Image Narrowing

To evaluate the performance of the proposed model to detect spatial image narrowing, the semi-artificial original stereo audio as described in Section 6.2.3 with a DSAR value of 15 dB is used. The spatial image narrowing is simulated by creating a set of test stereo audio whereby the ICLD of S_1 and S_2 deviate by $ICLD_{dev} = 1$ to 10 dB in 1-dB step, with respect to their reference ICLDs. The original and test audio are subsequently input to the spatial distortion evaluation model created in Matlab.

Figure 6.4 shows the mean D_m and one standard deviation D_{std} range calculated from the spatial distortion model. The label '*simulated*' refers to the simulated spatial image distortion statistic. The evaluation method is able to identify the simulated spatial image deviation in the test audio. When all of the

calculated ICLDs are used in the statistical analysis, the output of the evaluation method does not relate well to the simulated distortion. It can be seen that by performing the spatial cue selection, the output can be significantly brought closer to the simulated distortion.

6.3.2 Spatial Image Instability

For a lot of audio recordings, the locations of the directional sources are stationary in the stereo ensemble. However, spatial processing might alter the stability of the source locations. Unstable spatial image is perceived when the directional source is moving while it is actually stationary in the reference stereo audio. Hereby, two types of instability are evaluated: unstable spatial image with varying *rate of instability* and varying *degree of instability*.

Figure 6.4: Spatial distortion evaluation result for spatial image narrowing.

For the first case, instability of the spatial image is simulated by modulating the panning gain of S_1 and S_2 such that their ICLD varies sinusoidally from the original value (± 12 dB) to 0 dB. The modulation rate varies from 1 to 10 Hz in 1-Hz step. Figure 6.5 shows the output for unstable spatial image with varying instability rate. It can be seen that the evaluation method identifies the instability by outputting a large standard deviation range which approximates the reference distortion.

For the second case, the test audio are created by modulation of its panning gain such that the ICLD of S_1 and S_2 vary sinusoidally from 12 dB to $(12 - \textit{intensity})$ dB and -12 dB to $(-12 + \textit{intensity})$ dB respectively, where *intensity* varies in 2-dB step with a fix modulation rate of 2 Hz. Figure 6.6 shows the output for unstable spatial image with varying degree of instability at fixed gain modulation rate. As observed, the increase in the degree of instability is identified by the evaluation method by the increase in the standard deviation accordingly. The mean and standard deviation are also close to the reference distortion.

Figure 6.5: Spatial distortion analysis output for unstable spatial image with varying instability rate.

Figure 6.6: Spatial distortion analysis output for unstable spatial image with varying degree of instability.

Chapter 7

Summary, Conclusions and Recommendations

7.1 Summary and Conclusions

The implementation of MPEG-4 Parametric Stereo encoder has been presented. The implementation takes the non-compliant baseline PS encoder by 3GPP as the reference point. This reference encoder only implements the baseline version of the MPEG-4 encoder, whereby it only supports limited PS configuration and does not support the phase parameter calculation. It has been shown how to directly quantize the phase parameter using simple arithmetic and efficiently calculate the OPD parameter.

The problem of signal cancellation during stereo-to-mono downmixing was identified. A downmixing scheme that minimizes this problem has been proposed. The phase alignment performed prior to the downmixing is able to minimize the phase cancellation when the signals are subsequently mixed, hence improving the decoded audio quality. This scheme is especially advantageous for stereo audio that contain significant amount of out-of-phase signal components.

As the SBR and AAC transient detectors in the eaacPlus encoder basically perform detection on the same signal, it is identified that both detectors may be replaced by just one transient detector to reduce the complexity of the encoder.

A low-complexity transient detection algorithm that is integrated into the PS spatial parameter extraction has been presented. The detector detects the jump of energy level at each stereo band and combines the detection information across frequency to come out with the transient decision. It maintains its low complexity by utilizing of the energy values that are already calculated during the spatial parameter extraction. The unified transient detector is able to provide a good match to the transient information generated by both original SBR and AAC transient detectors.

During the implementation and optimizations work of the PS encoder, the author faced some difficulty in assessing the effect of the different PS configurations and optimizations to the spatial image of the decoded audio. Similarly, many audio processing may result in stereo spatial image distortion that is difficult and tedious to be assessed subjectively. Motivated by this problem, an objective evaluation concept for spatial image distortion has been presented. The method uses similar frequency decomposition utilized in PS encoder. Here, the ICLD deviation is used as the distortion indicator, assuming that the stereo audio recordings are created by panning of the various directional sources into the stereo ensemble.

Spatial cue selection is performed to select the significant ICLD values to be considered in the analysis. The selection is based on the ICC values and an example of the selection process has been presented. Statistical analysis is performed by first calculating the error between the selected ICLD of the original and processed stereo audio. Subsequently the mean and standard deviation of the error is calculated as the output. The output metrics relates to spatial distortion cases that were tested: the mean corresponds to spatial image *narrowing* and the standard deviation corresponds to spatial image *instability*. Furthermore, the spatial cue selection has been shown to be an important step for the analysis, as it is able to bring the output metrics closer to the simulated distortion.

The complexity analysis of the implementation and optimizations revealed that the optimizations indeed results in reduction of encoding complexity with respect to the reference encoder. In particular, the unified transient detector significantly contributes to the reduction. Finally, a subjective listening test shows that the optimized encoder results in audio quality that is similar to the reference encoder.

7.2 Recommendation for Future Research

The author feels that the objective evaluation of spatial image distortion concept will find a useful application in the field of audio coding. With the recent development in spatial audio coding such as *MPEG Surround* [54] and three-dimensional audio (3D audio), the method will provide a quick and efficient way to assess the effect of the algorithm to the spatial image quality.

However, the method and results presented in this thesis is still in a conceptual stage. For example the statistical analysis, though able to identify the spatial image distortion, performs a general statistical calculation without considering the contribution from each directional source separately. More research effort may be put into finding the methods to analyze the distortion of each audio source separately, and then weighting and combining these contributions into a final spatial image distortion output metrics. A few methods may be considered to cluster each individual source in the time - frequency plane. Such method could be *K-means clustering* [55] of the ICLD values or even performing audio source separations [56].

To analyze the significance of the contribution from various sources to the evaluation output metrics, the author recommends performing psychoacoustic experiments that is designed to reveal the features of stereo audio signal that significantly contributes to the perception of the spatial image distortion. Also

Chapter 7. Summary, Conclusion and Recommendations

for a more complete evaluation model, the value of *inter-channel time difference* (ICTD) may be considered in the evaluation method.

List of Publications

1. Samsudin, E. Kurniawati, B. P. Ng, F. Sattar and S. George, "A Stereo to Mono Downmixing Scheme for MPEG-4 Parametric Stereo Encoder," presented at the 31st IEEE International Conference on Acoustic, Speech and Signal Processing, Toulouse, France, 2006.
2. Samsudin, E. Kurniawati, F. Sattar, B. P. Ng, and S. George, "A Subband Domain Downmixing Scheme for Parametric Stereo Encoder," presented at the 120th Audio Engineering Society Convention, Paris, France, 2006.
3. Samsudin, B. P. Ng, E. Kurniawati, F. Sattar and S. George, "A Unified Transient Detector for Enhanced aacPlus Encoder," presented at the 120th Audio Engineering Society Convention, Paris, France, 2006.
4. Samsudin, B. P. Ng and F. Sattar, "An Objective Evaluation Method for Spatial Image Distortion in Stereo Audio Processing," submitted to IEEE Signal Processing Letters.
5. Samsudin, B. P. Ng and F. Sattar, "Optimizations of Parametric Stereo Encoding in Enhanced aacPlus Encoder," in preparation to be submitted to a journal.

References

- [1] H. Autti and J. Bistrom, "Mobile Audio – from MP3 to AAC and further," Helsinki University of Technology, T-111.550 Multimedia Seminar: Mobile Multimedia Application Platforms, Fall 2004.
- [2] R. Koenen, "Overview of the MPEG-4 Standard," ISO/IEC JTC1/SC29/WG11 N4668, March 2002.
<http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm#3.4>
- [3] O. Kunz and K. Brandenburg, "An Overview of MPEG-4 Audio," in *Proceedings of Audio Engineering Society UK Conference 'Audio: The Second Century'*, 1999, pp. 137-145.
- [4] Y. Wang, "Selected Advances in Audio Compression and Compressed Domain Processing," PhD thesis, Tampere University of Technology, 2002.
- [5] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio" in *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451-513, 2000.
- [6] B. Edler and H. Purnhagen, "Parametric Audio Coding," in *Proceedings of the Fifth International Conference on Signal Processing*, vol. 1, 2000, pp. 21 – 24.
- [7] H. Purnhagen and N. Meine, "HILN--The MPEG-4 Parametric Audio Coding Tools," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, 2000, pp. III-201 – III-204.
- [8] E. Schuijers, W. Oomen, B. den Brinker and J. Breebaart, "Advances in Parametric Coding for High-Quality Audio," in *114th Audio Engineering Society Convention*, 2003, preprint no. 5852.

- [9] S. Bruhn, "Bridging the gap between speech and audio coding - AMR-WB+ - The codec for mobile audio," 2E1380 S3 Seminar Series, Royal Institute of Technology, Stockholm, Sweden, May 2004.
[http://www.s3.kth.se/radio/COURSES/S3_SEMINAR_2E1380_2004/Se
m040506.shtml](http://www.s3.kth.se/radio/COURSES/S3_SEMINAR_2E1380_2004/Se
m040506.shtml)
- [10] Dolby Laboratories, Apple Computer, AT&T Wireless Services and RealNetworks, "Work Item Description for Higher Bitrate Audio Codec," 3GPP TSGS#19(03)0164, 2003.
- [11] International Standard ISO/IEC 14496-3:2001/Amd 1:2003, "Information technology - Coding of audio-visual objects - Part 3: Audio - Amendment 1: Bandwidth extension."
- [12] International Standard ISO/IEC 14496-3:2001/Amd 1:2004, "Information technology - Coding of audio-visual objects - Part 3: Audio - Amendment 2: Parametric coding for high-quality audio."
- [13] A. Ehret, M. Dietz and K. Kjorling, "State-of-the-Art Audio Coding for Broadcasting and Mobile Applications," in *114th Audio Engineering Society Convention*, 2003, preprint no. 5834.
- [14] 3GPP TS 26.401, "Enhanced aacPlus General Audio Codec - General Description (Release 6)," 3GPP Technical Specification Group Services and System Aspects, 2004.
- [15] <http://www.codingtechnologies.com/products/aacPlus.htm>
- [16] E. Schuijers, J. Breebaart, H. Purnhagen and J. Engdegard, "Low complexity parametric stereo coding," in *116th Audio Engineering Society Convention*, 2004, preprint no. 6073.
- [17] J.D. Johnston, and A.J. Ferreira, "Sum-difference stereo transform coding." in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, 1992, pp. II-569-572.
- [18] J. Herre, K. Brandenburg and D. Lederer, "Intensity Stereo Coding," in *96th Audio Engineering Society Convention*, 1994, preprint no. 3799.

- [19] H. Fuchs, "Improving Joint Stereo Audio Coding by Adaptive Inter-Channel Prediction," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustic*, 1993, pp. 39-42.
- [20] F. Baumgarte and C. Faller, "Binaural cue coding – Part I: Psychoacoustic fundamentals and design principles," *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 6, pp. 509 – 519, November 2003.
- [21] C. Faller and F. Baumgarte, "Binaural cue coding – Part II: Schemes and Applications," *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 6, pp. 520 – 531, November 2003.
- [22] J. Breebaart, S. van de Par, A. Kohlrausch and E. Schuijers, "High-quality parametric spatial audio coding at low bitrates," in *116th Audio Engineering Society Convention*, 2004, preprint no. 6072.
- [23] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997.
- [24] J. Breebaart, S. van de Par, A. Kohlrausch and E. Schuijers, "Parametric Coding of Stereo Audio," *EURASIP Journal on Applied Signal Processing*, Issue 9, pp. 1305 – 1322, 2005.
- [25] P. Ekstrand, "Bandwidth Extension of audio signals by spectral band replication," in *Proceedings of 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio*, 2002, pp. 73 – 79.
- [26] F. Mintzer, "On half-band, third-band and Nth band FIR filters and their design," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. ASSP-30, pp. 734 – 738, October 1982.
- [27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research* vol. 47, pp. 103 – 138, 1990.
- [28] J. Engdegard, H. Purnhagen, J. Roden and L. Liljeryd, "Synthetic ambience in parametric stereo coding," in *116th Audio Engineering Society Convention*, 2004, preprint no. 6074.

- [29] International Standard ISO/IEC 14496-3: 2001. *Information technology – Generic coding of moving pictures and associated audio information: - Part 3: Advanced Audio Coding (AAC)*.
- [30] L. G. Chen, S. W. Huang and T. H. Tsai, “Fast Filterbanks for the Low Power MPEG High Efficiency Advanced Audio Coding Decoder,” in *118th Audio Engineering Society Convention*, 2005, preprint no. 6336.
- [31] www.audiocoding.com
- [32] C. Faller, “Parametric Coding of Spatial Audio,” Ph.D. thesis, École Polytechnique Fédérale de Lausanne, France, 2004.
- [33] Samsudin, *Matlab-based Parametric Stereo Model: Source Code Listing*.
- [34] J. Wuttke, “Surround Recording of Music: Problems and Solutions,” in *119th Audio Engineering Society Convention*, 2005, preprint no. 6556.
- [35] R. Dressler, *Dolby Surround Pro Logic Decoder Principles of Operation*.
www.dolby.com/assets/pdf/tech_library/208_Dolby_Surround_Pro_Logic_Decoder.pdf
- [36] ITU-R Recommendation BS.1387, *Method for Objective Measurements of Perceived Audio Quality*.
- [37] T. W. Chang, H. W. Hsu, K. C. Lee, W. C. Lee, C. M. Liu, and C. H. Yang, “Efficient Design of Time-Frequency Stereo Parameter Sets for Parametric HE-AAC,” in *119th Audio Engineering Society Convention*, 2005, preprint no. 6600.
- [38] 3GPP TS 26.404, “Enhanced aacPlus general audio codec; Encoder Specification SBR part,” 3GPP Technical Specification Group Services and System Aspects, 2004
- [39] 3GPP TS 26.403, “Enhanced aacPlus general audio codec; Encoder Specification AAC part,” 3GPP Technical Specification Group Services and System Aspects, 2004.
- [40] M. Erne, “Perceptual Audio Coders “What to Listen for,” in *111st Audio Engineering Society Convention*, 2001, preprint no. 5489.

- [41] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and Sandler, M.B. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035 – 1047, September 2005.
- [42] V. S. Babu, A. K. Malot, V. M. Vijayachandran and M. K. Vinay, "Transient Detection for Transform Domain Coders," in *116th Audio Engineering Society Convention*, 2004, preprint no. 6175.
- [43] 3GPP TS 26.405, "Enhanced aacPlus general audio codec; Encoder Specification Parametric Stereo part," 3GPP Technical Specification Group Services and System Aspects, 2004.
- [44] 3GPP TS 26.406, "Enhanced aacPlus general audio codec; Conformance testing," 3GPP Technical Specification Group Services and System Aspects, 2004.
- [45] 3GPP TS 26.411, "Enhanced aacPlus general audio codec; Fixed-point ANSI-C code," 3GPP Technical Specification Group Services and System Aspects, 2004.
- [46] www.3gpp.org/ftp/Specs/html-info/26411.htm
- [47] ITU-R Recommendation P.830, "Subjective performance assessment of telephone-band and wide-bandwidth digital codecs."
- [48] E. A. Macpherson, "A Computer Model of Binaural Localization for Stereo Imaging Measurement," *Journal of Audio Engineering Society*, vol. 39, no. 9, pp. 604–622, September 1991.
- [49] J. Breebaart, S. van de Par and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. Model structure," *Journal of the Acoustical Society of America*, vol. 110, pp. 1074–1088, August 2001.
- [50] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *Journal of the Acoustical Society of America*, vol. 116, pp. 3075–3089, November 2004.

- [51] C. Avendano, "Frequency-domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression and Re-panning Applications," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 55–58.
- [52] Y. Jiao, Z. Zhang, T. Qu and X. Wu, "A Novel Objective Localization Quality Evaluation Method for Reproduced Sound," in *118th Audio Engineering Society Convention*, 2005, preprint no. 6440.
- [53] M. Briand, D. Virette and N. Martin, "Parametric Representation of Multichannel Audio Based on Principal Component Analysis," in *120th Audio Engineering Society Convention*, 2006, preprint no. 6813.
- [54] J. Breebaart, S. Disch, C. Faller, *et al.*, "MPEG Spatial Audio Coding / MPEG Surround: Overview and Current Status," in *119th Audio Engineering Society Convention*, 2005, preprint no. 6599.
- [55] K. Teknomo, *K-Means Clustering Tutorials*.
<http://people.revoledu.com/kardi/tutorial/kMean/>
- [56] N. Mitianoudis and M. E. Davies, "Audio source separation: solutions and problems," *International Journal of Adaptive Control and Signal Processing*, vol. 18, pp 299 – 314, 2004.