2011

Spatial squeezing techniques for low bit-rate multichannel audio coding

Bin Cheng

University of Wollongong
NOTE

This online version of the thesis may have different page formatting and pagination from the paper copy held in the University of Wollongong Library.

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.
Spatial Squeezing Techniques for Low Bit-Rate Multichannel Audio Coding

A thesis submitted in fulfilment of the requirements for the award of the degree

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

by

Bin Cheng
Bachelor of Engineering
Beijing University of Aeronautics and Astronautics, 2003
Master of Engineering Studies
University of Wollongong, 2004

SCHOOL OF ELECTRICAL, COMPUTER AND TELECOMMUNICATIONS ENGINEERING
2010
Abstract

In recent years, significant research has been focused on efficient compression and representation of multichannel spatial audio signals. Recent developments in this area exploit spatial audio cues representing inter-channel mathematical relationships. The original multichannel signal is downmixed to a backward compatible mono/stereo signal, while the spatial cues are utilized for recovering the surround sound. It is shown that, these approaches provide efficient coding of multichannel spatial audio signals, in terms of both bit-rate reduction and perceptual quality. However, drawbacks can be found in these approaches as the spatial cues do not represent perceptually relevant information, which can result in inefficient quantisation, as well as perceptual distortion of the localisation characteristics of the soundfield. Furthermore, in these approaches, as the downmixing and spatial cue derivation algorithm is specifically designed to suit a certain multichannel audio format, the flexibility and extensibility for coding future multichannel audio formats is limited.

The Spatially Squeezing Surround Audio Coding (S\(^3\)AC) is presented in this thesis as an alternative efficient solution for the representation of spatial audio signals. Based on estimating sound source and localisation information in the spatial soundfield, the fundamental idea in S\(^3\)AC is to represent a surround soundfield with a ‘squeezed’ soundfield by exploiting perceptual localisation irrelevancy. In particular, it is shown that, while limited perceptual precision is required for representing localisation information of a surround soundfield without perceptual distortion, the localisation precision computationally derivable for a small soundfield is adequate
to save the perceptual localisation information of a surround soundfield. Thus, a multichannel spatial audio signal rendering a surround soundfield can be represented by a small soundfield rendered with less channels, while additional spatial cues are not required. A typical $S^3AC$ application is then introduced, where a 5.1-channel surround audio signal is efficiently represented by a stereo downmix signal, which renders a ‘squeezed’ version of the original surround soundfield. This stereo signal is backward compatible to a conventional audio system, but can also be exploited to recover the original surround soundfield.

The proposed $S^3AC$ approach is then further analyzed. The localisation resolution in the $S^3AC$ squeezed soundfield is analyzed and is shown to be frequency and sound source dependent. The limitation of the squeezing process is then derived and evaluated. To further reduce the required bandwidth, a mono downmixing is introduced for $S^3AC$, with the source localisation information represented by $S^3AC$ cues. Compared with cues in other spatial audio coding approaches, the $S^3AC$ cues benefit from its feature of representing direct localisation information. Thus, an efficient $S^3AC$ cue quantisation solution based on psychoacoustical localisation principle is presented. In addition, a sound source localisation estimation algorithm is introduced, which can be used for any arbitrary multichannel audio format for extended flexibility.

Several additional $S^3AC$ applications are introduced. An efficient compression solution for Ambisonics B-format surround soundfield recording is presented based on $S^3AC$, which also extends the backward compatibility of Ambisonics signals. A binaural reproduction technique is also described for any $S^3AC$ encoded signal, for providing virtual surround sound experience over headphones. The $S^3AC$ soundfield squeezing idea is then exploited for multi-party teleconferencing scenario, where soundfields from different remote sites are perceptually discriminated when played back at the local site. The $S^3AC$ squeezing limitation derived earlier is then further exploited for representing multiple surround soundfields with one $S^3AC$ downmix.
Finally, the S³AC approach is extended for compressing multichannel three-dimensional audio signals. A source localisation estimation algorithm for any arbitrary 3D audio format is developed. The resulting source localisation is quantised based on a 3D source localisation quantisation approach, which exploits psychoacoustical principles for minimum localisation distortion. An extended S³AC spatial squeezing algorithm for is introduced for efficient and backward compatible representation of a 3D soundfield with a stereo downmix, while the 3D soundfield can also be represented by a mono downmix with accompanying S³AC cues that directly represent source localisation information.
Statement of Originality

This is to certify that the work described in this thesis is entirely my own, except where due reference is made in the text.

No work in this thesis has been submitted for a degree to any other university or institution.

Signed

Bin Cheng
Date Month Year
Acknowledgments

I would like to sincerely thank my supervisors Dr. Christian Ritz and Professor Ian Burnett for their guidance, encouragement, support and inspiring suggestions.

感谢爸爸妈妈，感谢你们抚养我长大成人，身体力行地教我做人做事的道理，并且这么多年来一直不求回报地支持我的学业。没有你们的支持和鼓励，以及从你们身上学习到的精神，我无法完成这个博士。寒窗苦读二十多年，现在终于算是告一个段落，一直在努力能让你们为我感到骄傲，并且会努力下去。

A big thank you all to my Whisper Labs colleagues, Eva, Kevin, Xiguang, Stephen and Shujau, for creating an active, happy and inspiring atmosphere while we are always being moved around, and of course, those great coffee times.

Last but not least, many thanks to my friends in Australia, for all those weekend dinners and get-togethers full of joy and happiness.
To Billie
# Contents

1 Introduction

1.1 Background ................................................................. 1
1.2 Spatially Squeezed Surround Audio Coding ............................ 4
1.3 Thesis Outline .............................................................. 6
1.4 Contributions ............................................................... 7
1.5 Publications ................................................................. 9
   1.5.1 Journal Publications ................................................ 9
   1.5.2 Book Chapters ........................................................ 9
   1.5.3 Conference Publications ............................................. 10

2 Literature Review

2.1 Introduction ................................................................. 11
2.2 Digital Audio Processing Fundamentals ............................... 12
   2.2.1 Digital Representation of Audio Signals ....................... 12
   2.2.2 Mono, Stereo and 2D Multichannel Audio Formats .......... 13
   2.2.3 Time-Frequency Representation of Audio Signals .......... 15
2.3 Fundamental Psychoacoustics .......................................... 19
   2.3.1 Lossless and Lossy Audio Coding ............................... 19
   2.3.2 Hearing Threshold .................................................. 20
2.3.3 Perceptual Filterbanks ........................................ 21
2.3.4 The Masking Phenomenon ................................... 23
2.3.5 Binaural Listening and HRTF ................................. 24
2.3.6 Localisation Blur .............................................. 27
2.3.7 Spatial Unmasking ............................................. 28
2.4 Perceptual Coding of Mono/Stereo Audio ....................... 29
  2.4.1 The MPEG-1 Layer III ....................................... 29
  2.4.2 The MPEG-2 AAC ........................................... 31
  2.4.3 High-Efficiency AAC ....................................... 33
2.5 Parametric Coding of 2D Spatial Audio ......................... 34
  2.5.1 Binaural Cue Coding ....................................... 35
  2.5.2 Parametric Stereo ........................................... 39
  2.5.3 Directional Audio Coding and Vector Based Amplitude Panning ................................. 41
  2.5.4 What about the LFE? ....................................... 45
  2.5.5 MPEG Surround ............................................. 45
  2.5.6 Binaural Reproduction of Spatial Audio ..................... 52
  2.5.7 The ‘Downmix + Cues’ Framework ......................... 53
2.6 Ambisonics ....................................................... 55
  2.6.1 First Order Ambisonics and B-Format ....................... 55
  2.6.2 UHJ .......................................................... 57
  2.6.3 Higher Order Ambisonics .................................. 58
  2.6.4 Ambisonics Reproduction over Loudspeakers ................. 59
2.7 3D Spatial Audio .................................................. 61
2.8 Evaluation of Audio Quality ..................................... 63
  2.8.1 Mean Opinion Score ....................................... 63
2.8.2 Perceptual Evaluation of Audio Quality .......................... 63
2.8.3 MUSHRA .................................................... 64
2.9 Summary ......................................................... 65

3 Spatially Squeezed Surround Audio Coding: Fundamental Principles and Applications 67
3.1 Introduction .................................................... 67
3.2 Fundamental Principles of $S^3$AC ................................ 68
  3.2.1 Squeezing the Auditory Space ................................. 68
  3.2.2 Sound Source Localisation Estimation ......................... 70
3.3 $S^3$AC Applied to ITU 5.1-Channel Surround Audio ............ 72
  3.3.1 Time-Frequency Transform ................................... 72
  3.3.2 $S^3$AC Encoding ............................................ 76
  3.3.3 $S^3$AC Decoding ............................................ 80
3.4 Evaluation ....................................................... 82
  3.4.1 Objective Evaluation ......................................... 82
  3.4.2 Subjective Evaluation ........................................ 84
3.5 Summary ......................................................... 89

4 Further Analysis and Improvements .................................. 92
4.1 Window Aliasing in $S^3$AC ...................................... 93
4.2 Localisation Resolution in $S^3$AC ................................ 95
  4.2.1 Frequency Dependent Localisation Resolution ............... 95
  4.2.2 Localisation Resolution of the Low Amplitude Components and Perceptual Relevancy .......................... 97
  4.2.3 $S^3$AC Squeezing Limitation ................................ 101
4.3 $S^3$AC Mono Downmixing and Localisation Side Information .... 103
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Using Mono Downmix and Cues in S³AC</td>
<td>103</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Location Dependent Perceptual Source Localisation Precision</td>
<td>105</td>
</tr>
<tr>
<td>4.3.3</td>
<td>S³AC Psychoacoustics Based Cue Quantisation and Coding</td>
<td>108</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Using S³AC Side Information for Coding of Sources Overlapped in Time-Frequency</td>
<td>112</td>
</tr>
<tr>
<td>4.4</td>
<td>S³AC Source Localisation Estimation based on Orthogonal Analysis</td>
<td>118</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Tangent Panning and Vector Panning</td>
<td>121</td>
</tr>
<tr>
<td>4.4.2</td>
<td>2D orthogonal analysis for virtual sound source localisation</td>
<td>125</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>S³AC Extended Applications</td>
<td>129</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>129</td>
</tr>
<tr>
<td>5.2</td>
<td>S³AC Compression of Ambisonics Signals</td>
<td>129</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Ambisonics and Amplitude Panning</td>
<td>129</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Conventional UHJ compression of Ambisonics</td>
<td>132</td>
</tr>
<tr>
<td>5.2.3</td>
<td>The S³AC approach to compressing Ambisonics</td>
<td>132</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Evaluation</td>
<td>137</td>
</tr>
<tr>
<td>5.3</td>
<td>S³AC Binaural Reproduction</td>
<td>140</td>
</tr>
<tr>
<td>5.3.1</td>
<td>System</td>
<td>140</td>
</tr>
<tr>
<td>5.3.2</td>
<td>HRTF Interpolation</td>
<td>143</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Evaluation</td>
<td>143</td>
</tr>
<tr>
<td>5.4</td>
<td>S³AC in Multi-Party Teleconferencing</td>
<td>147</td>
</tr>
<tr>
<td>5.4.1</td>
<td>System</td>
<td>147</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Recording and Representation of Localised Speech Signal</td>
<td>149</td>
</tr>
<tr>
<td>5.4.3</td>
<td>S³AC Squeezed Soundfield Reproduction</td>
<td>151</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Evaluation</td>
<td>154</td>
</tr>
<tr>
<td>CONTENTS</td>
<td>xi</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>----</td>
<td></td>
</tr>
<tr>
<td>5.5 S³AC Representation of Multiple Soundfields</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>5.5.1 System</td>
<td>157</td>
<td></td>
</tr>
<tr>
<td>5.5.2 Evaluation</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>5.6 Summary</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>6 Compressing the Multichannel Three Dimensional Audio</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>6.2 3D orthogonal analysis for virtual sound source localisation</td>
<td>169</td>
<td></td>
</tr>
<tr>
<td>6.3 S³AC Compression of 3D Multi-Channel Spatial Audio</td>
<td>171</td>
<td></td>
</tr>
<tr>
<td>6.3.1 System Overview</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>6.3.2 Time-Frequency Orthogonal Analysis for 3D Source Azimuth-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elevation Information</td>
<td>175</td>
<td></td>
</tr>
<tr>
<td>6.3.3 Spatial Localisation Quantisation Points</td>
<td>176</td>
<td></td>
</tr>
<tr>
<td>6.3.4 Stereo downmixing and S³AC spatial squeezing for 3D sound-field</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>6.3.5 Mono downmixing and further processing on SLQP</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>6.3.6 Decoding and reproduction</td>
<td>184</td>
<td></td>
</tr>
<tr>
<td>6.4 Evaluations</td>
<td>186</td>
<td></td>
</tr>
<tr>
<td>6.4.1 16-Channel 3D audio files for evaluation</td>
<td>187</td>
<td></td>
</tr>
<tr>
<td>6.4.2 Objective evaluation</td>
<td>189</td>
<td></td>
</tr>
<tr>
<td>6.4.3 Subjective evaluation</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>6.5 Summary</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>7 Conclusions and Further Work</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>7.1 Conclusions</td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>7.2 A ‘Sound Source + Localisation’ Framework</td>
<td>203</td>
<td></td>
</tr>
<tr>
<td>7.3 Further Work</td>
<td>204</td>
<td></td>
</tr>
</tbody>
</table>
Bibliography

A Higher Order Ambisonics and the Related Spherical Harmonics Functions
# List of Figures

2.1 Pulse Code Modulation .................................................. 13
2.2 ITU 5.1 multichannel loudspeaker layout .......................... 14
2.3 50% overlapping and windowing in STFT .......................... 16
2.4 The absolute hearing threshold ....................................... 22
2.5 Critical Bandwidth and ERB as functions of center frequency (between 10Hz and 20kHz) .................................................. 24
2.6 Free-field single sound source propagation on a horizontal surface and the left/right ear entrance signals .......................... 25
2.7 Localisation blur ........................................................... 27
2.8 Block Diagram of MPEG-1 Layer III (MP3) Encoder [3] ........ 29
2.9 Block Diagram of MPEG-2 AAC Encoder [27] .................... 32
2.10 Binaural Cue Coding [62] .............................................. 36
2.13 Directional Audio Coding [68] ........................................ 42
2.14 Two Dimensional VBAP .............................................. 43
2.15 Three Dimensional VBAP ............................................. 44
2.16 MPEG Surround Encoding ........................................... 46
2.17 MPEG Surround Downmixing to Stereo ............................ 49
2.18 MPEG Surround Downmixing to Mono ............................. 49
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.19</td>
<td>MPEG Surround Decoding</td>
<td>50</td>
</tr>
<tr>
<td>2.20</td>
<td>Soundfield Microphone</td>
<td>56</td>
</tr>
<tr>
<td>2.21</td>
<td>A sound source localised at azimuth $\mu$, elevation $\eta$</td>
<td>57</td>
</tr>
<tr>
<td>3.1</td>
<td>$S^3AC$ 360° to 60° Spatial Squeezing</td>
<td>70</td>
</tr>
<tr>
<td>3.2</td>
<td>Tangent Amplitude Panning</td>
<td>71</td>
</tr>
<tr>
<td>3.3</td>
<td>$S^3AC$ 5-Channel Encoding System</td>
<td>73</td>
</tr>
<tr>
<td>3.4</td>
<td>$S^3AC$ 5-Channel Decoding System</td>
<td>74</td>
</tr>
<tr>
<td>3.5</td>
<td>1024-Point Kaiser-Bessel Derived (KBD) window with $\alpha = 4$</td>
<td>75</td>
</tr>
<tr>
<td>3.6</td>
<td>5-Channel Sound System and Auditory Sub-Spaces</td>
<td>79</td>
</tr>
<tr>
<td>3.7</td>
<td>A ‘Squeezed’ Space Represented by Two Channels</td>
<td>79</td>
</tr>
<tr>
<td>3.8</td>
<td>A 5-channel signal rendering a mosquito flying around effect</td>
<td>83</td>
</tr>
<tr>
<td>3.9</td>
<td>A Time-Frequency-Azimuth Mesh of the Original Signal</td>
<td>85</td>
</tr>
<tr>
<td>3.10</td>
<td>A Time-Frequency-Azimuth Mesh of the $S^3AC$ Coded Signal</td>
<td>85</td>
</tr>
<tr>
<td>3.11</td>
<td>A Time-Frequency-Azimuth Mesh of the MPEG Surround 525 Coded Signal</td>
<td>86</td>
</tr>
<tr>
<td>3.12</td>
<td>A Time-Frequency-Azimuth Mesh of the MPEG Surround Non-Guided Coded Signal</td>
<td>86</td>
</tr>
<tr>
<td>3.13</td>
<td>Listening Test Results on General 5-Channel Audio Items</td>
<td>88</td>
</tr>
<tr>
<td>3.14</td>
<td>Listening Test Results on 5-Channel Audio Items with Strong Localisation Feature</td>
<td>91</td>
</tr>
<tr>
<td>4.1</td>
<td>Aliased Window During $S^3AC$ Analysis</td>
<td>93</td>
</tr>
<tr>
<td>4.2</td>
<td>Absolute Hearing Threshold, Source with 61dB Spectral Amplitude, Source with 49dB Spectral Amplitude and Minimum Masking Curve of a 76dB White Noise</td>
<td>100</td>
</tr>
<tr>
<td>4.3</td>
<td>Listening Test Results Comparing Different Sizes of the $S^3AC$ Downmix Soundfield</td>
<td>103</td>
</tr>
<tr>
<td>4.4</td>
<td>$S^3AC$ ‘Mono + Side Information’ Encoding</td>
<td>104</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.5</td>
<td>S³AC ‘Mono + Side Information’ Decoding</td>
<td>104</td>
</tr>
<tr>
<td>4.6</td>
<td>Listening Tests Results of Using Difference Localisation Precision at Different Listening Region</td>
<td>107</td>
</tr>
<tr>
<td>4.7</td>
<td>Illustration of the 5-Bit Codebook Design</td>
<td>109</td>
</tr>
<tr>
<td>4.8</td>
<td>Histogram of Codebook Azimuth Indices and Codebook Differential Indices</td>
<td>112</td>
</tr>
<tr>
<td>4.9</td>
<td>Average Cue Bit-Rate (in kbps) and 95% Confidence Intervals of Test Files for Different Prediction Lengths using 6-bit and 5-bit Codebook</td>
<td>113</td>
</tr>
<tr>
<td>4.10</td>
<td>Overlapping Sound Source in the 360° Soundfield</td>
<td>115</td>
</tr>
<tr>
<td>4.11</td>
<td>Overlapping Sound Source in the 60° Squeezed Soundfield and the Resulting Phantom Source</td>
<td>116</td>
</tr>
<tr>
<td>4.12</td>
<td>Localisation for Original, S³AC Coded and MPEG Surround Coded 1kHz tone source in the front region (azimuth between [-30° 30°]) in Test File 1</td>
<td>118</td>
</tr>
<tr>
<td>4.13</td>
<td>Localisation for Original, S³AC Coded and MPEG Surround Coded 1kHz tone source in the rear region (azimuth between [-180° -110°]∪[110° 180°]) in Test File 1</td>
<td>119</td>
</tr>
<tr>
<td>4.14</td>
<td>The Original, S³AC Coded and MPEG Surround 525 Coded Localisation for both Front and Rear Sources in Test File 2</td>
<td>120</td>
</tr>
<tr>
<td>4.15</td>
<td>2D Vector Analysis of Loudspeaker Signal Positioned at α, β</td>
<td>122</td>
</tr>
<tr>
<td>4.16</td>
<td>Tangent Amplitude Panning</td>
<td>124</td>
</tr>
<tr>
<td>4.17</td>
<td>2D Orthogonal Analysis of Loudspeaker Signal Positioned at α, β</td>
<td>126</td>
</tr>
<tr>
<td>5.1</td>
<td>S³AC Encoding System for Ambisonics Signals</td>
<td>133</td>
</tr>
<tr>
<td>5.2</td>
<td>S³AC Decoding System for Ambisonics Signals</td>
<td>134</td>
</tr>
<tr>
<td>5.3</td>
<td>Listening Tests Results Comparing S³AC and UHJ Compression of 2D Ambisonics</td>
<td>140</td>
</tr>
<tr>
<td>5.4</td>
<td>S³AC Binaural Reproduction System</td>
<td>142</td>
</tr>
<tr>
<td>5.5</td>
<td>HRTF Interpolation</td>
<td>144</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.6</td>
<td>Listening test results for S³AC binaural reproduction, comparing with original multi-channel playback</td>
<td>147</td>
</tr>
<tr>
<td>5.7</td>
<td>S³AC Based Spatialised Multi-Party Teleconferencing System</td>
<td>149</td>
</tr>
<tr>
<td>5.8</td>
<td>Four-Element Microphone Array for Speech Recording and Estimation of Speaker Azimuth</td>
<td>150</td>
</tr>
<tr>
<td>5.9</td>
<td>S³AC based Spatialised Teleconferencing Scenario with Three Participating Sites</td>
<td>153</td>
</tr>
<tr>
<td>5.10</td>
<td>S³AC based Spatialised Teleconferencing Scenario with Five Participating Sites</td>
<td>153</td>
</tr>
<tr>
<td>5.11</td>
<td>Scenario 1: Original Speaker Azimuths</td>
<td>158</td>
</tr>
<tr>
<td>5.12</td>
<td>Scenario 1: Estimated Speaker Azimuths from Two Remote Sites with Different Simulated Room Reverberation Time (Note: Legend from Figure 5.11 applies)</td>
<td>158</td>
</tr>
<tr>
<td>5.13</td>
<td>Scenario 1: ‘Squeezed’ Speaker Azimuths from Two Remote Sites Rendered at Site 1 (Note: Legend from Figure 5.11 applies)</td>
<td>159</td>
</tr>
<tr>
<td>5.14</td>
<td>Scenario 2: Original Speaker Azimuths (Note: microphones are hidden at circle center)</td>
<td>159</td>
</tr>
<tr>
<td>5.15</td>
<td>Scenario 2: Estimated Speaker Azimuths from Four Remote Sites with Different Simulated Room Reverberation Time (Note: Legend from Figure 5.14 applies)</td>
<td>160</td>
</tr>
<tr>
<td>5.16</td>
<td>Scenario 2: ‘Squeezed’ Speaker Azimuths from Four Remote Sites Rendered at Site 1 (Note: Legend from Figure 5.11 applies)</td>
<td>160</td>
</tr>
<tr>
<td>5.17</td>
<td>Scenario 2: Re-spatialised with Non-Uniform ‘Squeezed’ Sector Size Dependent on Number of Speaker for Each Remote Site</td>
<td>161</td>
</tr>
<tr>
<td>5.18</td>
<td>Squeezing Two Surround Soundfields into One Downmix</td>
<td>162</td>
</tr>
<tr>
<td>5.19</td>
<td>Listening test results for S³AC two-soundfields-to-one downmix compression of multiple spatial speech scenes</td>
<td>164</td>
</tr>
<tr>
<td>6.1</td>
<td>3D Loudspeaker Signal Positioned at azimuth $\mu$, elevation $\eta$</td>
<td>170</td>
</tr>
<tr>
<td>6.2</td>
<td>S³AC 3D Encoding System</td>
<td>173</td>
</tr>
<tr>
<td>6.3</td>
<td>S³AC 3D Decoding System</td>
<td>174</td>
</tr>
<tr>
<td>6.4</td>
<td>S³AC SLQP with High Precision</td>
<td>178</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.5</td>
<td>$S^3$AC SLQP with Low Precision</td>
<td>178</td>
</tr>
<tr>
<td>6.6</td>
<td>$S^3$AC SLQP with Low Precision while each SLQP is distinguished by color</td>
<td>181</td>
</tr>
<tr>
<td>6.7</td>
<td>$S^3$AC SLQP in a stereo downmix</td>
<td>181</td>
</tr>
<tr>
<td>6.8</td>
<td>$S^3$AC SLQP in a stereo downmix, zoom-in view of the left-hand side of Figure 6.7</td>
<td>182</td>
</tr>
<tr>
<td>6.9</td>
<td>16-Channel loudspeaker array CHESS in an anechoic environment</td>
<td>188</td>
</tr>
<tr>
<td>6.10</td>
<td>Designed source movement simulating airplane fly over-head effect</td>
<td>191</td>
</tr>
<tr>
<td>6.11</td>
<td>Time-frequency-elevation mesh of the original signal</td>
<td>191</td>
</tr>
<tr>
<td>6.12</td>
<td>Elevation feature of the 80$^{th}$ frequency</td>
<td>192</td>
</tr>
<tr>
<td>6.13</td>
<td>Azimuth/elevation 3D localisation feature of the 80$^{th}$ frequency</td>
<td>192</td>
</tr>
<tr>
<td>6.14</td>
<td>Time-frequency-elevation mesh of the signal encoded by $S^3$AC 3D mono downmix with High-precision SLQP quantisation</td>
<td>193</td>
</tr>
<tr>
<td>6.15</td>
<td>Time-frequency-elevation mesh of the signal encoded by $S^3$AC 3D mono downmix with Low-precision SLQP quantisation</td>
<td>193</td>
</tr>
<tr>
<td>6.16</td>
<td>Time-frequency-elevation mesh of the signal encoded by $S^3$AC 3D stereo downmix, SLQP recovered from stereo downmix soundfield</td>
<td>195</td>
</tr>
<tr>
<td>6.17</td>
<td>$S^3$AC 3D listening test results</td>
<td>198</td>
</tr>
<tr>
<td>7.1</td>
<td>A ‘Sound Source + Localisation’ Framework</td>
<td>205</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Mean Opinion Score ........................................ 64
2.2 MUSHRA Scoring Recommendation .......................... 65
3.1 An Example of S\(^3\)AC 360° to 60° Azimuth Squeezing Criteria .... 78
3.2 Average Decoded Azimuth Error of Three Evaluated Coders ........ 84
4.1 Average Decoded Azimuth Error of in the S\(^3\)AC ‘Squeezed’ Sound-field caused by Window Aliasing .......................... 95
4.2 6-Bit and 5-Bit Codebook Design for Location Dependent S\(^3\)AC Cue Quantisation ........................................ 108
4.3 Azimuth Error (in degrees) for Quantised S\(^3\)AC Cues, Compared with Original Derived Cues ....................... 114
4.4 Average Azimuth Error in Degrees, Comparing S\(^3\)AC, MPEG Surround and Original Front and Rear Source for Two Test Files ........ 121
5.1 W, X, Y Channel and Average Kullback-Leibler Spectral Distance of Eight 2D Ambisonics Recordings ......................... 139
6.1 CHESS loudspeaker configuration ............................. 188
6.2 Average azimuth/elevation error (in degrees) from S\(^3\)AC 3D mono coding with high-precision SLQP design ....................... 195
6.3 Average azimuth/elevation error (in degrees) from S\(^3\)AC 3D mono coding with low-precision SLQP design ....................... 195
6.4 Average azimuth/elevation error (in degrees) from S\(^3\)AC 3D stereo coding ........................................ 198
A.1 Higher order Ambisonics up to $4^{th}$ order and the related spherical harmonics functions [83] .......................... 220
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>Two Dimensional</td>
</tr>
<tr>
<td>3D</td>
<td>Three Dimensional</td>
</tr>
<tr>
<td>AAC</td>
<td>Advanced Audio Coding</td>
</tr>
<tr>
<td>BCC</td>
<td>Binaural Cue Coding</td>
</tr>
<tr>
<td>CD</td>
<td>Compact Disc</td>
</tr>
<tr>
<td>CHESS</td>
<td>Configurable Hemisphere Environment for Surround Sound</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DirAC</td>
<td>Directional Audio Coding</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>GCC-PHAT</td>
<td>Generalized Cross Correlation PHAse Transform</td>
</tr>
<tr>
<td>HE-AAC</td>
<td>High-Efficiency Advanced Audio Coding</td>
</tr>
<tr>
<td>HRIR</td>
<td>Head-Related Impulse Response</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head-Related Transfer Functions</td>
</tr>
<tr>
<td>ICC</td>
<td>Inter-Channel Coherence</td>
</tr>
<tr>
<td>ICLD</td>
<td>Inter-Channel Level Difference</td>
</tr>
<tr>
<td>ICPD</td>
<td>Inter-Channel Phase Difference</td>
</tr>
<tr>
<td>ICTD</td>
<td>Inter-Channel Time Difference</td>
</tr>
<tr>
<td>iDFT</td>
<td>inverse Discrete Fourier Transform</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>ILD</td>
<td>Inter-Aural Level Difference</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ITD</td>
<td>Inter-Aural Time Difference</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>KBD</td>
<td>Kaiser-Bessel Derived</td>
</tr>
<tr>
<td>kbps</td>
<td>kilo bits per second</td>
</tr>
<tr>
<td>LFE</td>
<td>Low Frequency Effect</td>
</tr>
<tr>
<td>MDCT</td>
<td>Modified Discrete Cosine Transform</td>
</tr>
<tr>
<td>MOS</td>
<td>Mean Opinion Score</td>
</tr>
<tr>
<td>MP3</td>
<td>MPEG-1 Layer 3</td>
</tr>
<tr>
<td>MPEG</td>
<td>Moving Pictures Expert Group</td>
</tr>
<tr>
<td>MPS</td>
<td>MPEG Surround</td>
</tr>
<tr>
<td>MUSHRA</td>
<td>MUltiple Stimuli with Hidden Reference and Anchor</td>
</tr>
<tr>
<td>PEALQ</td>
<td>Perceptual Evaluation of Audio Quality</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>PS</td>
<td>Parametric Stereo</td>
</tr>
<tr>
<td>PQMF</td>
<td>Pseudo Quadrature Mirror Filters</td>
</tr>
<tr>
<td>QMF</td>
<td>Quadrature Mirror Filters</td>
</tr>
<tr>
<td>S³ AC</td>
<td>Spatially Squeezed Surround Audio Coding</td>
</tr>
<tr>
<td>SLQP</td>
<td>Spatial Localisation Quantisation Points</td>
</tr>
<tr>
<td>SRP-PHAT</td>
<td>Steered Response Power with PHAse Transform</td>
</tr>
<tr>
<td>STFT</td>
<td>Short-Time Fourier Transform</td>
</tr>
<tr>
<td>TDE</td>
<td>Time-Delay Estimation</td>
</tr>
<tr>
<td>VBAP</td>
<td>Vector Based Amplitude Panning</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

Efficient compression of auditory content has always been a very active research area. In the last few decades, extensive researches has been made into compressing mono/stereo audio formats, utilizing human ear perception models. This led to several efficient and widely used techniques, e.g. MPEG-1 Layer 3 [1] (also well known as MP3), Advanced Audio Coding [2] [3] (also known as AAC), which can achieve near transparent decoding quality with bit-rate as low as 64kbps per audio channel.

In recent years, there is significant increase in the deployment of audio content and playback systems with multiple channels. Such application scenarios include movie theatre, home entertainment, Internet music and movie distribution, teleconferencing, etc. While conventional audio compression techniques, such as MP3 and AAC, process and transmit each audio channel individually, it is not desirable to use these techniques in multi-channel audio applications, especially when the following facts are considered:

- The transmission and storage cost increase linearly with the number of audio channels. For example, extending an audio application from stereo to the ITU 5.1-channel format [4] requires three times as much as the bandwidth
required in stereo applications, if conventional MP3 and AAC techniques are used. Future 3D audio applications will require further increases in the number of channels

• Similar to transmission and storage costs, the computational cost in processing multi-channel audio content will also increase linearly with the number of channels. It is especially undesired in real-time applications, such as teleconferencing.

• While soundfield dynamics and source localisation in multi-channel audio is the key improvement over conventional stereo/mono audio, and have to be rendered across multiple channels, these features can be distorted by coding each channel separately.

Recent development in audio coding techniques have focused on exploiting the differences and correlation between multiple channels. Investigations started from removing the redundancy between two stereo channels, so that less bandwidth is required for transmitting a stereo signal, e.g Parametric Stereo [5] [6] and the resulting MPEG HE-AAC [7]. Further studies were performed on the compression of multi-channel audio signals, typically ITU 5.1-channel audio, e.g. Binaural Cue Coding [8], MPEG Surround[9]. In these techniques, a typical approach is to derive side information (also called spatial cues) in multi-channel audio content representing inter-channel time difference, inter-channel level difference, and inter-channel coherence, while a downmix version of the original audio is created by mixing all original channel into less channels, typically stereo or mono. By analyzing the inter-channel relations saved in the spatial cues, the decoder is able to expand the downmix to multi-channel formats with the original surround sound listening experience recovered. This approach has the following advantages over conventional coding techniques:

• The total transmission bandwidth is determined by the bandwidth of the down-
mix signal and the size of the side information. Since the side information represents arithmetic relationships between audio channels, it can be quantized and compressed into a compact set of metadata, requiring a bandwidth less than 10kpbs [10]. In addition, an increased number of original audio channels does not significantly increase the size of the side information. By downmixing multiple audio channels into a stereo/mono format, the transmission bandwidth of this approach is hence less dependent on the number of original audio channels, compared with the conventional approaches where each channel is coded individually.

- While little bandwidth is required for transmitting the side information, the stereo/mono downmix can also be further compressed by conventional coding techniques such as MP3 or AAC. This results in transmitting multi-channel audio at a bit-rate only slightly higher than the conventional stereo/mono audio.

- The stereo/mono downmix, as well as in the format coded by conventional MP3/AAC, can also be played back seamlessly in a conventional system, in the circumstances that a multi-channel speaker system is not fitted. Backward compatibility is hence achieved.

These techniques open a new research area in audio processing called Spatial Audio Coding (SAC), aiming at providing efficient compression of multi-channel audio signals with spatial sound content. Essential considerations when designing a spatial audio coder include, bandwidth requirement being independent of numbers of channels, backward compatibility to conventional audio systems and efficient recoverability of the original surround sound scene, especially the spatial information and sound source localisation content. Further considerations include computational complexity, algorithm flexibility between multiple spatial audio formats, as well as extensibility for future formats.

Whilst having advantages mentioned above, existing SAC approaches have some
drawbacks. In these techniques, all original spatial information is saved in the spatial cues as side information. Hence, the original surround sound scene cannot be decoded with only the existence of the downmix, since no spatial information can be retrieved without the spatial cues. Being the critical part of these approaches, the spatial cues not only add extra bandwidth, but also lead to other problems. For example, these approaches fail to analyse the spatial sound field and sound source localisation from the viewpoint of perceptual sound localisation based on psychoacoustics. Rather, the spatial cues containing localisation information in these approaches are extracted by exploiting and deriving the mathematical relationships between audio channels. As a result, the spatial cues do not represent any sound source localisation information in the original soundfield. Hence, quantisation and compression on the spatial cues can cause unpredictable localisation errors. In addition, while inter-channel spatial cue derivation is the critical process in such approaches, these techniques become less extendable when the number of audio channels changes or increases. In other words, the spatial cue derivation algorithm in these techniques has to be significantly altered when the spatial audio signal format changes, e.g. the coding algorithm for 5.1-channel audio [4] is not suitable for coding 7.1-channel audio [11] [12] [13] in these approaches. In addition, computation complexity in these approaches increases linearly with increased number of audio channels, as spatial cue derivation is required between every pair of channels. Furthermore, adding extra bitstreams for spatial cues in audio frames results in non-backward-compatible frames for conventional audio decoding systems.

### 1.2 Spatially Squeezed Surround Audio Coding

As a result of this thesis, a novel spatial audio coding technique has been developed, named Spatially Squeezed Surround Audio Coding - (S3AC). Based on the analysis of spatial sound source localisation from the rendering of multichannel spatial audio signals, this technique saves sound source localisation information in a reproduced
downmix soundfield, which can be typically rendered by a stereo signal. As a re-
result, the required transmission bandwidth is reduced from multiple channels to two 
channels, similarly to other SAC solutions, while the stereo downmix signal is back-
ward compatible and can be further compressed by perceptual coders such as MP3 
or AAC. Nevertheless, rather than deriving inter-channel mathematical relationships,
the S\textsuperscript{3}AC technique derives surround soundfield and source localisation information 
from a spatial audio signal. The advantages of this approach, compared with existing 
SAC approaches, are briefed in the followings and will be further addressed in this 
thesis.

- Based on psychoacoustical principles, the S\textsuperscript{3}AC soundfield analysis on a spa-
tial audio signal provides angular information of surround sound source local-
isation contained in the original soundfield, typically azimuth/elevation inform-
ation. Hence, the resulting source localisation information can be further 
processed and compressed efficiently, based on perceptual localisation the-
ories, while the localisation distortion caused by quantisation is predictable. 
This also ensures that minimum localisation distortion is introduced during the 
compression by exploiting perceptual localisation theories.

- The stereo downmix soundfield is exploited in S\textsuperscript{3}AC to save the localisation in-
formation derived from the original surround soundfield. Hence, no extra side 
information is required, while stereo backward compatibility is maintained as 
well. In addition, S\textsuperscript{3}AC also provides a mono downmixing compression mode 
so as to further reduce the bandwidth, while source localisation information is 
saved as side information. In comparison with existing SAC approaches, the 
S\textsuperscript{3}AC side information directly represents perceptual localisation information 
of the original surround sound sources, which can be efficiently quantised and 
compressed as addressed above.

- Analyzing the input spatial audio signal from a ‘rendering soundfield’ point
of view, the $S^3$AC analysis algorithm has minimum dependency on the format of the original spatial audio signal. In fact, $S^3$AC is capable of analyzing and compressing any multichannel audio signal format as long as the loudspeaker channel location is given, while no change on $S^3$AC’s fundamental localisation analysis algorithm is required. In addition, the computational cost increases linearly with increased number of channels in $S^3$AC analysis, while efficient compression of 3D spatial audio signal with large number of channels using $S^3$AC is also introduced in this thesis.

1.3 Thesis Outline

This thesis is organized as follows:

Chapter 2 provides a literature review of relevant work in the area of digital audio coding techniques, especially in spatial and surround audio techniques and compression approaches. An investigation of fundamental digital signal processing and digital audio processing will be given, followed by a review of conventional perceptual audio coding techniques. A review on relevant psychoacoustical theories is given. This chapter then describes earlier and current developments in the field of spatial and surround audio processing.

Chapter 3 starts by presenting the fundamental psychoacoustic principles exploited in the $S^3$AC technique. This is followed by a detailed introduction to compressing ITU 5.1-channel audio signals using $S^3$AC techniques, where implementation details are also presented. This chapter then presents evaluation results for the $S^3$AC approach to compressing ITU 5.1-channel signals.

Chapter 4 provides further analysis and investigation into the $S^3$AC spatial audio coding approach. Window aliasing problems discovered in the $S^3$AC coding process is analyzed. Source localisation resolution in $S^3$AC stereo downmix is investigated
as well as the limitation of the S\textsuperscript{3}AC spatial squeezing process. A ‘mono downmix + cues’ approach is then introduced based on S\textsuperscript{3}AC fundamental principles, while a psychoacoustical localisation based cue quantisation approach is developed. A novel source localisation estimation algorithm based on 2D Cartesian orthogonal analysis is derived, which provides extended flexibility for estimation of sound source localisation information in S\textsuperscript{3}AC.

Chapter 5 presents some extended applications of S\textsuperscript{3}AC. These include using S\textsuperscript{3}AC to compress Ambisonics [14] spatial audio signals, binaurally reproducing S\textsuperscript{3}AC compressed surround audio signals and the application of the S\textsuperscript{3}AC technique in a multi-party teleconferencing environment.

Chapter 6 extends the S\textsuperscript{3}AC technique from compressing 2D spatial audio to 3D spatial audio with increased number of loudspeaker channels. The 2D Cartesian orthogonal analysis is extended for deriving 3D source localisation rendered by arbitrary numbers of loudspeaker channels. Based on this, the S\textsuperscript{3}AC compression approach for 3D spatial audio is detailed and evaluated.

Chapter 7 provides a conclusion to the thesis and details future work.

1.4 Contributions

The main contributions as a result of this thesis are:

1. Introduced an efficient, low bit-rate and backward compatible approach to compressing spatial audio signals, named S\textsuperscript{3}AC.

2. Exploited psychoacoustical localisation theories to justify the fundamental algorithm in S\textsuperscript{3}AC.

3. Developed algorithms to exploit the stereo downmix soundfield, so as to save
localisation information from an original surround soundfield. Transmission of side information is avoided as a result.

4. Developed an efficient quantisation process based on psychoacoustical theories for compressing $S^3AC$ side information for scenarios where a mono downmix is required.

5. Developed an approach to utilize side information to address problem caused by overlapping time-frequency components found in $S^3AC$ compression.

6. Developed compression schemes for Ambisonics spatial audio recording signals, based on $S^3AC$. In addition to bandwidth reduction, this approach provides not only improved stereo backward compatibility for Ambisonics, but flexibility for decoding to multiple playback formats.

7. Developed reproduction techniques for $S^3AC$ compressed signals in binaural playback environments. This provides spatial audio playback experiences for users without multi-channel loudspeaker systems.

8. Investigated and justified perceptual localisation distortion caused by $S^3AC$ compression. The perceptual impact caused by $S^3AC$ compression, including both spatial squeezing to stereo downmix soundfield and psychoacoustical based side information quantisation, is investigated.

9. Exploited the limitation and capability of a stereo downmix soundfield. The limitation of $S^3AC$ spatial squeezing approach to a stereo downmix soundfield is justified. An approach to compressing multiple surround soundfields utilizing single stereo downmix signala is developed based on the $S^3AC$ algorithm.

10. Developed an $S^3AC$ based technique for low bit-rate transmission and efficient reproduction of localised speech content in a multi-site teleconferencing scenario.
11. Proved the fundamental equivalence between multiple spatial audio panning and reproduction techniques. A general panning algorithm for both 2D and 3D spatial audio is drawn as a result.

12. Developed an efficient, low bit-rate and backward compatible approach for compressing 3D spatial audio rendered by arbitrary number of loudspeaker channels, based on the S$^3$AC technique.

1.5 Publications

The following publications were a result from this thesis:

1.5.1 Journal Publications


1.5.2 Book Chapters


1.5.3 Conference Publications


Chapter 2

Literature Review

2.1 Introduction

This chapter investigates the background and current development in the area of spatial audio coding. The investigation starts from fundamental principles in digital audio processing, followed by a review on the related psychoacoustical theories.

This chapter then diverts its attention to the classical stereo/mono audio coders, including MP3, AAC and related MPEG Standards. Investigation on the fundamental perceptual theories of these techniques is also reviewed. This is followed by a focused investigation on recent researches on spatial audio coding, while major techniques and developments in this area are reviewed, including Binaural Cue Coding (BCC) [15] [8], Parametric Stereo [5], MPEG Surround [9] and Directional Audio Coding [16].

The Ambisonics [14] [17] spatial audio recording and reproduction technique is also reviewed in this chapter. Finally, some audio quality evaluation methodologies and approaches are reviewed.
2.2 Digital Audio Processing Fundamentals

2.2.1 Digital Representation of Audio Signals

Human ears perceive sounds from the surrounding environments as one of our vital ways of perception. Physiologically, sound is a pressure wave that causes vibration on our eardrum, which is transformed into neural information recognizable for our brain. The need for storing and replaying sounds was first satisfied by the analog recording and storage techniques introduced more than a century ago. Following the rapid development of digital and computer technologies in the last few decades, sound has been measured, recorded and stored in the form of digital information, fundamentally as simple as some zeros and ones. This not only significantly improves the efficiency of recording, storing and playing back sounds, but introduces computer powers for processing sounds in any way that an intelligent individual can design. This includes but not limited to compression (for reducing the storage cost), manipulating (to change the characteristics of sounds), mixing (combining several sounds together) and even artificially synthesize sounds that do not naturally exists.

There are two steps to transfer a continuous sound wave into digital form, sampling and quantisation, which briefly refer to the digitalizing technique for transforming time-varying information and pressure information, respectively, of a sound wave into numerically finite information.

**Sampling** refers to the process that a continuous-time sound wave is transformed into a discrete-time signal by taking a sample of the sound wave between every time-interval. The reciprocal of a fixed signal sampling time-interval is called the sampling frequency of a signal. In audio processing, for example, a typical sampling frequency of 44.1kHz refers that forty-four thousand and one hundred samples are taken every second.

**Quantisation** refers to the process that a finite set of pre-defined values is used to
describe a signal with infinite value. The bigger a set of finite values is used in a
quantisation process, the higher quantisation precision is achieved. In audio process-
ing, a quantisation precision is generally described in binary terms. For example,
a 16-bit quantisation precision refers that 16 bits are assigned to a signal sample to
describe the sound pressure level information.

Figure 2.1 gives an illustration of the sampling and quantisation process, which ef-
effectively transform an analog audio to digital signal. The process of sampling and
quantisation is also named pulse code modulation, PCM [3], which is the signal for-
mat widely used in the Compact Disk (well known as CD) of digital audio, where
sampling frequency and quantisation precision are standardly defined as 44.1kHz and
16-bit, respectively.

2.2.2 Mono, Stereo and 2D Multichannel Audio Formats

The sampling and quantisation process described in Section 2.2.1 provides the typi-
cal approach for digital representation of a signal sound wave. This results in a single
channel of digital audio signal, referred as mono signal. A stereo signal refers to the
signal format containing two independent channels of audio information, generally
left and right. Compared with mono audio, stereo audio provides improved impres-
sion of a soundfield capable of rendering sound localisation and wideness informa-
tion. The stereo audio is especially beneficial for portable personal audio devices,
where audio is played back by earphone or headphone (as we only have two ears).
It is also widely used in modern worlds, e.g. FM broadcasting, TV broadcasting,
home audio system, computer music and online music distribution, due to its limited
Multi-channel audio usually refers to the audio formats that contain more than two channels. Introduced after stereo audio, multi-channel audio especially aims at further improving the sound localisation capability. A typical multi-channel audio format is the ITU 5.1-channel audio [4], which is illustrated in Figure 2.2. Besides the front left/front right channels located at ±30° as in stereo audio, the center channel is usually used for playing back voice contents, e.g. for movie audio, while the rear left/rear right channels located at ±110° are used for extended sound source localisation rendering. An extra subwoofer channel (the .1 channel, also known as LFE channel) is dedicated for playing back low-frequency contents.

Beside the ITU standard definition on 5.1-channel formats, there are various definitions on multi-channel audio formats given by organizations. A commonly known format, as an improvement to ITU-5.1, is a 7.1-channel system, while similar but slightly different definition is given by different companies, e.g. Dolby [13], DTS [12] and Microsoft [11]. However, this thesis will be mainly focused on the standardized ITU-5.1 format for the 2D spatial audio, while the extensibility to other 2D multichannel formats of presented approaches will be further addressed in later chapters.
2.2.3 Time-Frequency Representation of Audio Signals

Waveform represents the time domain varying information of a sound and it is the form in which a sound propagates as well as can be played back by equipments or perceived by ear. In order to obtain the frequency information of a signal, a time-frequency transform is required. A basic approach is the Fourier Transform defined as:

\[ X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt \]  \hspace{1cm} (2.1)

where \( x(t) \) and \( X(f) \) are the time and frequency domain representation of a continuous signal respectively. The inverse Fourier Transform that transforms frequency-domain signal to time-domain is given by:

\[ x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi ft}df \]  \hspace{1cm} (2.2)

For digital signals, the time-frequency transform can be achieved using Discrete Fourier Transform (DFT):

\[ X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N} \]  \hspace{1cm} (2.3)

where \( x[n] \) and \( X[k] \) are the time and frequency domain representation of a discrete signal respectively, \( N \) is the total number of samples. The frequency-time transform is achieved using inverse Discrete Fourier Transform (iDFT):

\[ x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]e^{j2\pi kn/N} \]  \hspace{1cm} (2.4)

While the DFT representation provides only the frequency-domain information of
a signal, both time and frequency domain information is required for audio signal processing. This can be achieved using Short-Time Fourier Transform (STFT). In STFT, a discrete audio signal is divided into frames with equal length of $2M$ samples, while each frame is overlapped with adjacent frames. A typical approach is to have 50% overlapping between adjacent frames, i.e. a following frame is acquired by shifting $M$ samples from the last frame, as illustrated in Figure 2.3. In this process, each frame is multiplied by a analysis window function with the same length on a sample-by-sample basis, such that, for the $p^{th}$ frame, the windowed signal is:

$$y[n] = x[n] \cdot w[n] \quad \text{for } n = (p - 1) \cdot M + 1 \ldots (p - 1) \cdot M + 2M \quad (2.5)$$

where $x[n]$ and $w[n]$ represent the original signal and the analysis window respectively. The resulting frame signal is transformed by a $2M$ point DFT similar as in Eq. 2.3 for further analysis:

$$Y[k] = \sum_{n=0}^{2M-1} y[n] e^{-j2\pi kn/2M} \quad (2.6)$$

The time-domain frame signal can be recovered from the frequency domain by:

\footnote{assuming that $M$ is an even number}
which contains both the original signal and the analysis window. Generally, the window function is symmetrical and designed to remove frequency domain aliasing caused by sharp cut-off on the time-domain frame edges. In order to remove the signal distortion caused by applying the analysis window, a synthesis window is applied on the resulting time-domain signal from Eq. 2.7 and an overlap-and-add approach is introduced. Typically, it is designed that the analysis window and synthesis window are identical, and in the 50% overlapping implementation, the window function satisfies:

\[ w^2[n] + w^2[n + M] = 1 \quad \text{for} \quad n = (p - 1)\cdot M + 1 \ldots (p - 1)\cdot M + M \quad (2.8) \]

Based on this, the original signal samples from \((p - 1)\cdot M + 1\) to \((p - 1)\cdot M + M\) can be recovered by adding the second half of the \((p - 1)^{th}\) frame and the first half of the \(p^{th}\) frame after the synthesis windowing, and pro-rata for other samples. It is noted that, zero padding is required at the beginning and tail of the original signal so that the whole original signal can be recovered.

While providing a good time-frequency representation of a signal, the STFT analysis described above increases the data rate of the signal due to the procedure of taking overlapping samples between adjacent frames. In the given example where a 50% overlapping is used, the data rate becomes twice as much as the original signal. To avoid this, the Modified Discrete Cosine Transform (MDCT) was introduced [18] [19], which provides the same data rate as the input signal (critical sampling) in the 50% overlap-and-add implementation. This is achieved by further constraining the analysis and synthesis windows to be time-reversed copies of each other [3] as well as using a real valued cosine basis function instead of the complex valued basis function.
for time-frequency transform. This results in that each frame with $2M$ time-domain samples can be represented by $M$ frequency-domain coefficients, while time-domain perfection reconstruction can be achieved by overlap-and-add procedure after inverse transform. Details of MDCT can be found in [3].

Besides the transform based time-frequency analysis techniques described above, some modern audio coding solutions utilize sub-band filter bank techniques. Similarly, the filter bank analysis also requires perfect reconstruction condition for the signal. A typical solution is the Quadrature Mirror Filters (QMF) [20], in which the input signal is separated by a pair of low-pass and high-pass filters having "mirroring" frequency features. This can be expressed as the Z transform relation between the low-pass filter $h_0[n]$ and the high-pass filter $h_1[n]$:

$$H_1[z] = -H_0[-z] \quad (2.9)$$

Perfect reconstruction can be achieved by applying synthesis filters satisfying the following:

$$G_0[z] = -H_0[z]$$
$$G_1[z] = -H_0[-z] \quad (2.10)$$

This results in a pair of two-channel QMF filters, while each channel can be further decomposed by sub-band filters for additional frequency resolution. However, inefficient filter implementation and high computational cost is required if sufficient frequency resolution for perceptual audio coding analysis is achieved using QMF filter banks. A solution called pseudo-QMF (PQMF) filter banks was introduced [21] [22], which provides efficient implementation and low complexity for subband filtering while near-perfect reconstruction is achieved. The PQMF utilizes a narrow band
low-pass filter and its modulated versions for frequency decomposition, resulting in multiple frequency bands with even bandwidth. It is widely used in modern audio coding techniques, such as MP3 [1] and AAC [2].

Similar as existing audio coders, time-frequency transform approaches presented in this subsection is used in the \( S^3 \)AC techniques proposed in this thesis for estimating time-frequency representation of an input audio signal for further analysis.

2.3 Fundamental Psychoacoustics

2.3.1 Lossless and Lossy Audio Coding

For audio signals, distortion is introduced during the sampling and quantisation process, when comparing a digital signal to its original analog waveform. The distortion can be limited by increasing the sampling rate and quantisation precision. In the audio CD standard, mentioned in Section 2.2.1, a sampling rate of 44.1kHz is used, which is adequate to preserve frequency information up to 22.05kHz, according to Nyquist theorem [23]. This well covers the average perceptual hearing range from approximately 20Hz to 20kHz, as suggested in [24]. Though the quantisation precision of 16-bit used in audio CD is not sufficient for all listeners [3], the 44.1kHz sampling rate and 16-bit quantisation precision in the audio CD standard is widely used as the original un-coded condition for digital audio signals. In this thesis, the CD quality of digital audio signal described above is used as the reference original condition for each channel unless otherwise described.

Generally, modern audio compression techniques can be categorized into two types, lossless compression and lossy compression. In lossless audio compression, e.g. Monkey’s Audio [25], Apple Lossless [26], only the signal redundancy is removed so that the original signal can be perfectly recovered after compression. Based on this, modern lossless audio coder can reduce the data rate to approximately half of
the bandwidth required for a channel of PCM audio signal. In addition, for multi-channel audio signals, a lossless compression further refers that each original channel is recovered perfectly after compression.

In comparison, lossy audio coders can achieve significant higher compression ratio compared with lossless audio coders by removing the audio signal irrelevancy. Although the original signal cannot be perfectly recovered, the goal of designing a lossy audio coder is to ensure minimum perceptual distortion is introduced during compression by intelligently exploiting psychoacoustical principles. In other words, a lossy audio coder removes the auditory information that cannot be perceived by human ears and reduces the bandwidth cost for the auditory information that has less perceptual importance, so that a significant compression can be achieved. For example, an AAC coder requires 64kbps for transmitting one channel of audio signal in ‘perceptually transparent’ quality [27], compared with more than 700kbps for one channel of CD format audio (Section 2.2.1). Here, ‘perceptually transparent’ quality is commonly used in literatures to describe that no perceptual distortion is introduced by a lossy audio coder. Generally, a lossy audio coder that exploits perceptual principles for signal compression is also referred as a perceptual audio coder.

Based on the success of conventional mono/stereo coders such as MP3, AAC, exploiting the right psychoacoustical principle in the right way is one of the main challenges in developing an efficient perceptual audio coder. This applies to developing a spatial audio coder as well. In this thesis, some psychoacoustical principles on spatial and localised sound perception will be investigated and further exploited in developing the proposed $S^3$AC technique.

2.3.2 Hearing Threshold

Human ear has limited perception range of sound intensity, from the bottom level of detecting a quiet stimulus to the top threshold of a loud stimulus causing pain. In
order to quantify the level of a sound stimuli, a sound pressure level (SPL) in decibels is defined as:

$$L_{SPL} = 20\log_{10} \frac{p}{p_0}$$  \hspace{1cm} (2.11)

where $L_{SPL}$ is the SPL, $p$ is the sound pressure in Pascals and $p_0$ is the standard reference level of $20 \mu Pa$ [28].

Based on this, the absolute hearing threshold (in dB SPL) for pure tone sound source in a noiseless environment can be approximately modeled by a frequency dependent function [28]:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$  \hspace{1cm} (2.12)

where $T_q$ is the absolute hearing threshold as a function of frequency $f$ in Hertz. This is also illustrated in Figure 2.4.

While this curve indicates the ideal perceptual limits for signal quantisation without introducing audible noise, it is worthwhile mentioning that the absolute hearing threshold is derived based on experiments carried out in an ideal noiseless environment with pure tone stimulus [29] [28], while realistic audio compression faces significantly more complicated auditory environment. In addition, with no prior knowledge of actual hardware information and playback level, the lowest level point in the absolute hearing threshold curve (at approximately 4kHz as shown in Figure 2.4) is usually referenced to $\pm 1$ bit of the audio signal amplitude [28].

### 2.3.3 Perceptual Filterbanks

Considering the frequency dependent property of human ear, perceptual audio coders utilizes proper filter bank techniques that well match the frequency features of the
human auditory perception. One well-known example is the critical banks, which mathematically models the frequency characteristics of the cochlea. In this model, the perceptual property of cochlea is modeled as a discrete set of band-pass filters, while sound sources within pass-band is considered to remain constant loudness for perception. According to [30], the bandwidth of a critical band can be expressed as a function of the center frequency as:

$$BW_c(f) = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69}$$  \hspace{1cm} (2.13)$$

where $BW_c$ is the bandwidth in Hertz as a function of the center frequency $f$ in Hertz. In perceptual audio coding, a set of filter banks (with information of center frequencies and related bandwidths) is usually predefined while satisfying Eq. 2.13. A common approach is to transform frequency scale in Hertz into a ‘Bark’ scale, which refers to discrete indices of critical bands. This is calculated as:
\[ z(f) = 31\arctan(0.00076f) + 3.5\arctan^2(f/7500) \]  

(2.14)

where \( z(f) \) is the Bark scale as a function of center frequency \( f \) in Hertz. In this approach, for instance, the center frequency of the fifth critical band can be calculated by substituting \( z(f) \) by 5 in Eq. 2.14, resulting in a center frequency of approximately 530Hz and a critical bandwidth of approximately 170Hz.

Besides the critical bands, the equivalent rectangular bandwidth (ERB) is another widely used perceptual filterbank technique in perceptual audio coding, which can be modeled as a function of center frequency \( f \) [31]:

\[ ERB(f) = 24.7(4.37(f/1000) + 1) \]  

(2.15)

In comparison with critical bands, the ERB filterbank provides higher frequency resolution for low frequencies, especially for frequencies below 500Hz. This provides benefits in better low frequency selectivity than critical bands, which results in more efficient frequency dependent quantisation when designing a perceptual audio coder. Figure 2.5 gives an illustration of the comparison between critical bandwidth (calculated by Eq. (2.13)) and ERB (calculated by Eq. (2.15)) as functions of center frequency.

### 2.3.4 The Masking Phenomenon

In audio coding, masking refers to the situation that one sound source becomes inaudible due to the presence of another sound source. In general, the masking phenomenon can be categorized into two types, simultaneous masking and non-simultaneous masking, based on the temporal feature of the masker and maskee. Simultaneous masking refers to the situation that, within a temporal frame, a strong sound component (the masker) makes weak components with nearby frequency (the
maskees) inaudible. Based on the spectral characteristics of the masker and maskee, simultaneous masking can be further categorized into three types, noise-masking-tone, tone-masking-noise and noise-masking-noise, details of which have been addressed in [32] [33] [34] [28]. On the other hand, non-simultaneous masking refers to the situation that the masking effect introduced by a strong sound components spreads to prior frames and later frames, named respectively as backward masking and forward masking [35]. The simultaneous masking and non-simultaneous are exploited in many modern audio codecs, such as MP3 and AAC [28], which are used in this thesis for compressing a downmix signal generated by the proposed S³AC technique.

### 2.3.5 Binaural Listening and HRTF

A localised sound source results in differences between the arrival signals at the left and right ears, i.e. the left/right ear entrance signals. By processing the difference
Figure 2.6 Free-field single sound source propagation on a horizontal surface and the left/right ear entrance signals

and correlation between the left and right ear entrance signals, the human auditory system is able to identify the direction of the sound source. Figure 2.6 illustrates a simplest scenario, where a sound source propagates horizontally to the left and right ear without reflection and reverberation (i.e. in a free-field environment). In this case, the difference between the sound propagation distance from the source to the left and right ear, $d_2 - d_1$, results in an arrival time difference at the left and right ears. This is referred as inter-aural time difference (ITD). Additionally, the head shadowing effect causes reduced sound pressure level at the right ear entrance signal, compared with the left ear entrance signal, which is referred as inter-aural level difference (ILD) [36]. Psychoacoustical principles further indicate that, when comparing with each other, ITD has higher significance for perceptual localisation of low frequency sources, while ILD plays a more important role at higher frequencies. This gives a fundamental study of binaural listening and localisation in an ideal
environment, while a common listening environment usually introduces reflection, reverberation and multiple sources. For such scenario with higher complication, a good simulation model can be found at [37].

In fact, the left/right ear entrance signals can be modeled as filtered versions of the original sound source signal. The filter that characterizes this model is referred as head-related transfer function (HRTF) [38]. In the example given in Figure 2.6, the left/right ear entrance signals, $S_{left}$ and $S_{right}$, can be described in the frequency domain as:

\[
\begin{bmatrix}
S_{left} \\
S_{right}
\end{bmatrix} = \begin{bmatrix}
H_{left} \\
H_{right}
\end{bmatrix} \cdot S^T
\]  

(2.16)

where $S$ is the source signal; $H_{left}$ and $H_{right}$ are the arrays representing the left and right HRTF respectively. Eq. (2.16) can be represented in time-domain equivalently as:

\[
\begin{align*}
S_{left} &= h_{left} * S \\
S_{right} &= h_{right} * S
\end{align*}
\]

(2.17)

where symbol $*$ denotes convolution; $h_{left}$ and $h_{right}$ are the left and right head-related impulse response (HRIR).

In modern audio processing, the principles of HRTF are utilized to synthesize stereo signals for localised headphone playback. By filtering a monophonic source signal with a pair of left/right HRTFs, the resulting stereo headphone signal can virtually simulate the perception of a localised sound source. While the actual HRTF varies between each individual, a set of HRTF database based on pre-measurements over dummy-head can provide efficient approximation for general audio coding applica-
Displacement of a Sound Source

![Figure 2.7 Localisation blur](image)

The human auditory system has limited localisation resolution. In other words, a slight displacement of a sound object can not be perceptually discovered, if the displacement is small enough. In this case, the perceptual auditory system recognizes

Figure 2.7 Localisation blur

...
this sound source at a fixed position. This phenomenon is referred as localisation blur and is illustrated in Figure 2.7. In addition, psychoacoustics indicate that human sound localisation precision is highly dependent on source location [36]. In an ideal listening environment, the localisation precision on the horizontal plane is approximately 1° in front of a listener and reduces to less than 10° at the sides and rear, while vertical precision reduces from approximately 3° at 0° elevation to approximately 10° at 90° elevation [36].

The localisation blur is the fundamental principle exploited by the proposed S3AC technique for efficiently removing perceptually irrelevant localisation information. This will be further discussed in Chapter 3.

2.3.7 Spatial Unmasking

The masking phenomenon reviewed in Section 2.3.4 only considers the scenarios where the masker and maskee are spatially overlapped, i.e. at a same location. When the masker and maskee are spatially displaced, it is possible that the maskee becomes audible. The amount of spatial unmasking is defined as the change in maskee energy at threshold of being detected for a particular location, compared to when the maskee is at the same location as the masker [41]. In other words, the spatial unmasking describes the ‘improvement’ of the detection threshold of the maskee, when it moves away from the masker. Measurements of spatial unmasking properties of some typical frequency components are described in [41] and [42]. Due to the natural complication of the spatial unmasking, no literature is found to fully quantify the amount of spatial unmasking for all frequencies and all spatial locations, nor an extensive spatial unmasking database suitable for coding application has been built. A computational model is described in [42], but individual measurement of binaural model is required as an input variable. Hence, the spatial unmasking phenomenon is not further investigated in this thesis.
2.4 Perceptual Coding of Mono/Stereo Audio

In general, the classical lossy audio coders that compress mono/stereo signals are referred as the perceptual audio coders, including the well known MP3 [1] and AAC [27]. In these coders, perceptual theories, such as masking (see Section 2.3.4) and perceptual bands (see Section 2.3.3), are exploited to remove irrelevant and inaudible information, while no or little localisation theories are considered. This section gives brief overviews on the two most widely used perceptual audio coders, MP3 and AAC.

2.4.1 The MPEG-1 Layer III

The wide-spreadly used MP3 audio file format on the Internet in fact refers to the MPEG-1 Layer III standard. Among the three layers in MPEG-1 (Layer I, II and III), the Layer III format offers highest output quality while minimum additional complexity and delay is introduced when compared with the other two layers. Figure 2.8 illustrates the system block diagram of the MPEG-1 Layer III encoder.

In the MP3 encoder, the input signal is filtered by a 32-band PQMF filter resulting in 32 channels of subband coefficients with equal bandwidth. For instances, for a signal at 48kHz sampling rate, the PQMF provides a frequency resolution of 750Hz per channel. In order to further improve the frequency resolution, the 32 channels of subband coefficients are subsequently multiplied by a sine window and then processed by the MDCT transform. During this step, the MDCT transform block length...
is switching between a long block and a short one. The long block utilizes a 18-point MDCT in order to maximize frequency resolution, e.g. this results in a frequency resolution of 41.66Hz for a 48kHz sampled signal. With increased frequency resolution, the temporal resolution is reduced by using such long transform block. This causes temporal noise, typically pre-echo, if the input signal contains transient components, e.g. castanet excerpts [3]. If transients are detected by the psychoacoustical model, which uses FFT for time-frequency transform, the MP3 MDCT transform utilizes a short block of 6-point MDCT to minimize the spreading of temporal quantisation. In order to maintain the frame length, a sequence of three short blocks is used to replace one long block during the signal transient period.

In addition to the transient detection for block switching, the psychoacoustical model provides perceptual analysis for bit allocation. While there are two psychoacoustical models according to MPEG standard [3], both models use FFT in parallel to the subband filtering in the main loop for time-frequency transform. The output from both models are the signal to mask ratios (SMRs) for each subband in the main loop. In general, both models exploits perceptual masking theories (discussed in Section 2.3.4) for the derivation of a masking curve across the whole frequency range, typically by the detection of masker and maskee components. The resulting masking curve is used to derive SMRs for each subband, which is then utilized by the bit-allocation stage to ensure minimum quantisation noise. A detailed description of both models is addressed in [3].

An iterative loop is adopted during the quantisation and entropy coding step. A non-uniform quantizer is applied to the sub-band coefficients as well as a scaling procedure for ensuring higher signal to noise ratio at low level input. Distortion caused by quantisation is computed and the iterative loop alters the quantisation step until the distortion for each band is within requirement. Based on this, Huffman coding is applied to the quantised sub-band coefficients. The bitstream is then multiplexed with the side information such as bit allocation definition and control parameter [3].
which results in the final MP3 data for transmission.

At the time MP3 was introduced in early 1990’s, it provided an efficient solution for the storage and transmission of audio signals, by well compromising between quality, complexity and bandwidth requirement. For example, it was reported to achieve average mean opinion score (MOS) [43] of 3.7 by using 64kbps/channel data rate [44]. Following the explosive development of Internet, MP3 has become a de facto standard of audio formats and been widely used in almost every type of digital media devices, e.g. computer, portable music player, cell phones, etc. However, in recent years, it has been challenged by later standards providing better quality at lower bit-rate, e.g. the MPEG-2 AAC.

2.4.2 The MPEG-2 AAC

While developments in audio research until late 1980’s has been well exploited in MP3, later achievements in audio research can not be further incorporated into the standard for improved performance while maintaining backward compatibility. As a result, the MPEG organization standardized a new non-backward compatible audio coder in 1997, named MPEG-2 Advanced Audio Coding, or AAC [2]. In the AAC standard, three different profiles are defined to offer different output quality, bit-rate and complexity, including Main, Low Complexity (LC) and Scalable Sample Rate (SSR). Figure 2.9 gives an overview of an AAC encoder with a complete set of available tool blocks defined in the standard [27].

Instead of the sub-band hybrid filter bank used in MP3, AAC utilizes MDCT transform only in its main coding loop, while a similar perceptual analysis model as the psychoacoustic model 2 used in MP3 [3] is kept. Transient detection is used to determine whether a long window of 2048 points or a consecutive set of eight 256-point windows is used for the MDCT transform. Effectively, this provides either a high frequency resolution of 23Hz or a high temporal resolution of 2.7ms for a sig-
nal sampled at 48kHz. Two types of analysis/synthesis window can be used in the MDCT transform. A sine window [3] is selected for better passband selectivity or a Kaiser-Bessel Derived (KBD) window [45] is selected for better stopband attenuation. Both windows satisfy the perfect reconstruction and time-domain anti-aliasing criteria for MDCT [18].

A gain control procedure is incorporated in the SSR profile of AAC. A PQMF filter bank is used to split the signal into four subbands with equal bandwidth. As a result, the original signal sampling rate can be efficiently reduced by quarters by discarding one or more subbands, e.g. reduced to 12kHz, 24kHz or 36kHz for a signal sampled at 48kHz [3].

In addition to the block switching procedure, AAC utilizes the temporal-noise shaping technique [46] to remove the pre-echo effect caused by transients. The TNS technique exploits the duality feature between a signal’s time and frequency domain representation and applies linear prediction on the spectral coefficients to achieved efficient coding of transient components when they are detected [3]. In addition, the TNS can be easily adopted into the AAC filterbank by replacing target spectral coefficients by prediction residuals [3].

Figure 2.9 Block Diagram of MPEG-2 AAC Encoder [27]
Prediction algorithms are also embedded in AAC to remove signal redundancy between transform blocks [47] and between audio channels [48]. In addition, Mid/Side stereo coding [49] and intensity stereo coding [50] are applied on each pair of audio channels, which are symmetrically placed on the two sides of the listener. These techniques provide efficient reduction of signal redundancy for both stereo signals and multichannel signals.

Similarly as in MP3 encoding, an iterative loop in AAC derives bit allocation strategy based on the masking threshold derived by the psychoacoustic model, so as to ensure minimum quantisation noise. The resulting quantised coefficients are processed by a Huffman coder. The bit stream is then multiplexed with the side information and control parameters from other blocks, e.g. gain control, TNS, Mid/Side stereo coding and prediction, for storage and transmission.

Based on the subjective evaluations, AAC provides 'indistinguishable' audio quality for a five-channel full-bandwidth signal at a bit-rate of 320kbps [51]. In recent years, AAC becomes a popular option for high quality online music distribution, e.g. the Apple iTunes music store [52], for being able to provide perceptually transparent quality of stereo audio at a bit-rate of 128kbps.

### 2.4.3 High-Efficiency AAC

After it was introduced in 1996 [2], AAC has been further improved by incorporating new technologies. This results in two enhanced versions of High-Efficiency AAC (HE-AAC) in the MPEG-4 Audio standard [53].

The first version of HE-AAC (HE-AAC v1) [7] optimizes the AAC by combining Spectral band replication (SBR) [54] with the AAC LC profile, for applications with constrains on bit rates [55]. In the HE-AAC v1, the classical AAC coder is used for compressing the low frequency components of the input signal, while the SBR technique derives a compact set of side information for reconstructing the high frequency
components during decoding [56]. This effectively results in reduction of transmis-
sion bandwidth by up to 50% [57]. The combined version of AAC and SBR is also
named aacPlus [58], while SBR is also adopted in MP3 resulting in an enhanced
version of MP3 named mp3PRO [59].

The second version of HE-AAC (HE-AAC v2) was standardized by MPEG at 2004
[6]. In addition to the SBR technology in HE-AAC v1, the HE-AAC v2 further
adopts a new technology named Parametric Stereo (PS) [5]. This technique derives
a compact set of cues representing inter-channel difference and correlation between
the two stereo channels. Hence, the transmission bandwidth is reduced from two full
bandwidth channels to one mono channel plus a compact set of side information.
In addition, the idea of exploiting inter-channel audio cues have been extensively
studied resulting in recent rapid development of spatial audio coding technologies.
Hence, the Parametric Stereo technology will be further detailed in Section 2.5.2.

2.5 Parametric Coding of 2D Spatial Audio

Following the significant achievements in perceptual coding of mono/stereo audio
signals (as reviewed in Section 2.4), research interests has been focused on efficient
compression of multichannel spatial audio in recent years. Although the MPEG-2
AAC can provide ‘indistinguishable’ quality for a five-full-bandwidth multichannel
audio signal at a bit-rate of 320kbps (see Section 2.4.2), lower bit-rate for coding
multichannel audio is desired while efficiency and quality is maintained. Typically,
the AAC compression of multichannel audio is achieved by coding each discrete
cchannel individually, e.g. 64kbps per channel resulting in 320kbps for five channels.
Nevertheless, the introduction of Parametric Stereo in HE-AAC v2 (see Section 2.4.3
opened a research area, which aims at efficiently removing cross-channel signal re-
dundancy by utilizing side information, or spatial cues.

Based on this, extended solutions aimed at the representation of multi-channel au-
dio channels using a downmixed signal with fewer channels plus side information have been introduced and proved to be successful[60]. Examples of such algorithms, besides Parametric Stereo, are Binaural Cue Coding (BCC) [15], Directional Audio Coding (DirAC) [16]. Basically, these techniques exploit the cross-channel relationships, including inter-channel level difference (ICLD), inter-channel time difference (ICTD) and inter-channel coherence (ICC), to form a set of spatial cues so that a decoder is able to recover the full channel signal by applying the inter-channel relationships contained in these cues to the downmixed signal. Typically, a five channel audio signal can be downmixed into stereo or mono audio accompanying with spatial cues requiring bandwidth of less than 10kbps [8], while the downmixed stereo or mono audio can be further coded by perceptual coders, such as MP3 or AAC. Thus, with a bit-rate only slightly higher than the normal stereo signal, the decoder can recover the original channel audio signal and provide the audience with a surround sound panorama. Furthermore, backward compatibility can be easily achieved since the stereo/mono downmix has no difference between any other conventional stereo/mono audio formats [61]. These approaches were merged with AAC, resulting in a new MPEG standard for compressing multichannel audio, named ‘MPEG Surround’ [9]. These spatial audio coding approaches will be studied in this section.

2.5.1 Binaural Cue Coding

The system diagram of the Binaural Cue Coding (BCC) is illustrated in Figure 2.10. In this system, a multichannel input signal \((x_1(n), x_2(n) \ldots x_c(n))\) is downmixed into a mono/stereo signal, while the ICTD, ICLD and ICC are derived as spatial cues. The size of the spatial cues is relatively compact, typically less than 10kbps after quantisation and entropy coding [8]. This effectively reduces the bandwidth by approximately 80\%, if a five-channel signal is downmixed to a mono signal. The downmix signal can be further compressed by a perceptual coder, e.g. AAC. This results in using bandwidth slightly higher (for the spatial cues) than conventional mono/stereo audio for transmitting a multichannel signal, while backward compatibility is also
maintained. The spatial cues are used by the BCC decoder for upmixing the downmix signal to a multichannel signal that approximates the original one.

Similar as most modern audio coding techniques, the BCC encoding/decoding process is performed in the frequency domain, which provides benefits of adopting psychoacoustical models and perceptual principles, e.g. critical bands. The time-frequency estimation can be achieved by transform techniques (e.g. STFT) or sub-band filtering (e.g. PQMF). For an input signal with $C$ number of channels, the BCC downmix is calculated in the frequency domain for each time-frequency coefficient as:

$$S(n, k) = e(n, m) \cdot \sum_{c=1}^{C} X_c(n, k)$$  \hspace{1cm} (2.18)

In this equation, time, frequency and perceptual bands are indexed as $n$, $k$ and $m$ respectively. $X_c(n, k)$ and $S(n, k)$ denote coefficients of the input channel signal and downmix signal respectively. $e(n, m)$ is a scale factor for preserving the overall signal energy within each perceptual band calculated as:
where $p_{xc}(n,m)$ and $p_x(n,m)$ are the power estimation for the $c^{th}$ input channel and all input channels respectively, calculated in the $m^{th}$ perceptual band where frequency $k$ is located:

$$
p_{xc}(n,m) = \sum_{k=m_l}^{m_r} |X_c(n,k)|^2
$$

$$
p_x(n,m) = \sum_{k=m_l}^{m_r} \left| \sum_{c=1}^{C} X_c(n,k) \right|^2
$$

where $m_l$ and $m_r$ are the left and right boundary frequency for the $m^{th}$ perceptual band.

In parallel, spatial cues including, ICLD, ICTD and ICC are derived between a reference channel and each of the other channels. For each derivation pair, ICLD, ICTD and ICC are derived for each time index $n$ and each subband index $k$, for a pair of channels $x_1(n,k)$ and $x_2(n,k)$ such that [8]:

- Inter-Channel Time Difference ICTD:

$$
\hat{\tau} = \arg \max_d \{ \Phi_{12}(d) \}
$$

while $\Phi_{12}(d)$ is defined as the normalized cross-correlation function:

$$
\hat{\Phi}_{12}(d) = \lim_{l \to \infty} \frac{\sum_{n=-l}^{l} x_1(n,k)x_2(n+d,k)}{\sqrt{\sum_{n=-l}^{l} x_1^2(n,k) \sum_{n=-l}^{l} x_2^2(n,k)}}
$$
• Inter-Channel Level Difference ICLD in dB:

\[ \Delta \hat{L} = \lim_{l \to \infty} 10 \log_{10} \left( \frac{\sum_{n=-l}^{l} x_2^2(n, k)}{\sum_{n=-l}^{l} x_1^2(n, k)} \right) \]  

(2.23)

• Inter-Channel Coherence ICC:

\[ \hat{\Gamma} = \max_{d} |\hat{\Phi}_{12}(d)| \]  

(2.24)

After these derivation, post-processing and quantisation are performed on these spatial cues [62], which results in average bit-rate of 2kbps for ICLD and ICTD per channel pair and 1.5kbps for ICC [8].

When decoding, the downmix signal is modified based on the spatial cues to generate each of the multichannel signal. For each time-frequency coefficient, ICTD is utilized to determine the delay or phase modification between the reference channel and the generated channel, while ICLD controls the level modification. Correlation between subbands is reduced by applied ICC synthesis so that ICTD and ICLD are effectively as a function of frequency [62].

The performance of BCC compression of stereo signal at a bit rate of 56kbps was reported in [63], where improvement of quality compared with conventional stereo audio coder at a same bit rate is achieved. By combining BCC and MP3, a new multichannel audio coder named MP3 Surround was introduced [64], which operates at a bit rate slightly higher than conventional MP3 for transmitting multichannel surround sound. Subjective evaluations performed on MP3 Surround show that [64] [60], when compared with MPEG-4 AAC operated at 320kbps, the MP3 Surround coder provides minimum quality reduction at a bit rate of 192kbps.
2.5.2 Parametric Stereo

Based on a similar fundamental idea as BCC, the Parametric Stereo (PS) [65] [5] [66] compresses a stereo audio signal into a mono signal by downmixing and extracting inter-channel cues. As mentioned in Section 2.4.3, the deployment of PS in AAC results in the MPEG-4 HE-AAC v2 as it reduces the bandwidth by nearly 50% while minimum distortion in perceptual quality is introduced [6] [67]. Similarly as BCC, the encoding process of PS contains two main procedures, downmixing and spatial cue derivation (as illustrated in Figure 2.11), while the decoder performs upmixing on the downmix signal based on the transmitted cues (as illustrated in Figure 2.12).

In PS, Interchannel Intensity Difference (IID), Interchannel Phase Difference (IPD) and Interchannel Coherence (IC) are derived in the frequency domain as spatial cues.
representing level, delay and correlation information between the two stereo channels. Based on a FFT time-frequency implementation, PS derives one set of cues for each perceptual band by regrouping FFT bins. For the $m^{th}$ perceptual band with frequency boundaries of $[m_l, m_r]$, PS derives IID as:

$$IID(m) = 10 \log_{10} \frac{\sum_{k=m_l}^{m_r} |X_1(k)|^2}{\sum_{k=m_l}^{m_r} |X_2(k)|^2}$$

(2.25)

where $X_1(k)$ and $X_2(k)$ are the FFT coefficients of the two channels. The IPD representing relative phase information between the two channels is derived as:

$$IPD(m) = \angle \left( \sum_{k=m_l}^{m_r} X_1(k)X_2^*(k) \right)$$

(2.26)

where $*$ denotes complex conjugation and $\angle$ represents phase derivation of a complex value. And the IC is defined as the normalized cross-correlation such that:

$$IC(m) = \frac{\left| \sum_{k=m_l}^{m_r} X_1(k)X_2^*(k) \right|}{\sqrt{\left( \sum_{k=m_l}^{m_r} |X_1(k)|^2 \right) \left( \sum_{k=m_l}^{m_r} |X_2(k)|^2 \right)}}$$

(2.27)

The mono downmix signal $S(k)$ is then generated by weighted summation between the two channels:

$$S(k) = \omega_1 X_1(k) + \omega_2 X_2(k)$$

(2.28)

where $\omega_1$ and $\omega_2$ are coefficient dependent magnitude control parameters for preserving constant overall power. Note that $\omega_1$ and $\omega_2$ are also complex-valued to ensure no
out-of-phase cancellation between the two channels. In addition, an extra phase cue is derived in PS to preserve correct phase relationship between the input and output signals, named Overall Phase Difference (OPD), calculated as:

$$OPD(m) = \arg \left( \sum_{k=m_l}^{m_r} X_1(k) S^*(k) \right)$$  \hspace{1cm} (2.29)

The PS decoder rebuilds the stereo signal from the transmitted mono downmix and cues. The absolute delay/phase modification for the decoded channels is derived based on the two phase-related cues, IPD and OPD, while the level modification and de-correlation is derived based on IID and IC respectively. Subjective evaluations show that, by incorporating the PS scheme, the HE-AAC v2 achieves significant improvement on subjective quality when compared with HE-AAC v1, while both coders are at a bit-rate of 24kbps [5].

2.5.3 Directional Audio Coding and Vector Based Amplitude Panning

Directional Audio Coding (DirAC) is a method for efficient compression and soundfield representation of a spatial microphone recording signal, e.g. an Ambisonics B-format signal (further reviewed in Section 2.6). Figure 2.13 gives a system illustration of the DirAC.

In DirAC, a spatial microphone recording signal is analyzed in frequency domain to derive the soundfield information including both localisation (azimuth and elevation) and diffuseness information. This information is transmitted as accompanying side information for a downmix audio channel. For a B-format signal with four channels of input, denoted as $W, X, Y, Z$, DirAC estimates sound source azimuth $\theta$ and elevation $\phi$ by analyzing the intensity relation between the 3D Cartesian coordinate axis. This estimation is carried out based on the signal’s time-frequency representation, such that [69]:

Figure 2.13 Directional Audio Coding [68]

\[
\begin{align*}
\theta(n,k) &= \tan^{-1} \left[ \frac{-I_y(n,k)}{-I_x(n,k)} \right] \\
\phi(n,k) &= \tan^{-1} \left[ \frac{-I_z(n,k)}{\sqrt{I_x^2(n,k) + I_y^2(n,k)}} \right]
\end{align*}
\] (2.30)

where \( n \) and \( k \) are time and frequency indices respectively. The intensity for each axis, e.g. \( I_x(n,k) \) for the x-axis, is calculated as:

\[
I_x(n,k) = \sqrt{2} \frac{\Gamma}{Re \{C_W^*(n,k)C_X(n,k)}
\] (2.31)

where \( Re \{ \cdot \} \) denotes taking real part of a complex value, \( C_W(n,k) \) and \( C_X(n,k) \) are the time-frequency coefficients of the \( W \) and \( X \) input channels, and \( \Gamma \) is the acoustic impedance defined as:

\[
\Gamma = \rho \cdot c
\] (2.32)

where \( \rho \) is the density of medium and \( c \) is the speed of sound [70]. The diffuseness information is calculated in DirAC as:
The omnidirectional channel in the B-format recording, $W$-channel, is transmitted as the downmix signal in DirAC, while direction and diffuseness information derived above is transmitted as side information. When decoding, the directional sound source in the original soundfield is re-rendered by amplitude panning the sound source to the location specified by the DirAC directional cues, while the diffuseness cues are used to reproduce surround image with no perceptual localisation feature.

The inventor of DirAC, V. Pulkki, also proposed a amplitude panning method for rendering localised sound sources using multiple speakers, named Vector Based Amplitude Panning (VBAP) [71]. This panning algorithm can be applied for both 2D and 3D panning. In 2D, to pan a sound source over a pair of loudspeakers located on a horizontal 2D surface, as illustrated in Figure 2.14, a vector $\mathbf{p}$ representing the panned sound source is expressed as:

$$
\psi(n, k) = 1 - \frac{\sqrt{2 \cdot \sum_{i=X,Y,Z} \text{Re} \{C_i^*(n, k)C_i(n, k)\}}^2}{\sum_{i=W,X,Y,Z} |C_i(n, k)|^2}
$$

Figure 2.14 Two Dimensional VBAP
\[ p = g_1 l_1 + g_2 l_2 \] (2.34)

where \( l_1 \) and \( l_2 \) are the unit length vectors pointing the directions of the two loudspeakers and \( g_1 \) and \( g_2 \) are the gain factors. For a pair of loudspeakers symmetrically placed at \( \pm \theta \) (the situation given in Figure 2.14), the VBAP vector analysis in equivalent to tangent amplitude panning law [72], where the relation between loudspeaker directions \( \pm \theta \), loudspeaker gains \( g_1 \) and \( g_2 \) and the resulting source direction \( \varphi \) can be expressed by:

\[ \frac{\tan \varphi}{\tan \theta} = \frac{g_1 - g_2}{g_1 + g_2} \] (2.35)

While not exploited in VBAP, a general derivation will be given in Section 4.4.2, where the fundamental equivalence between 2D vector based panning analysis, 2D Cartesian orthogonal analysis and tangent panning law can be proved, which is further extended for 3D source localisation analysis in Section 6.2. For 3D amplitude
panning, e.g. over three loudspeakers as illustrated in Figure 2.15, the VBAP vector analysis can be expressed as:

\[ \mathbf{p} = g_1 \mathbf{l}_1 + g_2 \mathbf{l}_2 + g_3 \mathbf{l}_3 \]  

(2.36)

where \( \mathbf{p} \) is the vector representing the panned sound source in the 3D soundfield, \( \mathbf{l}_1 \), \( \mathbf{l}_2 \) and \( \mathbf{l}_3 \) are 3D unit length vectors pointing the directions of the three loudspeakers and \( g_1 \), \( g_2 \) and \( g_3 \) are the gain factors.

### 2.5.4 What about the LFE?

The spatial audio coding approaches described in previous sections are focused on efficient compression of the full bandwidth channels in the surround audio signal, e.g. the ‘5’ in the ITU 5.1-channel audio [4]. The ‘.1’ in the ITU 5.1-channel audio refers to a reserved channel containing low frequency components up to 120Hz, named Low Frequency Effect channel or LFE, which is generally used for feeding subwoofers. Compared with a full bandwidth signal, e.g. 44.1kHz sampled signal, the LFE channel requires less than 1% of the full bandwidth. Hence, a simple low-pass filtering can achieve significant bit-rate reduction for compress an LFE channel in PCM format. While the spatial audio coding approaches can be also applied on the LFE channel, little bit-rate reduction is achieved at the cost of additional complexity.

### 2.5.5 MPEG Surround

By further exploiting the fundamental ideas in the spatial audio coding techniques described in this section, the MPEG Surround is a recently standardized compression scheme for multichannel surround audio while compatibility to MPEG-2 AAC is maintained [73]. The encoding/decoding structure of MPEG Surround is similar to other spatial audio coding system, while a multichannel input signal is downmixed into a stereo or mono signal with spatial side information extracted. For instance,
Figure 2.19 illustrates the MPEG Surround Encoding system for stereo downmixing.

Two types of analysis modules are used in MPEG Surround encoding, the two-to-one (TTO) block and the three-to-two (TTT) block. In a TTO block, a pair of input channels is downmixed into a mono channel, while side information is extracted. Specifically, for one perceptual band \( m \), a set of spatial cues are derived between the two input channels \( x_1 \) and \( x_2 \) as [74]:

- **Inter-Channel Level Difference ICLD \( \Delta L_{x_1,x_2,m} \):**

\[
\Delta L_{x_1,x_2,m} = 10\log_{10} \frac{p_{x_1,m}}{p_{x_2,m}}
\]  

where \( p_{x_i,m} \) is the power estimation defined as:

\[
p_{x_i,m} = \sum_{n} \sum_{k=m_l}^{m_r} x_{i,k}(n)x_{i,k}^*(n)
\]  

where \( m_l \) and \( m_r \) are the left and right boundary frequency for the \( m^{th} \) perceptual band, \( k \) is the frequency index, \( n \) is the time index of the analyzed temporal frame.

- **Inter-Channel Correlation ICC \( \rho_{x_1,x_2,m} \):**
\[ \rho_{x_1x_2,m} = \text{Re} \left\{ \frac{\sum_{n} \sum_{k=m+1}^{m_r} x_{1,k}(n)x^*_2,k(n)}{\sqrt{p_{x_1,m}p_{x_2,m}}} \right\} \]  

(2.39)

- Power preservation coefficients \( \psi_1, m \) and \( \psi_2, m \) satisfying:

\[ \psi_1 + \psi_2 = \sqrt{p_{x_1,m} + p_{x_2,m}^2 + 2\rho_{x_1,x_2,m} \sqrt{p_{x_1,m}p_{x_2,m}}} \]  

(2.40)

The one channel downmix signal is then calculated by:

\[ s_{1,k}(n) = \frac{x_{1,k}(n) + x_{2,k}(n)}{\psi_1 + \psi_2} \]  

(2.41)

and an out-of-phase component \( d_{1,k}(n) \) for reconstructing the original two channels are defined such that:

\[ x_{1,k}(n) = \psi_1 \cdot s_{1,k}(n) + d_{1,k}(n) \]

\[ x_{2,k}(n) = \psi_2 \cdot s_{1,k}(n) - d_{1,k}(n) \]  

(2.42)

In a TTT block, three input channels \( x_l, x_r \) and \( x_c \) are downmixed into two channels by a matrixing approach, such that [74]:

\[
\begin{bmatrix}
  s_{l,k}(n) \\
  s_{r,k}(n) \\
  s_{c,k}(n)
\end{bmatrix} =
\begin{bmatrix}
  1 & 0 & 1 \\
  0 & 1 & 1 \\
  1 & 1 & -1
\end{bmatrix}
\begin{bmatrix}
  x_{l,k}(n) \\
  x_{r,k}(n) \\
  \frac{\sqrt{2}}{2} x_{c,k}(n)
\end{bmatrix}
\]  

(2.43)

where \( s_l \) and \( s_r \) are the actual two channel output signals. The third channel, \( s_c \), which is called the auxiliary channel, is discarded but can be recovered by either a
prediction model or a level parameter estimation model. In the prediction model, two prediction coefficients $\gamma_{1,m}$ and $\gamma_{2,m}$ are derived for each perceptual band, such that an estimation of $s_c$ is calculated based on the two downmix channels as:

$$\hat{s}_{c,k}(n) = \gamma_{1,m} \cdot s_{l,k}(n) + \gamma_{2,m} \cdot s_{r,k}(n)$$  \hspace{1cm} (2.44)

The reconstruction of the auxiliary channel can be achieved by either transmitting an extra set of side information containing prediction error of:

$$e_{1,k}(n) = s_{c,k}(n) - \hat{s}_{c,k}(n)$$  \hspace{1cm} (2.45)

or transmitting a correlation parameter:

$$\rho^2_m = 1 - \frac{p_{c_{1,k},m}}{p_{s_{l,m}} + p_{s_{c,m}} + 0.5 \cdot p_{s_{r,m}}}$$  \hspace{1cm} (2.46)

In the level estimation model, two ICLD parameters are estimated to describe the relative energy between the three input channels, such that:

$$\Delta L_{1,m} = 10 \log_{10} \frac{p_{s_{l,m}} + p_{s_{r,m}}}{0.5 \cdot p_{s_{c,m}}}$$

$$\Delta L_{2,m} = 10 \log_{10} \frac{p_{s_{c,m}}}{p_{s_{c,m}}}$$  \hspace{1cm} (2.47)

For an ITU 5.1-channel input signal, the MPEG surround downmixing process utilizing the TTO and TTT blocks is illustrated in Figure 2.17 for stereo output and in Figure 2.18 for mono output.

An example of the MPEG Surround decoder system is given in Figure 2.19, where a stereo downmix is transmitted. During decoding, the MPEG Surround decoder
follows the inverse procedure of the encoding tree structures given in Figure 2.18 or Figure 2.17. Each TTO and TTT block is reversed utilizing the transmitted channels and parameters, resulting in multichannel recovery. For a TTO block, the single channel input of $s_{1,k}$ is decorrelated as $D(s_{1,k})$, using the transmitted ICC cues, then upmixed to two channels by [74]:

$$
\begin{bmatrix}
\hat{x}_{1,k}(n) \\
\hat{x}_{2,k}(n)
\end{bmatrix} = \begin{bmatrix}
\lambda_1 \cos(\alpha + \beta) & \lambda_1 \sin(\alpha + \beta) \\
\lambda_2 \cos(-\alpha + \beta) & \lambda_2 \sin(-\alpha + \beta)
\end{bmatrix} \begin{bmatrix}
s_{1,k}(n) \\
D(s_{1,k}(n))
\end{bmatrix}
$$

(2.48)

The parameters $\alpha$, $\beta$, $\lambda_1$ and $\lambda_2$ in Eq. 2.48 are defined based on the transmitted ICLD, ICTD and ICC cues such that:
For a TTT block, the upmixing from two input channels to three channels is achieved by either using the prediction model or the level estimation model. While using the prediction model, Eq. 2.43 is inversed such that [74]:

\[
\begin{bmatrix}
\hat{x}_{l,k}(n) \\
\hat{x}_{r,k}(n) \\
\hat{x}_{c,k}(n)
\end{bmatrix} = \frac{1}{3}
\begin{bmatrix}
2 & -1 & 1 \\
-1 & 2 & 1 \\
\sqrt{2} & \sqrt{2} & -\sqrt{2}
\end{bmatrix}
\begin{bmatrix}
s_{l,k}(n) \\
s_{r,k}(n) \\
\hat{s}_{c,k}(n)
\end{bmatrix}
\tag{2.53}
\]

where \(\hat{s}_{c,k}(n)\) can be either estimated by the prediction model as:

\[
\hat{s}_{c,k}(n) = \gamma_{1,m} \cdot x_{l,k}(n) + \gamma_{2,m} \cdot x_{r,k}(n) + e_{1,k}(n)
\tag{2.54}
\]

In the level estimation model, upmixing from two channels to three channels is achieved by:
\[
\begin{bmatrix}
\hat{x}_{l,k}(n) \\
\hat{x}_{r,k}(n) \\
\hat{x}_{c,k}(n)
\end{bmatrix} = 
\begin{bmatrix}
w_{11,m} & 0 \\
0 & w_{22,m} \\
w_{31,m} & w_{32,m}
\end{bmatrix}
\begin{bmatrix}
s_{l,k}(n) \\
s_{r,k}(n)
\end{bmatrix}
\tag{2.55}
\]

where \(w_{11,m}, w_{22,m}, w_{31,m}\) and \(w_{32,m}\) are defined as:

\[
w_{11,m} = \sqrt{\frac{\kappa_{1,m} \cdot \kappa_{2,m}}{\kappa_{1,m} \cdot \kappa_{2,m} + \kappa_{2,m} + 1}}
\tag{2.56}
\]

\[
w_{22,m} = \sqrt{\frac{\kappa_{1,m}}{\kappa_{1,m} + \kappa_{2,m} + 1}}
\tag{2.57}
\]

\[
w_{31,m} = \frac{1}{2} \sqrt{\frac{2 \cdot \kappa_{2,m} + 1}{\kappa_{1,m} \cdot \kappa_{2,m} + \kappa_{2,m} + 1}}
\tag{2.58}
\]

\[
w_{32,m} = \frac{1}{2} \sqrt{\frac{2 \cdot \kappa_{2,m} + 1}{\kappa_{1,m} + \kappa_{2,m} + 1}}
\tag{2.59}
\]

with

\[
\kappa_{i,m} = 10^{\Delta L_{i,m}/10}
\tag{2.60}
\]

In order to minimize the overall system complexity, MPEG Surround shares the same QMF filterbank from the AAC encoding/decoder, while the downmixing and spatial cue estimation process can be carried out in AAC’s subband domain [9]. Based on this, subjective evaluations on MPEG Surround show that [73] [75] [76], when combined with HE-AAC, MPEG Surround can achieve average MUSHRA marks [77] of approximately 90 (Excellent quality) for compressing 5.1-channel audio at a bit rate of 150kbps. This is comparable with discrete HE-AAC coding for each channel at the same bit rate. At a lower bit rate of 64kbps, MPEG Surround performs better than discrete HE-AAC coding for each channel, and achieve average MUSHRA [77] (further reviewed in Section 2.8.3) marks of approximately 80 (Good to Excellent quality), comparing with approximately 60 for discrete HE-AAC. When combined
with an MPEG-1 Layer II coder, MPEG Surround can achieve average MUSHRA marks of approximately 90 at a bit rate of 256kbps.

The development of the spatial audio coding techniques and standardization of MPEG Surround provides an efficient solution for extending the application of multichannel surround audio in the consumers market while backward compatibility is maintained. By utilizing these approaches, multichannel audio can be stored and transmitted without introducing additional bandwidth and complexity significantly, although perceptual distortion is minimized when comparing the decoded multichannel signal with the original signal.

### 2.5.6 Binaural Reproduction of Spatial Audio

In addition to efficient compressing and recovery of multichannel spatial audio signals, the spatial audio coding techniques adopt binaural rendering principles for virtual surround sound synthesis over binaural renderings [78] [69]. In these approaches, while no additional modification is required for the encoding, the transmitted spatial cues are derived into binaural source localisation information so that a pair of HRTFs with matching location is used to filter the transmitted downmix signal [79]. This results in stereo output signals with virtual surround scene similar as the original multichannel surround scene for binaural playback. Subjective evaluations of MPEG Surround binaural reproduction coder have been carried out [80] [78] [73], where direct HRTF filtered version of the original multichannel signal is used as the reference signal. It is shown that, when compared to the reference signal, the MPEG Surround binaural reproduced signal can achieve ‘Good’ quality at a bit rate of 160kbps and ‘Excellent’ at 191kbps. Based on this, the binaural reproduction of a spatial audio coded signal can be used as a backward compatible solution to provide virtual surround sound rendering for users with conventional stereo system or portable music players. However, while the spatial cues in these spatial audio coding approaches do not directly represents sound source localisation information, addi-
tional complexity is required to transform spatial cues to localisation information so that the HRTF pair with correct localisation can be derived.

2.5.7 The ‘Downmix + Cues’ Framework

It is obvious that, the spatial audio coding approaches described in this section share a similar fundamental idea: downmixing a multichannel spatial audio signal into mono/stereo for achieving significant bandwidth reduction, while extracting cross-channel arithmetical relationships as spatial cues for multichannel recovery. This approach can by viewed as a ‘downmix + cues’ framework, which has advantages in the following aspects:

- Bandwidth is reduced from $N$ channels (e.g. five-channel) to mono or stereo.
- The mono/stereo downmix can be further compressed by MP3/AC.
- The mono/stereo downmix is fully backward compatible to conventional non-multichannel audio systems.
- Spatial cues can be efficiently compressed and quantised as a compact set of side information requiring minimum bandwidth.
- Compatible with existing perceptual coders such as AAC, while spatial cue derivation can be performed in AAC’s subband domain.

In comparison with conventional approaches, e.g. using AAC to compress each channel individually, this ‘downmix + cues’ framework provides near transparent quality at a bit-rate comparable to mono/stereo signal [73], [76]. However, the following drawbacks are found in this framework:

- Correct multichannel signal recovery critically relies on the spatial cue side information, without which the sound localisation property in the original signal cannot be retrieved.
• The spatial cues do not directly represent sound source localisation information, rather, they are cross-channel mathematical relationships.

• Following the previous point, additional complexity is introduced for translating cues to surround localisation information, if perceptual localisation theories should be applied for improving the efficiency of spatial cues quantisation.

• As the downmixing process and cue derivation are dependent to each other, and are carried out in pre-defined channel pair, a specific algorithm has to be designed for each type of multichannel format. This results in limited flexibility for deploying this framework over different multichannel audio formats.

• As the spatial cues are derived for each pair of channels, computational cost will increase significantly if the number of input channels increases.

In order to overcome the limitation of using extra side information, in addition to the two standard modes using mono and stereo downmix with accompanying side information, MPEG Surround also provides a ‘blind upmixing’ (also referred as ‘Non-Guided’) decoding approach, where the downmix signal is decoded to multichannel without using the side information [9]. However, while using the ‘Non-Guided’ mode, the surround soundfield is significantly distorted after MPEG Surround decoding, due to the lack of critical soundfield information transformed and saved in the spatial cues.

In comparison with existing spatial audio coding approaches based on the ‘downmix+cues’ framework, the S$^3$AC technique proposed in this thesis addresses above shortcomings by analyzing the spatial soundfield to efficiently represent source localisation information rendered by multichannel signal and inherently avoids extra metadata. This will be further addressed in following chapters.
2.6 Ambisonics

Ambisonics [14] [17], introduced in the 1970’s, is known as one of the best spatial audio recording techniques and provides excellent soundfield and source location recoverability. The application of Ambisonics involves spatial sound recording with a soundfield microphone and reproduction technique to reproduce surround soundfield over a number of loudspeakers [81]. Ambisonics has been widely used in professional studios for surround soundfield recording and reproduction. This section reviews several typical Ambisonics principles and applications, including first order Ambisonics, B-format, UHJ, higher order Ambisonics and Ambisonics reproduction over loudspeakers.

2.6.1 First Order Ambisonics and B-Format

The fundamental principle of Ambisonics is that, information of a 3D surround soundfield can be recorded and stored as four channels, named the W-X-Y-Z channels. The W channel contains omnidirectional sound pressure information, while the remaining three channels, X, Y and Z, contain directional sound velocity information over the three according axes in a 3D Cartesian coordinates. This is known as the first order Ambisonics and the W-X-Y-Z channels are known as the Ambisonics B-format [82]. The soundfield microphone is designed to record an Ambisonics B-format, in which an omnidirectional microphone is used to record the W channel and three orthogonally placed figure-of-eight microphones are used to record the X-Y-Z channels. Figure 2.20 illustrate an example of a soundfield microphone. Ideally, a B-format representation of a three-dimensionally localised sound source can be expressed by:
Figure 2.20 Soundfield Microphone

\[ W = \frac{\sqrt{2}}{2} S \]
\[ X = \cos \mu \cdot \cos \eta \cdot S \]
\[ Y = \sin \mu \cdot \cos \eta \cdot S \]
\[ Z = \sin \eta \cdot S \]

(2.61)

where \( \mu \) and \( \eta \) are the azimuth and elevation of the sound source location, \( S \) represents the localised sound source. This is also illustrated in Figure 2.21. A perfect B-format recording cannot be achieved due to the size and non-ideal hardware property of each microphone, as well as the encapsulation of four individual microphones. However, an artificial B-format signal can be synthesized by up-mixing a monaural signal to B-format using user specified sound source localisation.

As the Ambisonics B-format signal represents the surround soundfield and sound
source localisation information, the recording or synthesizing of B-format Ambisonics signal is independent of the playback loudspeaker configuration. This introduces high flexibility for Ambisonics reproduction system, which the producer of the original surround soundfield has no need to concern. However, a geometrically symmetrical loudspeaker layout, e.g. cubic or dodecahedron, is required for optimal Ambisonics reproduction [83].

2.6.2 UHJ

The backward compatibility to conventional stereo for Ambisonics can be achieved by a UHJ downmixing [84]. This is achieved by discarding the Z channel and downmixing the remaining three channels by:

\[
L = (0.4699 - 0.171j)W + (0.0928 + 0.255j)X + 0.3277Y \\
R = (0.4699 + 0.171j)W + (0.0928 - 0.255j)X - 0.3277Y \quad (2.62)
\]
where \( j \) is the imaginary unit. The decoding from UHJ to three-channel Ambisonics B-format is calculated as:

\[
\begin{align*}
W &= 0.491(L + R) + 0.082j(L - R) \\
X &= 0.209(L + R) - 0.414j(L - R) \\
Y &= 0.381(L - R) + 0.192j(L + R)
\end{align*}
\] (2.63)

Although stereo backward compatibility is achieve by using the UHJ approach, a large amount of soundfield localisation content is discarded due to the fix-parameter downmixing. Thus, UHJ is considered as a low-quality solution for achieving Ambisonics backward compatibility as significant distortion is introduced after decoded to Ambisonics for reproduction.

### 2.6.3 Higher Order Ambisonics

In addition to the first order B-format, Ambisonics can achieve improved source localisation accuracy and sound image stability by improving its order. Higher order Ambisonics (HoA) [85] signal is achieved by increasing the number of channels in addition to the W-X-Y-Z in the first order, as shown in Appendix A. The number of channels required for a \( m^{th} \) order Ambisonics is \((m + 1)^2\), which effectively means great challenge in designing and manufacturing HoA microphone. Although research shows possibility of producing HoA microphone [86], efficient and reliable production of HoA signal can be achieved by artificial synthesis. To produce an \( m^{th} \) order Ambisonics signal \( B(t) \) for a source signal located at azimuth \( \mu \) and elevation \( \eta \), the source signal \( s(t) \) is multiplied by a vector \( y(\mu, \eta) \) containing a series of spherical harmonics functions [83]:

\[
B(t) = s(t) \cdot y(\mu, \eta)
\] (2.64)
while $y(\mu, \eta)$ is defined as:

$$ y(\mu, \eta) = \left[ Y_1(\mu, \eta), Y_2(\mu, \eta), \ldots, Y_{(m+1)^2}(\mu, \eta) \right] $$

(2.65)

while the spherical harmonics function $Y_1(\mu, \eta)$ is defined in Appendix A. Theoretically, Ambisonics requires infinite order (i.e. infinite number of spherical harmonics functions) to describe a soundfield information perfectly [85]. However, this is limited by available signal bandwidth and playback loudspeaker configuration. The number of loudspeaker channels required for reproducing an $m^{th}$ order Ambisonics signal is $m^2$. Thus, practical Ambisonics approach limits the utilized order based on available bandwidth and playback system, based on which efficient reproduction of source localisation and soundfield image can be achieved [83].

2.6.4 Ambisonics Reproduction over Loudspeakers

The Ambisonics encoded signal has to be decoded for loudspeaker playback. The decoding process is dependent on the number of available loudspeakers and the configuration of the loudspeaker array. For an $m^{th}$ order Ambisonics encoded signal, a total number of $m^2$ loudspeakers are required for reproducing the soundfield efficiently. However, the order of an encoded Ambisonics signal can be easily reduced by discarding higher order channels if reproduction loudspeaker system does not have enough channels. To decode an Ambisonics signal $B(t)$, a decoding matrix $D$ is applied so that the loudspeaker signal $L(t)$ is generated by [83]:

$$ L(t) = D \cdot B(t) $$

(2.66)

The decoding matrix $D$ can be derived as the pseudo-inverse of a ‘re-encoding’ matrix $C$, such that:
\[ D = \text{pinv}(C) = C^T \cdot (C \cdot C^T)^{-1} \] (2.67)

where \( \cdot^T \) denotes transposition operation. The re-encoding matrix \( C \) is defined as:

\[ C = [c_1, c_2, \ldots, c_i, \ldots, c_N] \] (2.68)

where \( i \) is the speaker index and \( N \) the total number of loudspeakers in the reproduction system. The vector \( c_i \) is defined as the series of spherical harmonics functions \( Y_n(\mu, \eta) \) in Appendix A.1, where the azimuth \( \mu_i \) and elevation \( \eta_i \) of the \( i^{th} \) loudspeaker is used:

\[ c_i = [Y_1(\mu_i, \eta_i), Y_2(\mu_i, \eta_i), \ldots, Y_{(m+1)^2}(\mu_i, \eta_i)] \] (2.69)

In addition, it is required that a regular loudspeaker array is utilized for Ambisonics decoding, so that the pseudo-inverse operation is applicable in Eq.2.67. This is equivalent to ensuring the re-encoding matrix \( C \) to satisfy:

\[ \frac{1}{N} \cdot C \cdot C^T = I \] (2.70)

where \( I \) is a diagonal unitary matrix [83].

In general, Ambisonics can efficiently capture a soundfield, as well as its localisation information, by using a first-order B-format signal. The source localisation accuracy and soundfield stability can be improved by synthesizing higher order Ambisonics for reproduction. It also provides an efficient solution for artificial rendering of a three-dimensionally localised sound source based on a monophonic signal and user specified locations. Ambisonics is also flexible in terms of reproduction. The order
of encoded Ambisonics signal can be easily enhanced or reduced to match the available loudspeaker configuration. As long as the loudspeaker array is regular, efficient decoding and reproduction can be achieved to recover the original surround sound scene.

In this thesis, the proposed S\(^3\)AC techniques is used as an efficient, stereo backward compatible, compression approach for Ambisonics B-format recordings, while compared with the UHJ approach (Section 5.2). In addition, the 3D Ambisonics reproduction approach is exploited in Chapter 6 as a reproduction method for S\(^3\)AC encoded 3D multichannel audio signals.

### 2.7 3D Spatial Audio

Two-dimensional multichannel audio formats, such as ITU-5.1 [4], was introduced decades ago as a significant improvement over traditional stereo system for scenarios where better sound localisation feature than stereo system is desired, e.g. cinema, home theater. In recent years, research interests in 3D spatial audio are developing and some multi-channel 3D audio systems are being deployed, for instance, in commercial applications such as IMAX [87], academic research such as Allosphere [88]. However, due to the natural complication of 3D audio and increased bandwidth cost introduced by more numbers of loudspeakers as well as hardware and software, the application of 3D audio is still limited. While significant research interests have focused on efficient representation of 2D spatial audio, especially coding 5.1-channel audio, as reviewed in Section 2.5, less interest can be found in the area of compressing 3D spatial audio. In addition, no international standard for 3D loudspeaker format is recognized, neither for 3D audio recording, encoding or reproduction.

In the scenarios of 3D audio, the number of audio channels can go significantly higher than those found in 2D audio, e.g. 16-channel in 3D compared to 5.1-channel in 2D, resulting in demanding bandwidth even with compressing each channel with
MP3/AAC. Current achievements in 2D horizontal only spatial audio based on parametric analysis, such as MPEG Surround (see Section 2.5), cannot be efficiently extended for compressing 3D spatial audio. This is because the number of parameters representing arithmetical relationships between channels increases in linearly with the number of loudspeakers, which are significantly higher in 3D. Another recently introduced 3D audio technique called Vector Based Amplitude Panning (VBAP) [89] (see Section 2.5.3) focuses on re-producing localised audio contents in 3D loudspeaker arrays. However, when encoding a multi-channel 3D audio signal, reversing VBAP to analyze a number of channels is not applicable as the VBAP algorithm is based on analyzing a set of three triangularly placed loudspeakers. The Ambisonics [17] (see Section 2.6) approach introduced decades ago has been used for many 3D audio recording and reproduction works, in both scientific researches and creative arts. However, practical application of Ambisonics is limited by requiring symmetrical loudspeaker array for accurate Ambisonics reproduction, which is not applicable in real-world scenarios such as cinemas. On the other hand, in practical Ambisonics, where sound source and localisation is represented by a 4-channel microphone recording called B-format (or 1st order Ambisonics) [81], the source localisation derived from B-format recording is inconsistent as ideal characteristics of the recording hardware is not practical. In addition, while higher order Ambisonics provides significant improvement on the localisation sharpness for reproduction, its recording techniques are complex and offers significant practical challenges.

In this thesis, an $S^3$-AC based approach is proposed for efficient representation of 3D multichannel audio. Unlike Ambisonics, which is fundamentally a recording technique, the proposed approach is focused on encoding a 3D multichannel audio signal, i.e. the signal for loudspeaker playback, to significantly reduce the bandwidth requirement for 3D audio while distortion caused by recording hardware can be avoided. This approach will be further addressed in Chapter 6.
2.8 Evaluation of Audio Quality

For a lossless audio coder, the performance of the coder is evaluated by complexity and the compression ratio achieved, as only the signal redundancy is removed and the original signal can be perfectly reconstructed. For lossy audio coders, while perceptual irrelevancy is removed from the signal, it is importance to evaluate that the discarded signal contents are indeed with no or less perceptual importance. This evaluation can only be achieve by subjective comparison between the original signal and the coded signal, i.e. doing listening tests. In addition, subjective evaluation can utilize statistical analysis so that the perceptual degradation introduced by an audio coder can be quantified. In this section, several widely used subjective evaluation methodologies are reviewed.

2.8.1 Mean Opinion Score

The Mean Opinion Score (MOS) is a typical subjective evaluation methodology to rate the quality of one audio or speech coder [90] [43]. In a MOS evaluation, a group of listeners are asked to compare the original signal condition and the coded condition following the marking scheme given in Table 2.1. A number of test files should be used in the MOS test and the average of all individual scores are calculated as the resulting MOS mark for the candidate audio codec.

2.8.2 Perceptual Evaluation of Audio Quality

The Perceptual Evaluation of Audio Quality is an ITU standard for computational audio quality assessment, designed to overcome the time-consuming drawback of listening test [91]. Although the PEAQ is an objective method, the evaluation algorithm used in PEAQ is based on perceptual analysis on the evaluated signal. A PEAQ evaluation program compares a coded signal with a reference signal resulting in a quality mark similar as the MOS mark specified in Table 2.1. In the PEAQ eval-
<table>
<thead>
<tr>
<th>MOS</th>
<th>Quality</th>
<th>Impairment</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Excellent</td>
<td>Imperceptible</td>
</tr>
<tr>
<td>4</td>
<td>Good</td>
<td>Perceptible but not annoying</td>
</tr>
<tr>
<td>3</td>
<td>Fair</td>
<td>Slightly Annoying</td>
</tr>
<tr>
<td>2</td>
<td>Poor</td>
<td>Annoying</td>
</tr>
<tr>
<td>1</td>
<td>Bad</td>
<td>Very Annoying</td>
</tr>
</tbody>
</table>

Table 2.1 Mean Opinion Score

ulation, the two input signal (the reference and the candidate) are both transformed into frequency domain, where a computational model of human ear perception [92] is utilized to derive perceptual quality variables. The variables are then derived into MOS marks for the candidate coding condition [91]. While PEAQ is an efficient methodology for computationally evaluate perceptual quality, it can be only applied for mono/stereo signal. Hence, it is not used in this thesis.

2.8.3 MUSHRA

MUSHRA [77] stands for Multiple Stimuli with Hidden Reference and Anchor. Compared with the MOS methodology, the MUSHRA methodology utilizes statistical analysis such as confidence intervals to achieve statistical significant results with fewer listeners. According to the MUSHRA recommendation, listener is guided to listen to a labeled reference signal followed by a group of un-labeled signals, including a hidden reference, a 3.5kHz low-pass filtered anchor signal and one or more coded conditions. The listener rates the un-labeled signals based on the perceived quality comparing the labeled reference. A general quality guideline based on a 0-100 marking scale is recommended in MUSHRA, as shown in Table 2.2, based on which the listener can distinguish obvious differences and small differences. For each candidate coding condition, the average score is calculated by taking the arith-
<table>
<thead>
<tr>
<th>Quality</th>
<th>MUSHRA Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>[100 80]</td>
</tr>
<tr>
<td>Good</td>
<td>[80 60]</td>
</tr>
<tr>
<td>Fair</td>
<td>[60 40]</td>
</tr>
<tr>
<td>Poor</td>
<td>[40 20]</td>
</tr>
<tr>
<td>Bad</td>
<td>[20 0]</td>
</tr>
</tbody>
</table>

Table 2.2 MUSHRA Scoring Recommendation

metrical mean between all listeners, and a 95% confidence interval is also calculated based on the standard deviation and the number of listeners.

Being a comprehensive subjective evaluation methodology as well as taking statistical significance into evaluation consideration, MUSHRA has been widely used for evaluating perceptual audio coders, including spatial audio coders, [8] [65] [10] [75]. The subjective evaluation methodology used in this thesis is also based on MUSHRA. As MUSHRA is designed based on the evaluation of mono/stereo coders, additional considerations are adopted into MUSHRA for some listening test projects in this thesis, in order to meet the requirement of perceptual evaluation of localised sound sources and surround audio. These will be detailed in the following chapters.

2.9 Summary

This chapter has been focused on the fundamentals of digital audio processing and the state-of-the-art spatial audio coding approaches. The digital representation of audio signals and audio signal formats from mono to multichannel has been reviewed, followed by investigation on time-frequency representation of audio signals. Some important psychoacoustical principles have also been reviewed. These principles have been extensively exploited in the conventional perceptual audio coders, as well
as the spatial audio coders, to achieve efficient compression of audio signals at low bit rates. Well-known conventional perceptual audio coders including MPEG-1 Layer III and MPEG-2 AAC are also reviewed, followed by detailed investigation on recent developments in spatial audio coding techniques. Approaches such as BCC, PS, DirAC and MPEG Surround, are described. These approaches are based on a ‘downmix + cues’ framework, where multichannel spatial audio are downmixed to less channels (typically mono or stereo) while side information containing inter-channel arithmetical relationships is extracted for surround sound recovery. The advantages and limitations of this ‘downmix + cues’ framework is also discussed. The Ambisonics soundfield recording and reproduction technique is reviewed for efficient representation for both 2D and 3D soundfield. The encoded Ambisonics signal can be either a soundfield microphone recording or a synthesized signal, while Ambisonics provides high flexibility in terms of reproduction configuration. This chapter finally reviews some methodologies for the evaluation of audio quality, which plays an importance role in the assessment of the performance and efficiency of an audio coder.

In the next chapter, the S3AC fundamental principles will be presented. This is followed by a detailed description of a typical S3AC encoding/decoding system for compressing ITU 5.1-channel audio signals, while both objective and subjective evaluation are presented.
Chapter 3

Spatially Squeezed Surround Audio Coding: Fundamental Principles and Applications

3.1 Introduction

Recent developments in efficient and backward compatible coding of spatial audio has been reviewed in Section 2.5. These approaches, based on a ‘downmix + cues’ framework, have been reported to have satisfactory performance in terms of significant bit-rate reduction and near-transparent perceptual audio quality, while maintaining backward compatibility to conventional stereo/mono audio systems [61] [9] [75]. Limitations in these approaches have been discussed in Section 2.5.7, while a major one is that additional side information (either incorporated into the binary format or as a separate transmission) is essential for the precise recovery of the spatialized auditory events and the full surround sound scene. This results in increased computational complexity for cue synthesis, as well as increasing storage and transmission requirements.

In this thesis, a new approach to coding multi-channel spatial audio signals is developed and presented, called Spatial Squeezing Surround Audio Coding ($S^3$AC), which fundamentally requires no additional side information to transmit surround
sound localisation information for recovering the full channel signal from a stereo downmix. Instead of exploiting the arithmetic relationships between audio channels found in the approaches based on the ‘downmix + cues’ framework, S\(^3\)AC achieves the compression of a spatial audio signal by representing it with a smaller sound field, typically, a stereo sound field, which is exploited to represent the perceptual localisation information in the original surround soundfield. Side information is avoided as the surround localisation can be directly derived from the transmitted stereo signal for reproducing the multichannel surround sound scene. In addition, the stereo downmix signal in S\(^3\)AC is also backward compatible to a conventional stereo audio system.

In this Chapter, the fundamental principles of S\(^3\)AC are presented, followed by a detailed description of an S\(^3\)AC encoding/decoding system for efficient and backward compatible compression of ITU 5.1-channel audio signals. Based on this, both subjective and objective evaluation results are presented for evaluating the proposed S\(^3\)AC system.

### 3.2 Fundamental Principles of S\(^3\)AC

#### 3.2.1 Squeezing the Auditory Space

Ideally, a typical 5-channel surround audio track can provide an audience, standing at or near the center ‘sweet-point’ of the imaginary circle, with a full 360° 2D surround sound scene. However, psychoacoustics show that human hearing has a limited resolution in locating sound objects [36]. This phenomenon is known as localization blur as discussed in Section 2.3.6. In addition, psychoacoustics also show that although different features may result from altering the frequency properties of the perceived sound, the highest perceptual localisation resolution is approximately 1° [36], which is found for the front-center area for a sound source with frequency component of approximately 1kHz. In other words, displacement of an auditory event of up to 1°
Spatially Squeezed Surround Audio Coding: Fundamental Principles and Applications

is perceptually unnoticeable. In addition, the perceptual localisation resolution degrades, when the sound source moves to the ‘less-sensitive’ area, e.g. sides, rear or top. Based on this, to represent surround localisation information of a 360° soundfield rendered by a 5.1-channel audio track without introducing perceptual distortion, it is required that 360 ‘points’ of localisation information can be efficiently derived and represented.

However, when mixing music to stereo or multi-channel speakers using a computational approached, e.g. amplitude panning laws, the precision of a panned source direction can be much higher than 1°; since it is based on numerical calculation, the precision of which is limited by the algorithm’s numerical precision. Inversely, the precision of the source localisation information that can be derived from a soundfield is limited by computational precision, which is much higher than perceptual localisation precision.

Based on these, the fundamental concept of S³AC is that a full surrounding auditory panorama for perceptual listening purposes can be represented by a smaller ‘squeezed’ soundfield for the purposes of transmission. The ‘squeezed’ soundfield contains sufficient sound localisation information, which can be derived by a computational algorithm, for full surround sound recovery without introducing perceptual loss. Thus, the data rate of the compressed signal has minimum independency of the number of original channels but is instead dependent on the size of the squeezed soundfield. A typical application of S³AC is the compression of ITU 5.1-channel audio. As illustrated in Figure 3.1, a 360° surround soundfield rendered by a 5.1-channel signal is squeezed into a 60° stereophonic soundfield, which contains sufficient perceptual localisation information for efficient full surround sound recovery.
3.2.2 Sound Source Localisation Estimation

The efficiency of the source localisation estimation algorithm has critical impact on the proposed S\textsuperscript{3}AC scheme, as S\textsuperscript{3}AC fundamentally requires accurate sound source localisation estimation for its encoding/decoding process. To render a sound source over loudspeakers at a specified position, an efficient method is to utilize pan-pot laws [93]. For example, to reproduce a source at azimuth $\varphi$ by a stereo pair of loudspeakers located at $\pm \theta$, as illustrated in Figure 3.2, it can be effectively achieved by utilizing the tangent amplitude panning law, such that:

$$\frac{g_L - g_R}{g_L + g_R} = \frac{\tan \varphi}{\tan \theta} \quad (3.1)$$

where $g_L$ and $g_R$ are the gain parameter of the left and right channel respectively. For a given source level, e.g. $g_S$, it should be also satisfied that:

$$g_L^2 + g_R^2 = g_S^2 \quad (3.2)$$
or equivalently:

\[
\begin{align*}
g_L &= g_S \cdot \frac{\tan \theta + \tan \varphi}{\sqrt{2 \cdot \tan^2 \theta + 2 \cdot \tan^2 \varphi}} \\
g_R &= g_S \cdot \frac{\tan \theta - \tan \varphi}{\sqrt{2 \cdot \tan^2 \theta + 2 \cdot \tan^2 \varphi}}
\end{align*}
\] (3.3)

Based on this, the source localisation rendered by a given pair of loudspeakers can be effectively estimated by:

\[
\varphi = \tan^{-1} \left[ \frac{g_L - g_R}{g_L + g_R \tan \theta} \right] 
\] (3.4)

Note that, for a pair of loudspeakers located on a 360° circle, the angles can be limited within ±90° by rotating the axis. The \(\tan^{-1}(\cdot)\) function in Eq. 3.4 is monotonic, so that the derived source azimuth \(\varphi\) in Eq. 3.4 has a single value.
While the inverse amplitude panning method described above provides efficient estimation of source localisation rendered by a pair of loudspeakers at low computational cost, other localisation estimation algorithms can be adopted in S\textsuperscript{3}AC. A further analysis of localisation estimation algorithms is given in Section 4.4.

3.3 S\textsuperscript{3}AC Applied to ITU 5.1-Channel Surround Audio

The application of S\textsuperscript{3}AC encoding/decoding system to ITU 5.1-channel surround audio is described in this section. This system aims at efficient and backward compatible compression of a 5.1-channel audio signal by exploiting the fundamental spatial squeezing principles of S\textsuperscript{3}AC. The block diagram of the encoding system is illustrated in Figure 3.3, while the decoding system is illustrated in Figure 3.4. Note that, in this system, the ‘.1’ channel is not considered, as significant bandwidth reduction of the LFE channel can be achieved by a simple low-pass filtering (as discussed in Section 2.5.4).

3.3.1 Time-Frequency Transform

The analysis and synthesis process of S\textsuperscript{3}AC is performed in the frequency domain. For a multichannel PCM waveform input signal, the time-frequency transform can be achieved by using any modern transform or subband filtering technique, such as those described in Section 2.2.3. In this implementation, the STFT described in Section 2.2.3 is utilized based on a 1024-point 50% overlapping implementation. The STFT is applied on each of the input channels, such that, for the \( i \)\textsuperscript{th} channel, the STFT coefficients of the \( p \)\textsuperscript{th} frame is achieved by:

\[
X_{i,p}(k) = \sum_{n=M(p-1)}^{M(p+1)-1} x_{i,p}(n) \cdot w(n) \cdot e^{-j2\pi kn/2M}
\]  

(3.5)
Figure 3.3 $S^3AC$ 5-Channel Encoding System
Figure 3.4 $S^3$AC 5-Channel Decoding System
where $k$ is the frequency index, $M = 512$, and $w(n)$ is the analysis window with a length of 1024 points. In this implementation, the Kaiser-Bessel derived (KBD) [3] [45] window with shape parameter $\alpha = 4$ (for a good trade-off between main-lobe selectivity and side lobe attenuation, as discussed in [3]) is applied as both the analysis and synthesis window, which is plotted in Figure 3.5.

The inverse STFT transform can be achieved by:

$$y_{i,p}(n) = \frac{1}{2M} \sum_{k=0}^{2M-1} X_{i,p}(k)e^{j2\pi kn/2M}$$  \hspace{1cm} (3.6)$$

And the original time-domain signal $x_{i,p}(n)$ is recovered by multiplying $y_{i,p}(n)$ with synthesis window followed by the overlap-add process. In the following analysis of this section, the frame index $p$ is ignored for simplicity purpose.
3.3.2 S\textsuperscript{3}AC Encoding

In the S\textsuperscript{3}AC encoding, the STFT coefficients of all input channels are analyzed by three steps: virtual source localisation estimation in surround soundfield, 360° to 60° mapping and re-panning to downmix, as illustrated in Figure 3.3. These three S\textsuperscript{3}AC analysis steps are detailed in the following.

**Virtual source localisation estimation in surround soundfield**

In this step, the ‘rendered’ surround sound scene and virtual sources are evaluated. The 360° sound field can be divided into a number of sub-regions based on the speaker setup, as illustrated in Figure 3.6, such that each sub-region is represented by amplitude panning between a pair of channels. Adjacent channels are used to render virtual sources located near the perimeter while diagonal channel pairs can be used to render some virtual sound localisation effects within the central soundfield area, e.g. an aircraft flying overhead effect. When S\textsuperscript{3}AC is applied to existing recordings, pairs of frequency coefficients from adjacent or diagonal channel pairs are assumed to represent a virtual source.

The amplitudes $|X_a[k]|$ and $|X_b[k]|$ of the virtual source of frequency $k$ rendered by two channels $a$ and $b$ are selected as:

$$\{a, b\} = \arg \max_{ij} [ |X_i(k)| + |X_j(k)| ]$$  \hspace{1cm} (3.7)

where $|X_i(k)|$ and $|X_j(k)|$ are the magnitudes of channel pair $i, j$ as a function of frequency $k$. This method identifies a virtual source as that rendered by the channel pair with the dominant energy for a given frequency component. The corresponding azimuth of the virtual source $\varphi_{ab}$ is computed based on Eq. 3.4 as:

$$\varphi_{ab}(k) = \tan^{-1} \left[ \frac{|X_a(k)| - |X_b(k)|}{|X_a(k)| + |X_b(k)| \tan \theta_{ab}(k)} \right]$$  \hspace{1cm} (3.8)
where $\theta_{ab}(k)$ is half the angular separation of the chosen channel pair $a$ and $b$ of Figure 3.6, e.g. $\theta_{ab}(k) = 15^\circ$ if the front-left channel and the center channel are selected. As the selection of channel pair is based on analyzing the frequency coefficients of each channel pair, as given in Eq. 3.7, $\theta_{ab}$ is frequency dependent. Note that, the derived azimuth $\phi_{ab}(k)$ is the azimuth in the coordinates where the selected $a$ and $b$ channels are symmetrically placed. It should be then transformed into the main coordinates (where the center channel is located at $0^\circ$) by adding an axis rotation parameter:

$$\phi_{360}(k) = \phi_{ab}(k) + \varpi_{ab}$$  \hspace{1cm} (3.9)

The axis rotation parameter $\varpi_{ab}$ is dependent on the selected channel pair. For example, if the front-left channel (located at 30°) and the surround-left channel (located at 110°) are selected, it can be derived that $\varpi_{ab} = 70^\circ$.

Based on this, a virtual source is thus represented by a single mono signal and its azimuth $\phi_{360}(k)$, while the mono signal can be derived based on Eq. 3.2, such that:

$$S(k) = \sqrt{X_a^2(k) + X_b^2(k)} \cdot e^{i\phi_{ab}}$$  \hspace{1cm} (3.10)

where the phase $\phi_{ab}$ is chosen as the phase of channel with the higher amplitude out of channel $a$ and $b$, i.e. the dominating channel. For example, in Figure 3.6, virtual source 1 (VS1) located at 15° in the front-left region (FL) is derived from the front-left channel (located at 30°) and the center channel (located at 0°) in the 5-channel setup. Virtual source 2 (VS2) located at -150° in the rear region is derived from the two surround channels (located at $\pm 110^\circ$). This process assumes sources are rendered in a similar way to the amplitude panning approach, as addressed in Section 3.2.2, however here panning is applied to individual of frequency coefficients. The identified virtual source is then squeezed into the smaller downmix auditory space.
Once frequency-domain virtual sources are derived for each channel pair, a new azimuth in the stereo sound field is assigned to this source according to a 360° to 60° linear mapping criteria such that:

$$\varphi_{dm}(k) = f(\varphi_{360}(k))$$  \hspace{1cm} (3.11)

where \( f(\cdot) \) is a function that represents the azimuth mapping process. An example of a squeezed soundfield from Figure 3.6 is given in Figure 3.7 also detailed in Table 3.1, where the VS1 and VS2 from Figure 3.6 are mapped to 25° and 2.5° respectively in Figure 3.7. Note that, although the 360° to 60° azimuth mapping criteria can be chosen flexibly, this mapping criteria has to be monotonic, i.e. a one-to-one mapping, so that each derived azimuth in the 360° surround soundfield has a unique representation in the stereo squeezed soundfield.

### Re-panning to downmix

Based on the previous two steps, the virtual sound source derived by Eq. 3.10 is then re-panned to the downmix azimuth derived by Eq. 3.11, based on tangent amplitude.
Figure 3.6 5-Channel Sound System and Auditory Sub-Spaces

Figure 3.7 A ‘Squeezed’ Space Represented by Two Channels
panning law given in Eq. 3.3. The frequency representation of left and right downmix channels, \( L_{dm}(k) \) and \( R_{dm}(k) \), are thus generated by:

\[
L_{dm}(k) = S(k) \cdot \frac{\tan \theta_{dm} + \tan \varphi_{dm}(k)}{\sqrt{2 \cdot \tan^2 \theta_{dm} + 2 \cdot \tan^2 \varphi_{dm}(k)}}
\]
\[
R_{dm}(k) = S(k) \cdot \frac{\tan \theta_{dm} - \tan \varphi_{dm}(k)}{\sqrt{2 \cdot \tan^2 \theta_{dm} + 2 \cdot \tan^2 \varphi_{dm}(k)}}
\]  

(3.12)

Here, in a stereo downmix soundfield, \( \theta_{dm} \) is typically 30°.

The resulting frequency representation of left and right downmix channels, \( L_{dm}(k) \) and \( R_{dm}(k) \), are transformed back to time domain based on the inverse STFT addressed in Section 3.3.1. This effectively results in a stereo downmix signal containing the auditory and localisation information of the original surround soundfield. In addition, the S\(^3\)AC stereo downmix can be further compressed by a conventional MP3 or AAC coder. Effectively, this results in using the same bandwidth as conventional stereo signal, e.g. 128kbps stereo AAC, to transmit a full surround sound.

### 3.3.3 \( S^3 \)AC Decoding

As illustrated in Figure 3.4, the \( S^3 \)AC decoding process starts from applying a time-frequency transform on the received stereo downmix signal (an MP3 or AAC decoding is performed first if further compression is applied on the \( S^3 \)AC downmix). Similar as in the encoding process, a STFT is applied on both of the left and right received downmix channels. This is followed by the following three steps of the \( S^3 \)AC synthesis process: virtual source localisation estimation within the downmix soundfield, 60° to 360° mapping and re-panning to surround.

### Virtual source localisation estimation within the downmix soundfield

The frequency domain virtual sound source and its azimuth in the squeezed soundfield is estimated from the transmitted stereo signal. Here, the virtual sound source
is estimated by:

\[ \hat{S}(k) = \sqrt{L_{dm}^2(k) + R_{dm}^2(k)} \cdot e^{\phi_{dm}} \]  (3.13)

where the phase \( \phi_{dm} \) is chosen as the phase of the channel with higher amplitude amount the left and right downmix channels, so that the resulting decoded signal has a consistent phase. The virtual source azimuth in the downmix soundfield is derived by:

\[ \hat{\phi}_{dm}(k) = \tan^{-1}\left[ \frac{|L_{dm}(k)| - |R_{dm}(k)|}{|L_{dm}(k)| + |R_{dm}(k)|} \tan \theta_{dm} \right] \]  (3.14)

while \( \theta_{dm} \) is typically 30°.

60° to bf 360° mapping

Based on the derived source and azimuth in the 60° stereo downmix soundfield, the source azimuth in the 360° surround soundfield can be derived by inverting Eq. 3.11, such that:

\[ \hat{\phi}_{360}(k) = f^{-1}(\hat{\phi}_{dm}(k)) \]  (3.15)

Re-panning to surround

Based on the recovered virtual source azimuth in the 360° surround soundfield \( \hat{\phi}_{360}(k) \), a corresponding pair of channels, channel \( a \) and \( b \), is chosen to render the virtual sound source at the desired azimuth. The azimuth in the 360° surround soundfield \( \hat{\phi}_{360}(k) \) is transformed to match the selected channel pair by inverting Eq. 3.9:

\[ \hat{\phi}_{ab}(k) = \hat{\phi}_{360}(k) - \varpi_{ab} \]  (3.16)
And the frequency coefficient of the two channels $X_a(k)$ and $X_b(k)$ is subsequently generated by amplitude panning, as:

$$\begin{align*}
X_a(k) &= \hat{S}(k) \cdot \frac{\tan \theta_{ab}(k) + \tan \hat{\phi}_{ab}(k)}{\sqrt{2 \cdot \tan^2 \theta_{ab}(k) + 2 \cdot \tan^2 \hat{\phi}_{ab}(k)}} \\
X_b(k) &= \hat{S}(k) \cdot \frac{\tan \theta_{ab}(k) - \tan \hat{\phi}_{ab}(k)}{\sqrt{2 \cdot \tan^2 \theta_{ab}(k) + 2 \cdot \tan^2 \hat{\phi}_{ab}(k)}}
\end{align*}$$

(3.17)

where $\theta_{ab}(k)$ is half the angular separation of the two chosen channels $a$ and $b$.

Once the S$^3$AC synthesis for all frequencies and all frames are completed, the resulting frequency domain coefficients of all channels are transformed into the time domain using the inverse STFT given in Eq. 3.6. This effectively recovers the 5-channel surround audio signal rendering the original surround sound scene and sound source localisation information.

### 3.4 Evaluation

In this section, the proposed S$^3$AC coding scheme for 5-channel spatial audio signal is evaluated. Existing spatial audio coding techniques, including MP3 Surround [64] and MPEG Surround [9] (also see Section 2.5.5) are used to compare with S$^3$AC. Objective evaluations, focused on preserving the source localisation accuracy, are performed to compared the proposed S$^3$AC and existing spatial audio coding techniques. This is followed by subjective evaluation.

#### 3.4.1 Objective Evaluation

In this section, the localization accuracy of the decoded multichannel audio of S$^3$AC and MPEG Surround is evaluated. A 5-channel example signal rendering a mosquito flying slowly around the horizontal plane is generated for objective evaluation purposes. Figure 3.8 illustrates the movement behavior of the sound source in the exam-
Figure 3.8 A 5-channel signal rendering a mosquito flying around effect

ple signal with the horizontal plane representing the 360° sound scene surface and the z-axis indicates the time frame. The azimuth of the major source in the soundfield is objectively identified by applying the inverse panning law of Eq. 3.8 to the two strongest channels as determined by Eq. 3.7. Three candidate coders, including $S^3$AC, MPEG Surround 525 (stereo downmix with side information) and MPEG Surround Non-Guided (stereo downmix with no side information) [9] are compared against the original soundfield using this measure. All three candidate coders use stereo downmix, while only MPEG Surround 525 uses extra side information. No further compression on the downmix signal is used for all candidate coders, in order to avoid additional distortion. In this evaluation, the virtual source and the respective azimuth are generated on a per-frame-per-frequency basis, using a 50% overlapped 1024-point STFT. The azimuth difference of corresponding frequency coefficients between the original signal and the three coding methods are calculated as the de-
coding azimuth error. Table 3.2 shows the average results of this evaluation for the mosquito signal. It can be seen that S\textsuperscript{3}AC, even though it requires no side information, offers a 7 fold improvement over MPEG Surround 525 and provides localization recoverability only marginally beyond the 1° requirement for perceptually indistinguishable localization (see Section 3.2, also Section 2.3.6). In addition, in comparison with MPEG Surround NonGuided mode, S\textsuperscript{3}AC offers a 30 fold improvement in average localization accuracy.

A more intuitive comparison is given in Figure 3.9 to Figure 3.12. In these figures, frame index, frequency index and azimuth are represented by x-axis, y-axis and z-axis, respectively. Thus, an overview of the localisation representation of a signal is given. Comparing with the localisation of the original signal in Figure 3.9, minimum localisation distortion is introduced in the S\textsuperscript{3}AC coded signal in Figure 3.10, while significant distortion is introduced in the MPEG Surround Non-Guided coded signal in Figure 3.12. Considering that both S\textsuperscript{3}AC and MPEG Surround Non-Guided have no side information, the performance improvements offered by S\textsuperscript{3}AC at no bit rate increase are again apparent.

### 3.4.2 Subjective Evaluation

Two sets of listening tests are performed for subjectively evaluating the proposed S\textsuperscript{3}AC spatial audio coding scheme.

In the first test, 11 multi-channel sound recordings are selected including three immersive recordings, four pop music clips, three classical music clips and one movie sound track. Except the proposed S\textsuperscript{3}AC scheme, two candidate spatial audio codes

<table>
<thead>
<tr>
<th>( S^3 )AC</th>
<th>MPEG Surround 525</th>
<th>MPEG Surround Non-Guided</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.39°</td>
<td>10.28°</td>
<td>42.75°</td>
</tr>
</tbody>
</table>

Table 3.2 Average Decoded Azimuth Error of Three Evaluated Coders

\[ 1.39° \quad 10.28° \quad 42.75° \]
Figure 3.9 A Time-Frequency-Azimuth Mesh of the Original Signal

Figure 3.10 A Time-Frequency-Azimuth Mesh of the S$^3$AC Coded Signal
Figure 3.11 A Time-Frequency-Azimuth Mesh of the MPEG Surround 525 Coded Signal

Figure 3.12 A Time-Frequency-Azimuth Mesh of the MPEG Surround Non-Guided Coded Signal
are used for comparison purpose, including MP3 Surround [64] and MPEG Surround Non-Guided. The listening test employs the MUSHRA methodology [77] (also see Section 2.8.3) with a hidden reference and a 3.5kHz low-pass filtered anchor for each file. Eight listeners including both expert and none expert listeners participated in the test. In this test, all candidate coders use stereo downmix, while the MP3 Surround is the only coder that uses extra side information. The MPEG Surround Non-Guided and S$^3$AC downmix files are compressed using 128kbps AAC and the MP3 Surround downmix is compressed by 192kbps MP3. Figure 3.13 shows the results of this listening test with mean and 95% confidence intervals indicated for each test item and each condition.

The results show that, while the MP3 Surround codec provides the highest quality performance in most items due to its usage of additional spatial cues and higher compressed bit rate for the downmix signal, the S$^3$AC provides a very close performance to the MPEG Surround Non-Guided codec. The overall average quality of the S$^3$AC coded signals is mostly rated around a grade of 80%, which corresponds to ‘good’ to ‘excellent’ quality according to the MUSHRA recommendation [77]. During the test, listeners reported that the S$^3$AC coded signal achieved more accurate localization recoverability, especially when compared with MPEG Surround Non-Guided. However, listeners claimed that the distortion of the monophonic perceptual quality introduced by S$^3$AC is slightly higher than other candidate coders, which results in lower marks than MPEG Surround Non-Guided in some items. This issue is caused by multiple sources overlapped in time-frequency, which will be further addressed in Section 4.3.4.

While the test items used in the first test are ‘background like’ signals with minimum sound source localisation content, five more files with focused localisation content are used to further evaluate the proposed S$^3$AC scheme. These files include:

- Aircraft flying over-head effect
Figure 3.13 Listening Test Results on General 5-Channel Audio Items
• Moving car siren
• Localised male speech
• Localised female speech
• Mosquito flying in a circle (used in Section 3.4.1)

The test methodology is based on MUSHRA [77]. However, the listeners are guided to focus on the localization accuracy of the coded signals. S^3AC, MPEG Surround 525 and MPEG Surround Non-Guided are evaluated and compared in this test. All three coders use the stereo downmix, which is further compressed by 128kbps AAC, while MPEG Surround 525 is the only coder using side information. For each file, a hidden reference and an anchor signal is added, while the anchor signal is generated by 3.5kHz low-pass filtering and equally mixing a mono signal to each channel to remove localisation. Nine listeners, including both expert and none expert listeners, participated in the tests. The results including mean and 95% confidence intervals are shown in Figure 3.14. The results indicate that, without any side information, the localization error of S^3AC cannot be perceptually detected. In addition, the results show that listeners have difficulty in distinguishing the hidden reference, S^3AC and MPEG Surround 525. However, the MPEG Surround Non-Guided coder is easily distinguished as it results in significant distortion of source position.

3.5 Summary

This chapter introduces an efficient approach to spatial audio coding, called S^3AC. Based on identifying frequency domain virtual sources and exploiting perceptual localization redundancy, S^3AC achieves efficient compression of a spatial audio signal by representing it with a ‘squeezed’ downmix signal. A typical application of S^3AC is the compression of ITU 5.1-channel spatial audio, while an implementation based on the STFT is described. Compared with other spatial audio coding techniques,
e.g. MPEG Surround, $S^3$AC requires no side information for efficient representation of a 5.1-channel audio signal as surround sound localisation in the original signal is saved in the $S^3$AC downmix, which is backward compatible to existing stereo audio compression systems. The proposed $S^3$AC system is evaluated both objectively and subjective. The evaluation results show that, in comparison with other spatial audio coding solutions, $S^3$AC has significant advantages in preserving source localisation accuracy without introducing additional bandwidth to a stereo transmission. Based on this, the $S^3$AC technique is further analyzed in the next Chapter.
**Figure 3.14** Listening Test Results on 5-Channel Audio Items with Strong Localisation Feature
Chapter 4

Further Analysis and Improvements

The fundamentals of the Spatially Squeezed Surround Audio Coding (S$^3$AC) have been presented in Chapter 3. Rather than being restricted to the usage of inter-channel spatial cues as found in existing spatial audio coding approaches, S$^3$AC exploits the localization redundancy of the surround sound based on human psychoacoustic principles. A STFT based implementation approach of S$^3$AC has been presented in Chapter 3, which provides efficient and backward compatible compression of an ITU 5.1-channel signal with no side information. In addition, S$^3$AC shows significant advantages in preserving the source localization information compared with existing approaches.

This chapter is focused on further analysis and improvements of S$^3$AC. A window aliasing problem found in the S$^3$AC time-frequency transform is analyzed and evaluated. This is followed by an analysis of the S$^3$AC localisation resolution. This chapter then presents an extension to S$^3$AC, where mono downmixing is used to further reduce the bandwidth while side information representing source localisation is adopted into S$^3$AC. A psychoacoustics based S$^3$AC cue quantisation approach is introduced, while S$^3$AC side information is further exploited for efficient representation of overlapping time-frequency source components. In addition, a source localisation estimation algorithm, which has minimum dependency on the input au-
dio channel format, is presented based on orthogonal analysis of each input channel in 2D Cartesian coordinates.

### 4.1 Window Aliasing in S$^3$AC

The S$^3$AC encoding/decoding system relies on a time-frequency transform, e.g. the STFT presented in Section 3.3.1. In the STFT, to avoid blocking artifacts, an analysis window is applied on each input time frame prior to the application of the time-to-frequency transform. The effect of the analysis window can be removed by applying a synthesis window prior to overlap add reconstruction after the frequency-to-time transform. Perfect reconstruction of the original time-domain can be achieve if the analysis/synthesis window satisfies the condition discussed in Section 2.2.3.

However, in S$^3$AC, the re-panning to downmix process, described in Eq. 3.12, applies a frequency dependent calculation on the STFT coefficients. This effectively
Further Analysis and Improvements

alters the shape of the analysis window on a frequency-dependent basis, i.e. aliasing on the analysis window. This is illustrated in Figure 4.1. The modification of the window function caused by S3AC analysis is dependent on the derived frequency domain source localisation and according S3AC spatial squeezing, as described in Eq. 3.8 and Eq. 3.11. Thus, it is not possible to predict the modification on the window for compensation. As a result, applying a traditional matched analysis/synthesis window approach will not result in complete aliasing cancellation in S3AC. This leads to errors in reconstruction of the stereo downmix signal and consequently errors in reconstruction of the time domain multichannel output signals during decoding.

While the overall distortion introduced by S3AC has been evaluated both objectively and subjectively in Section 3.4, additional analysis quantifies the error introduced by the window aliasing problem. While the filterbank induced error is important, it is more critical to consider the change in location of the virtual sources due to the changes in the frequency coefficients. Hence, the error is calculated as the absolute azimuth difference between the derived source azimuth in the ‘squeezed’ stereo soundfield during encoding (\(\phi_{dm}\) in Eq. 3.11) and the derived source azimuth in the ‘squeezed’ stereo soundfield during decoding (\(\hat{\phi}_{dm}\) in Eq. 3.14). To investigate these errors, three types of time-frequency transforms, including PQMF, STFT and MDCT, are applied on the five multichannel audio signals with focused localisation content used in Section 3.4.2. The total duration of the test files are 92 seconds, with 44.1kHz sampling rate, which results in approximately 8710 frames in a 1024-point 50% overlapping approach. The S3AC encoding/decoding are applied on these signals for error analysis and the average azimuth error for each file and each transform is given in Table 4.1. Since the average displacements are within 0.5° in the squeezed space at worst, the maximum source displacement after recovering to the 360° soundfield is approximately 3°. This is dependent on the ‘360°-to-60°’ squeezing function described in Eq. 3.11. For instance, based on the squeezing criteria described in Table 3.1, for a source localised in the front left region, where percep-
### Table 4.1

<table>
<thead>
<tr>
<th>File</th>
<th>PQMF</th>
<th>STFT</th>
<th>MDCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aeroplane</td>
<td>0.48°</td>
<td>0.35°</td>
<td>0.42°</td>
</tr>
<tr>
<td>Car Siren</td>
<td>0.25°</td>
<td>0.44°</td>
<td>0.39°</td>
</tr>
<tr>
<td>Female Speech</td>
<td>0.16°</td>
<td>0.18°</td>
<td>0.15°</td>
</tr>
<tr>
<td>Male Speech</td>
<td>0.20°</td>
<td>0.17°</td>
<td>0.16°</td>
</tr>
<tr>
<td>Mosquito</td>
<td>0.39°</td>
<td>0.31°</td>
<td>0.36°</td>
</tr>
</tbody>
</table>

Average Decoded Azimuth Error of in the S³AC ‘Squeezed’ Soundfield caused by Window Aliasing

Tual localisation has highest resolution of approximately 1°, an azimuth error of 0.5° in the squeezed downmix soundfield results in azimuth error of 1.5° after recovering to 360° surround soundfield.

### 4.2 Localisation Resolution in S³AC

#### 4.2.1 Frequency Dependent Localisation Resolution

In practice, limited signal quantisation precision is applied during the S³AC encoding/decoding process, which has an impact on the localisation resolution capability of the S³AC encoded signal. In other words, the number of azimuths that the downmix signal can be used to manipulate is limited. Assuming that a 16-bit integer quantisation is applied for the transmission of the encoded signal, i.e. the stereo downmix has an integer value in \( [0, 2^{16} - 1] \). The derivation of source azimuth in the squeezed soundfield during decoding, Eq. 3.14, is re-written as:

\[
\hat{\phi}_{dm}(k) = \tan^{-1} \frac{\|A_L(k)\| - \|A_R(k)\|}{\|A_L(k)\| + \|A_R(k)\|} \tan \theta_{dm} \tag{4.1}
\]

where \( A_L(k) \) and \( A_R(k) \) are the amplitude of the frequency coefficients of the left
Further Analysis and Improvements

and right downmix channels respectively, and \(\|X\|\) stands for rounding to the nearest 16-bit integer. In addition, according to Eq. 3.13, the quantised values of \(\|A_L(k)\|\) and \(\|A_R(k)\|\) are also restricted by the overall source energy \(\|A_S(k)\|\), such that:

\[
0 \leq \|A_L(k)\| \leq \|A_S(k)\|
\]

\[
\|A_R(k)\| = \|\sqrt{\|A_S(k)\|^2 - \|A_L(k)\|^2}\|
\]

(4.2)

This results in having a localisation resolution capability of \(\|A_S(k)\| + 1\) for each frequency, since each pair of \(\|A_L(k)\|\) and \(\|A_R(k)\|\) defined by Eq. 4.2 can be effectively utilized at the decoder to derive one virtual source azimuth. This further suggests that, for each frequency, the number of azimuths that can be stored in an S\(^3\)AC downmix is quantified by the signal amplitude.

In addition to the localisation resolution being limited by the virtual source energy, localisation loss is also introduced by the spatial squeezing step in the S\(^3\)AC analysis, which is modeled by Eq. 3.11. Assuming that, for a frequency \(k\), an S\(^3\)AC virtual source is derived from a channel pair with angular separation of \(2 \cdot \theta_{ab}(k)\), as defined in Eq. 3.8, and squeezed into a sector with size of \(\varepsilon_{dm}(k)\) in the downmix soundfield as defined in Table 3.1, while the total size of the downmix soundfield is \(2 \cdot \theta_{dm}\). For example, if the front-left region rendered by the front-left channel (located at 30°) and the center channel (located at 0°) is selected, the source rendered by this channel pair will eventually be squeezed into the region of \([20°, 30°]\) in the downmix soundfield. This results in that:

\[
\theta_{ab}(k) = 15°
\]

\[
\varepsilon_{dm}(k) = 10°
\]

The spatial squeezing process is effectively equivalent to limiting the range of \(\|A_L(k)\|\)
and $\|A_R(k)\|$, or equivalently $L_{dn}(k)$ and $R_{dn}(k)$ as defined in Eq. 3.12. The resulting localisation precision in the S$^3$AC squeezed soundfield can be modeled by a proportional function between the size of the selected channel pair $2 \cdot \theta_{ab}(k)$, from which the source is derived, and the size of the sector in the squeezed soundfield $\varepsilon_{dn}(k)$, representing the selected channel pair, which is:

$$\rho_S(k) = (\|A_S(k)\| + 1) \frac{\varepsilon_{dn}(k)}{2 \cdot \theta_{ab}(k)}$$  (4.3)

Recalling the localisation blur theory (see Section 2.3.6 and [36]) that indicates a perceptual localisation resolution limit of approximately 1°, the perceptual redundancy of the S$^3$AC downmix signal for transmitting an original soundfield with a size of $\psi$ degrees (typically $\psi = 360^\circ$) is:

$$r(k) = \rho_S(k) - \psi = (\|A_S(k)\| + 1) \frac{\varepsilon_{dn}(k)}{2 \cdot \theta_{ab}(k)} - \psi$$  (4.4)

### 4.2.2 Localisation Resolution of the Low Amplitude Components and Perceptual Relevancy

In order to achieve no loss in the perceived source localisation, a non-negative redundancy value is required in Eq. 4.4. For instance, in a given set of geometrical parameters in the S$^3$AC analysis, including the size of the original soundfield $\psi$, size of the sectors in the squeezed soundfields $\varepsilon_{dn}(k)$ and the selected source analysis channel pair separated by $2 \cdot \theta_{ab}(k)$, the minimum spectral amplitude required to maintain non-negative localisation redundancy can be derived by setting Eq. 4.4 to zero, resulting in:

$$\|A_S(k)\|_{min} = \frac{2 \cdot \psi \cdot \theta_{ab}(k)}{\varepsilon_{dn}(k)} - 1$$  (4.5)

In the typical S$^3$AC coding of ITU 5.1-channel signal, where a 360° surround sound-
Further Analysis and Improvements

field (as given in Figure 3.6) is squeezed into a $60^\circ$ downmix soundfield (as given in Figure 3.7) by following the criteria given in Table 3.1, a condition is given as:

$$\psi = 360^\circ$$

$$\theta_{ab}(k) \in [15^\circ, 40^\circ, 70^\circ] \quad (4.6)$$

$$\varepsilon_{dm}(k) \in [10^\circ, 5^\circ]$$

In addition, based on the squeezing criteria defined in Table 3.1, the ratio between $2 \cdot \theta_{ab}(k)$ and $\varepsilon_{dm}(k)$ can be limited to:

$$\frac{2 \cdot \theta_{ab}(k)}{\varepsilon_{dm}(k)} = \begin{cases} 
3 & \text{for the FL and FR regions} \\
8 & \text{for the L and R regions} \\
14 & \text{for the Rear regions} \\
28 & \text{for the Diagonal 1 and 2 regions}
\end{cases} \quad (4.7)$$

Note that, the highest perceptual localisation resolution of $1^\circ$ applies on the front listening regions only, where

$$\frac{2 \cdot \theta_{ab}(k)}{\varepsilon_{dm}(k)} = 3 \quad (4.8)$$

Hence, it can be derived based on Eq. 4.5 that, the minimum source spectral amplitude of

$$\|A_S(k)\|^\text{min} = 1079 \quad (4.9)$$

is required for a source located at the front region to ensure no localisation loss. For other regions with less localisation resolution, e.g. approximately $10^\circ$ localisation resolution at the L and R regions, Eq. 4.5 should be re-written as:
Further Analysis and Improvements

\[ \|A_S(k)\|_{\text{min}} = \frac{2 \cdot \psi / 10 \cdot \theta_{ab}(k)}{\varepsilon_{dm}(k)} - 1 \]  

(4.10)

As in the L and R regions,

\[ \frac{2 \cdot \theta_{ab}(k)}{\varepsilon_{dm}(k)} = 8 \]  

(4.11)

it results in a minimum source spectral amplitude of:

\[ \|A_S(k)\|_{\text{min}} = 287 \]  

(4.12)

for a source located at the two side regions to ensure no localisation loss.

While this indicates that, for some low amplitude components, the localisation redundancy can be negative resulting in insufficient localisation resolution, these low amplitude components, however, naturally receive less perceptual importance. In addition, spectral source components with low amplitude tend to be below the threshold of hearing or masked simultaneously or temporally by other nearby spectral components. For example, defining the lowest point of the absolute hearing threshold as 0dB with amplitude equivalent to the minimum integer value of 1 in the 16-bit format, Figure 4.2 plots the following four curves:

- Absolute Hearing Threshold

- Source spectral amplitude of 1079 in 16-bit format (61dB) for no localisation loss in the front region, as derived in Eq. 4.9

- Source spectral amplitude of 278 in 16-bit format (49dB) for no localisation loss in the side regions, as derived in Eq. 4.12
Further Analysis and Improvements

Figure 4.2 Absolute Hearing Threshold, Source with 61dB Spectral Amplitude, Source with 49dB Spectral Amplitude and Minimum Masking Curve of a 76dB White Noise

- Minimum masking curve of a 76dB white noise signal (1/10 of the maximum energy of a 16-bit audio signal)

It is shown that, both source amplitudes representing minimum amplitude requirement for no localisation loss in the front and side regions are well below the minimum masking curve of the 76dB white noise. In addition, the analysis here further indicates that,

- While the spatial resolution is given in proportion to the amplitude of the source, louder sources with higher perceptual importance inherently attract higher localisation resolution.

- When there is no priori knowledge of the pre-amplifying and post-amplifying in the actual playback system, quantifying a minimum size of the squeezed soundfield is only an estimation of the ideal condition. For a given target size
of the squeezed soundfield, a post-amplifying process can be applied on the encoded signal to ensure no perceptual localisation loss.

- As can be found in Eq. 4.3, increasing the size of the squeezed soundfield can also improve the localisation resolution if pre-amplifying and post-amplifying of the signal is not desired.

### 4.2.3 S³AC Squeezing Limitation

As discussed in this section, the size of the S³AC downmix soundfield has an impact on the localisation resolution. In Section 3.4.2, subjective experiments have shown that the standard S³AC 360°-to-60° analysis-synthesis process results in no perceived localisation distortion. Here, it is suggested that the S³AC soundfield squeezing can be performed in a more intensive way, i.e. squeeze the surround soundfield into a soundfield smaller than 60°. By setting the redundancy formula of Eq. 4.4 to zero, the smallest size of the sector in the squeezed soundfield in degrees, without causing localisation loss, can be derived as a function of the spectral energy:

\[
\varepsilon_{dm}(k) = \frac{2 \cdot \psi \cdot \theta_{ab}(k)}{\|A_S(k)\| + 1}
\]  \hspace{1cm} (4.13)

In this equation, while the numerator is defined by the size of the soundfield and loudspeaker layout, the source spectral energy component, \(\|A_S(k)\|\), is signal dependent and varies over time-frequency. In order to further quantify the minimum size of the S³AC downmix soundfield without introducing significant distortion in perceptual localisation, subjective analysis is performed.

In this test, the perceptual impact of using different sizes for the downmix soundfield in the S³AC spatial squeezed process is evaluated, in order to find the smallest size of the S³AC squeezed soundfield that does not introduce perceptual localisation loss. In the experiment, standard ITU 5.1-channel files with 16-bit quantisation precision
and 44.1kHz sampling rate were used. The five files with focused source localisation content used in Section 3.4.2 are utilized. Various S^3AC squeezing approaches are applied on the test signals to generate several coding conditions for evaluation, including S^3AC squeezing from 360° to 60°, 40°, 20°, 10°, 5° and 1°, respectively, i.e. the size of the squeezed soundfield, θ_{dm}, as described in Section 3.3.2 varies between these given value to generate different S^3AC downmix signal and decoded respectively. This effectively changes Eq. 3.11 and Eq. 3.12 during S^3AC encoding, and correspondingly changes Eq. 3.14 and Eq. 3.15 during S^3AC decoding. All the S^3AC encoded files are decoded to the standard ITU 5.1-channel format for playback. The volume of the playback systems is adjusted such that a relative 90dB white noise (by defining the source pressure of the minimum integer value of 1 in the 16-bit format as 0dB) is played back in an absolute sound pressure [94] of 90dB in every loudspeaker. The experiment is based on the MUSHRA methodology [77], where the six coding conditions described above were randomly mixed with a hidden reference and a 3.5kHz low-pass-filtered anchor, which is also un-localised by equally mixing a mono signal to each channel. Ten listeners participated in the experiment including both experienced and in-experienced listeners. During the test, listeners are instructed to compare both the perceptual quality and sound source localisation precision between the reference and candidate signals.

The average results with 95% confidence intervals are illustrated in Figure 4.3. It is shown that, for all the test files, different S^3AC spatial squeezing types from 360° to 60°, 40°, 20° and 10° do not introduce any distortion from a statistical point of view, when compared to the un-coded reference. More intensive squeezing types from 360° to 5° and 1° only cause degradation of less than 10 MUSHRA marks for the aeroplane sound scene, while no statistical difference is found for other test files. The results indicate that while the standard 360°-to-60° S^3AC soundfield squeezing does not introduce perceivable distortion, a more intensive squeezing approach such as 360°-to-10° also maintains perceptual and localisation equivalence between the
Further Analysis and Improvements

4.3 S\textsuperscript{3}AC Mono Downmixing and Localisation Side Information

4.3.1 Using Mono Downmix and Cues in S\textsuperscript{3}AC

In addition to the S\textsuperscript{3}AC encoding/decoding system described in Section 3.3, which utilizes a stereo downmix to transmit the auditory and localisation information of a surround soundfield, the S\textsuperscript{3}AC encoding/decoding algorithm can be easily altered to utilize a mono downmix to compress a 5.1-channel audio, while source localisation information can be saved as accompanying S\textsuperscript{3}AC cues.

The S\textsuperscript{3}AC ‘Mono + Side Information’ encoding/decoding system is illustrated in Figure 4.4 and Figure 4.5. Compared with the S\textsuperscript{3}AC stereo encoding/decoding system, presented in Section 3.3, the S\textsuperscript{3}AC mono encoding utilizes the source and azimuth derived from the virtual source localisation estimation process, and saves them as a mono downmix and accompanying side information. Specifically, the frequency

\[\text{Figure 4.3 Listening Test Results Comparing Different Sizes of the S}\textsuperscript{3}AC Downmix Soundfield}\]
domain virtual source, derived from Eq. 3.10 is transformed back to time domain as a mono downmix, while the source azimuth in 360° surround soundfield, derived from Eq. 3.8 and Eq. 3.9, is quantised and saved as the S^3AC side information. As the S^3AC spatial cues directly represents source localisation information (i.e. source azimuth in a 2D soundfield), an efficient quantisation based on perceptual localisation principles can be employed, resulting in a minimum bandwidth requirement for transmitting S^3AC cues, which will be further addressed in Section 4.3.2.

During decoding, the mono downmix is transformed into the frequency domain, while the S^3AC side information is decoded for representing the source localisation information. Similar as the process after deriving the frequency domain virtual source and localisation from the squeezed soundfield in the S^3AC stereo downmix decoding, the S^3AC mono decoder re-pans the virtual source to a selected channel pair based on the localisation derived from the S^3AC cues by utilizing an amplitude panning algorithm. The resulting five-channel signal is transformed back to the time domain. This then effectively recovers a five-channel surround audio signal.
Introducing a ‘mono + cues’ approach in $S^3$AC not only reduces the transmission bandwidth, but provides additional flexibility and extensibility for $S^3$AC. The applications for $S^3$AC ’mono + cues’ approach will be further investigated in Chapter 5.

4.3.2 Location Dependent Perceptual Source Localisation Precision

To minimize bit-rates, the spatial cue side information must be efficiently quantised for transmission. Existing spatial audio coders [8] [10] exploit human auditory frequency sensitivity for spectral quantization; but human auditory localization is not directly exploited. $S^3$AC utilizes localization blur to effect compression of the space through squeezing, but this approach can be improved by recognizing that localization blur is, in itself, location dependent. As addressed in Section 2.3.6, psychoacoustic principles show a localization precision of approximately 1° in front of a listener, reducing to more than 10° on the sides and to the rear of the listener [36]. This leads to approximately 7 bits and 3 bits effective azimuth precision for the 60° front region and 140° rear region respectively, in the ITU 5.1 channel setup.

To further investigate these theories and exploit them in coding applications, listening tests are performed. These tests aim at finding optimal localisation precision for quantisation of sound sources at different regions based on an ITU 5.1-channel setup, i.e. location dependent source localisation precision. Four types of moving sound sources are used, including:

- 500Hz tone
- 1KHz tone
- band-pass noise with two critical-band pass-band with central frequency at approximately 2KHz
Based on an ITU 5.1 channel setup (see Figure 2.2), each sound object is panned into four horizontal regions:

- 30° to -30° in the front
- 30° to 110° in the left
- -30° to -110° in the right
- 110° to -110° in the rear

This effectively results in 16 reference signals. For each reference signal, the source signal is synthesized utilizing amplitude panning to move within the region from one side to the other at a fix speed. For each reference signal, five coded conditions are generated by using different precisions for panning the source to discrete azimuths in the according region. The panning azimuth precisions are:

- 6 bits, equivalently 64 azimuths
- 5 bits, equivalently 32 azimuths
- 4 bits, equivalently 16 azimuths
- 3 bits, equivalently 8 azimuths
- 2 bits, equivalently 4 azimuths

In other words, in these coded conditions, compared with the reference condition, the source signal is only allowed to be panned to discrete azimuths based on a given precision. The listening tests are based on the MUSHRA methodology [77], while the listeners are asked to compare the localization accuracy between the reference

• moving car siren
Further Analysis and Improvements

Figure 4.6 Listening Tests Results of Using Difference Localisation Precision at Different Listening Region

and coded conditions. For each reference signal, a non-moving source signal located at the center of the region is used as the anchor signal. Six listeners participated in the tests, including both expert and non-expert listeners.

The results including mean and 95% confidence intervals are shown in Figure 4.6. In order to highlight location dependent source localisation precision, results from different source files in the same region are statistically analyzed together. It is shown that, for the front and side sources, the perceived distortion increases with the decreasing bit precision while there is strong ambiguity for rear source. According to the results, by using 5 or 4 bits (32 and 16 discrete azimuths respectively) for the front and side azimuth quantisation, the accuracy of the coded material is within 90% when compared to the original. However, in the ambiguous rear plane, azimuth precisions between 5 bits and 2 bits have no statistical difference. These results suggest that, while previous psychoacoustics research indicates higher precision for perceptually undistorted quantization, reduced precision is adequate in coding applications.
Further Analysis and Improvements

### Table 4.2

<table>
<thead>
<tr>
<th>Regions</th>
<th>Linear Azimuth Resolution</th>
<th>Numbers of Azimuth in Region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6-bit</td>
<td>5-bit</td>
</tr>
<tr>
<td><strong>Front</strong> ([-30°, 30°])</td>
<td>2°</td>
<td>3°</td>
</tr>
<tr>
<td><strong>Left</strong> ((30°, 110°])</td>
<td>6.67°</td>
<td>20°</td>
</tr>
<tr>
<td><strong>Right</strong> ((-30°, -110°])</td>
<td>6.67°</td>
<td>20°</td>
</tr>
<tr>
<td><strong>Rear Left</strong> (110°, -180°])</td>
<td>17.5°</td>
<td>35°</td>
</tr>
<tr>
<td><strong>Rear Right</strong> ((-180°, -110°])</td>
<td>17.5°</td>
<td>35°</td>
</tr>
<tr>
<td><strong>Total Number of Azimuths</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2
6-Bit and 5-Bit Codebook Design for Location Dependent \(S^3\)AC Cue Quantisation

#### 4.3.3 \(S^3\)AC Psychoacoustics Based Cue Quantisation and Coding

Based on results presented in Section 4.3.2, a location dependent \(S^3\)AC spatial cue quantisation approach is developed. Two types of codebooks are introduced to quantise each derived \(S^3\)AC cue. As described in Table 4.2, a 6-bit quantisation codebook is derived with 64 discrete azimuths non-uniformly distributed between regions in the 360° circle, while a 5-bit codebook giving a further bit-rate reduction is also derived. Figure 4.7 also illustrates the 5-bit codebook design in a 360° 5-channel surround audio setup. Based on a 50% overlapped 1024-point Short-Time-Fourier-Transform (STFT) implementation, with the frequency bins further grouped into 20 bands equivalent to double ERB frequencies [31] for each frame, the two codebooks result in a fixed bit-rate of 10.36kbps and 8.61kbps for \(S^3\)AC cues, respectively. According to the subjective evaluation presented in Figure 4.6, these codebook designs will result in no significant localisation distortion in the vital frontal region, while the degradation in other regions is less than 10 MUSHRA scores when compared to the multichannel reference where source locations are not quantised.

For further bit-rate reductions, lossless differential coding of the codebook quantised
spatial cues can be used. Due to the band perception property of the human auditory system [31] [94], each frequency band can be assumed to represent a single sound source. Hence, it is expected that the location of the source varies smoothly over time resulting in highly correlated cues. In contrast, spatial cues between adjacent frequency bands represent different sound sources and hence will show less correlation. Therefore, the redundancy remaining in the spatial cues of one frequency band can be further removed using frame-wise differential coding. In this approach, after a codebook index $C(p, k)$ is located for a source azimuth $\varphi_{360}(p, k)$ derived by Eq. 3.8 and Eq. 3.9, the difference between spatial cue codebook indices is derived for the same frequency band between two adjacent frames $p$ and $p - 1$ as,
Further Analysis and Improvements

\[ d(p, k) = C(p, k) - C(p - 1, k) \quad p = 2, 3, \ldots, N \quad (4.14) \]

where \( C(p, k) \) and \( C(p - 1, k) \) are the codebook indices for the \( k \)th frequency band of the \( p \)th and \((p - 1)\)th time frame. Here, \( N \) represents the differential prediction length; hence, a sequence of quantised spatial cue indices is represented by the anchor index, (the first quantised spatial cue index) followed by \( N - 1 \) differential values. The resulting anchor index and following differential values can then be entropy coded for transmission.

Five test signals are used for evaluating the \( S^3 \)AC spatial cue quantisation and coding scheme proposed above, including:

- Audience applause recorded in a concert hall
- Moving car siren
- Localised female speech
- Aircraft flying over-head effect
- Music

Figure 4.8 illustrates the histogram of both differential values \( d(p, k) \) and the original codebook azimuth indices \( C(p, k) \) using a 6-bit codebook, derived from all the test signals. Approximately 175,000 cues are evaluated. Note that the x-axis varies from 0 to 63 for codebook azimuth indices, while it varies from -31 to 31 for codebook differential indices in Figure 4.8. It is shown that, while the probability of the quantised azimuths is evenly distributed over all indices, the differential coded result has a highly centralized distribution, which can be exploited by entropy coding.

In addition, the proposed codebook quantisation and differential coding approaches are evaluated for bit-rate efficiency using the five test files. A 50% overlapped 1024-
Further Analysis and Improvements

point Short-Time-Fourier-Transform (STFT) is implemented on the test signal, with the frequency bins further grouped into 20 bands equivalent to double ERB frequencies [31] for each frame. For the derived $S^3$AC cues, differential prediction lengths $N$ from 5 frames to the same as total number of frames in a file are examined. The resulting cues and differential values are entropy coded using Rice coding [95]. Figure 4.9 illustrates the resulting average cue bit-rate and 95% confidence intervals for both 6 and 5 bit codebook designs. For the 6-bit and 5-bit codebook respectively, using 5 frames for the prediction length reduces the bit-rate from 10.36kbps and 8.61kbps to approximately 6.4kbps and 6kbps, while using 50 frames further reduces the bit-rate to 5.5kbps and 5.4kbps.

The localisation error introduced by the proposed cue quantisation approach is also investigated. For each $S^3$AC cue, the error is calculated as the absolute azimuth difference between the original derived azimuth $\theta_{360}(p, k)$ and the quantised azimuth $C(p, k)$, such that:

$$E(p, k) = |\theta_{360}(p, k) - C(p, k)|$$  \hspace{1cm} (4.15)

Table 4.3 shows the average error for each test file in different regions for both 6 and 5 bit codebook designs, while the overall average error for each file is also given. File 1, which is a recording of concert hall applause having widely distributed localisation, shows highest error, while the error in other signals is limited. Note that, as File 2 has no sound source content in the rear region, no error is calculated for this part. The perceptually vital front region has least error, while more errors are caused by the lower codebook precision in the sides and rear. However, as the codebook design is based on the location dependent perceptual localisation precision, suggested by the subjective experiment presented in Section 4.3.2, the perceptual distortion on localisation caused by these errors is limited.
4.3.4 Using S$^3$AC Side Information for Coding of Sources Overlapped in Time-Frequency

It is less likely that for a particular frequency bin or band, only one source ever occurs in each time frame. Overlapping sound sources are defined here as the occasion that, for each frequency of a certain time frame, there are two or more sound sources rendered by different channel pairs concurrently. For example, for a frequency $k$ in a time frame $p$, a primary source $S_{Pi}(p, k)$ is rendered by the front left/front right channel pair and a secondary source $S_{Se}(p, k)$ is rendered by the rear left/rear right channel pair, as shown in Figure 4.10. Although S$^3$AC provides a highly accurate recoverability of source localization, its performance will be affected if the soundfield contains significant overlapping sound source components. In particular, when the two overlapping sources have energy close to each other, the two sound sources with the same frequency component in the downmix will result in one phantom source that has an azimuth between the two actual ones, as shown in Figure 4.11. Similar problems exist for the inter-channel cue based spatial audio coding approaches, as the source overlapping makes it difficult to determine the transmitted cues; and the downmix process will mix the independent sources together without proper recoverability.
Further Analysis and Improvements

Figure 4.9 Average Cue Bit-Rate (in kbps) and 95% Confidence Intervals of Test Files for Different Prediction Lengths using 6-bit and 5-bit Codebook

The S$^3$AC spatial cues representing source localisation can be introduced to overcome the problem of overlapping sound sources described above. As shown in Figure 4.11, during soundfield analysis, overlapping sound sources are discovered in the same frequency bin, while they are rendered by different channel pairs and have different spectral components and azimuths. The primary source, with higher spectral energy, can be identified as the channel pair with highest energy using the normal S$^3$AC algorithm as presented in Eq. 3.7, such that:

$$|X_a(p, k)|, |X_b(p, k)| = \max_{ij} [|X_i(p, k)|, |X_j(p, k)|]$$  \hspace{1cm} (4.16)

where $i, j \in \{FL, FR, C, RL, RR\}$ and $i \neq j$

while the secondary source can be identified subsequently as:
<table>
<thead>
<tr>
<th>Region</th>
<th>Front</th>
<th>Sides</th>
<th>Rear</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6-bit</td>
<td>5-bit</td>
<td>6-bit</td>
<td>5-bit</td>
</tr>
<tr>
<td>Codebook</td>
<td>6-bit</td>
<td>5-bit</td>
<td>6-bit</td>
<td>5-bit</td>
</tr>
<tr>
<td>File 1</td>
<td>0.5</td>
<td>0.7</td>
<td>1.6</td>
<td>5.0</td>
</tr>
<tr>
<td>File 2</td>
<td>0.3</td>
<td>0.4</td>
<td>1.4</td>
<td>4.4</td>
</tr>
<tr>
<td>File 3</td>
<td>0.4</td>
<td>0.6</td>
<td>3.0</td>
<td>5.7</td>
</tr>
<tr>
<td>File 4</td>
<td>0.5</td>
<td>0.6</td>
<td>1.2</td>
<td>4.1</td>
</tr>
<tr>
<td>File 5</td>
<td>0.3</td>
<td>0.4</td>
<td>3.3</td>
<td>5.9</td>
</tr>
</tbody>
</table>

Table 4.3
Azimuth Error (in degrees) for Quantised S³AC Cues, Compared with Original Derived Cues

\[ |X'_c(p, k)|, |X'_a(p, k)| = \max_{mn} [|X_m(p, k)|, |X_n(p, k)|] \tag{4.17} \]

\[ m, n \in \{\{FL, FR, C, RL, RR\} - \{a, b\}\} \text{ and } m \neq n \]

This is effectively the channel pair with highest energy excluding the primary channel pair. The resulting two pairs are evaluated by inverse amplitude panning to determine the primary and secondary sound sources and their corresponding azimuths, using Eq. 3.8 3.9 and 3.10, resulting in the frequency domain coefficient of the primary source \( S_{Pi}(p, k) \) and its azimuth \( \phi_{Pi,360}(p, k) \), and the secondary source \( S_{Se}(p, k) \) and its azimuth \( \phi_{Se,360}(p, k) \). Subsequently, these two azimuths are mapped from 360° soundfield to the 60° downmix field, using Eq. 3.11, resulting in downmix azimuth for the primary and secondary sources as \( \phi_{Pi,dm}(p, k) \) and \( \phi_{Se,dm}(p, k) \). Based on these, the stereo downmix is synthesized by:
Further Analysis and Improvements

Figure 4.10 Overlapping Sound Source in the 360° Soundfield

\[
L_{dm}(p, k) = S_{Pi}(p, k) \cdot [\tan \theta_{dm} + \tan \varphi_{Pi,dm}(p, k)] \\
+ S_{Se}(p, k) \cdot [\tan \theta_{dm} + \tan \varphi_{Se,dm}(p, k)] \\
R_{dm}(p, k) = S_{Pi}(p, k) \cdot [\tan \theta_{dm} - \tan \varphi_{Pi,dm}(p, k)] \\
+ S_{Se}(p, k) \cdot [\tan \theta_{dm} - \tan \varphi_{Se,dm}(p, k)]
\]  

(4.18)

where $\theta_{dm}$ is typically 30° when using stereo downmix. This stereo downmix is then transformed back to time domain for transmission, while the azimuth information of the primary and secondary source $\varphi_{Pi,dm}(p, k)$ and $\varphi_{Se,dm}(p, k)$ can be quantised and transmitted as accompanying side information. Based on this, the decoder can recover the two sources using:
Further Analysis and Improvements

Figure 4.11 Overlapping Sound Source in the 60° Squeezed Soundfield and the Resulting Phantom Source

\[ \hat{S}_{Pi}(p, k) = \frac{L_{dm}(p, k) - \frac{L_{dm}(p, k) + R_{dm}(p, k)}{2\tan\theta_{dm}} \cdot [\tan\theta_{dm} + \tan\hat{\phi}_{Se,dm}(p, k)]}{\tan\hat{\phi}_{Pi,dm}(p, k) - \tan\hat{\phi}_{Se,dm}(p, k)} \]
\[ \hat{S}_{Se}(p, k) = \frac{L_{dm}(p, k) + R_{dm}(p, k)}{2 \cdot \tan\theta_{dm}} - \hat{S}_{Pi}(p, k) \] (4.19)

where \( \hat{\phi}_{Pi,dm}(p, k) \) and \( \hat{\phi}_{Se,dm}(p, k) \) are the quantised azimuths.

For evaluating the efficiency of this method, it is compared with the MPEG Surround 525 [9], for compressing two 5-channel test files with overlapping sources:

- Two moving sinusoidal tones, both at 1kHz: one rendered by the front left/front right channel pair and moving linearly from the front left to the front right; the other one rendered by the rear left/rear right channel pair and moving linearly from the rear right to the rear left. The energy of the front tone is 4.7dB higher than the rear tone, so that the front source has the dominant contribution to the overall sound scene.
• Two moving bandpass noise sources. The front noise source has a pass band between 500Hz and 1.5kHz, rendered by the front left/front right channel pair and moving from front left to front right. The rear noise source has a pass band between 1kHz and 2kHz, rendered by the rear left/rear right channel pair and moving from rear right to rear left. The two narrow band noise sources thus have overlapping frequency bands between 1kHz and 1.5kHz. The energy of the front noise source is 3dB higher than the rear noise so that the front noise is the dominant source.

This evaluation aims to evaluate the localisation accuracy for both the front and rear sources, by comparing the signal coded by S\(^3\)AC and MPEG Surround 525 with the original signal. For both original and coded signals, the localisation information of the front and rear sources is derived separately by applying inverse amplitude panning (Eq. 3.8) on the related channel pair.

The evaluation result, comparing the original, S\(^3\)AC coded and MPEG Surround coded localisation of the front source in test file 1 is shown in Figure 4.12, while the result for the rear source in test file 1 is given in Figure 4.13. The result shows that, while minimum localisation distortion is introduced by the S\(^3\)AC coding for both the front and rear original source azimuths, the MPEG Surround 525 introduces obvious localization distortion.

The evaluation result for test file 2 is given in Figure 4.14. In Figure 4.14, time-frequency-localisation mesh plot is given for both front and rear sources in the original condition, as well as the coded condition by S\(^3\)AC and MPEG Surround. The left column shows the original and the coded source in the front plane, while the right column shows the result for the rear source. The S\(^3\)AC advantage of preserving accurate source localisation is again revealed.

In addition, Table 4.4 shows the average azimuth error in degrees for the front and
Further Analysis and Improvements

Figure 4.12 Localisation for Original, $S^3$AC Coded and MPEG Surround Coded 1kHz tone source in the front region (azimuth between [-30° 30°]) in Test File 1

rear sources for both test file, after coded by $S^3$AC and MPEG Surround. This error is calculated as the azimuth difference between the original source azimuths and the source azimuths in the two coded versions. It is worth mentioning that, the test signals are synthesized tone and band-limited signal for objective evaluation purpose, while performance when coding real audio signal can degrade. However, this result indicates that $S^3$AC offers a significant higher accuracy for reproducing the localization information of independent overlapping sound sources in comparison with MPEG Surround.

4.4 $S^3$AC Source Localisation Estimation based on Orthogonal Analysis

The analysis for $S^3$AC source localisation estimation presented so far, utilizes pairwise inverse amplitude panning. Although the efficiency of this algorithm for preserving accurate localisation information has been proved by objective and subjective
Further Analysis and Improvements

Figure 4.13 Localisation for Original, $S^3$AC Coded and MPEG Surround Coded 1kHz tone source in the rear region (azimuth between $[-180^\circ, -110^\circ] \cup [110^\circ, 180^\circ]$) in Test File 1

evaluation, this algorithm has limited flexibility. In particular, the source estimation process developed for the ITU 5.1-channel signal format (as presented previously) cannot be used for other multichannel audio signal formats.

This Section presents a novel source localisation estimation algorithm that has minimum dependency on the input multichannel loudspeaker format. A derivation will be given to show that there is fundamental equivalence between two major types of 2D amplitude panning methods, tangent panning law (presented in Section 3.2.2) and vector based amplitude panning (e.g. VBAP [71]). It is further shown that, both methods can be generalized as a universal orthogonal amplitude analysis for any arbitrary 2D loudspeaker setup. This orthogonal analysis algorithm can be used on any multichannel audio signals for source localisation estimation, as long as the location (azimuth in 2D case) of each loudspeaker channel is given. Since it’s fundamentally equivalent to amplitude panning, this orthogonal analysis algorithm provides the same efficiency as the $S^3$AC analysis/synthesis process (based on amplitude pan-
Figure 4.14 The Original, $S^3$AC Coded and MPEG Surround 525 Coded Localization for both Front and Rear Sources in Test File 2
Further Analysis and Improvements

<table>
<thead>
<tr>
<th>Coder</th>
<th>Test File 1 1kHz Tone</th>
<th>Test File 2 Band Pass Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Front Source</td>
<td>Rear Source</td>
</tr>
<tr>
<td>S³AC</td>
<td>0.06°</td>
<td>0.01°</td>
</tr>
<tr>
<td>MPEG Surround</td>
<td>2.44°</td>
<td>25.82°</td>
</tr>
<tr>
<td></td>
<td>0.89°</td>
<td>10.01°</td>
</tr>
<tr>
<td></td>
<td>2.96°</td>
<td>45.4°</td>
</tr>
</tbody>
</table>

Table 4.4
Average Azimuth Error in Degrees, Comparing S³AC, MPEG Surround and Original Front and Rear Source for Two Test Files

... (continued from previous text)

4.4.1 Tangent Panning and Vector Panning

Figure 4.15 shows a circumstance that a source is being panned by two loudspeakers, located at $\alpha$, $\beta$ with different gains of $g_1$ and $g_2$. The resulting source azimuth, $\varphi$, can be estimated by geometrical vector analysis, as shown in the figure.

As used in 2D VBAP [89], a vector based amplitude panning analysis of this condition is represented as:

$$ p = g_1 l_1 + g_2 l_2 $$

where $l_1$ and $l_2$ are the unit-length vectors representing the direction of the two loudspeakers, $p$ represents the source vector, which can be reformed as:

$$ p = g s $$

... (continued from previous text)
where \( g \) represents the gain parameter and \( s \) represents unit-length directional vector pointing the source direction. Hence, Eq.4.20 can be re-written as:

\[
gs = \begin{bmatrix} g_1 & g_2 \end{bmatrix} \begin{bmatrix} l_1 & l_2 \end{bmatrix}^T \tag{4.22}
\]

where \( \cdot^T \) denotes transposition.

Given that:

\[
s = \begin{bmatrix} \cos \varphi & \sin \varphi \end{bmatrix}^T I
\]
\[
l_1 = \begin{bmatrix} \cos \alpha & \sin \alpha \end{bmatrix}^T I
\]
\[
l_2 = \begin{bmatrix} \cos \beta & \sin \beta \end{bmatrix}^T I \tag{4.23}
\]

Eq. 4.22 can be further derived into:
Further Analysis and Improvements

\[
g[\cos \varphi \quad \sin \varphi]^T = [g_1 \quad g_2] \cdot \begin{bmatrix} \cos \alpha & \sin \alpha \\ \cos \beta & \sin \beta \end{bmatrix} \quad (4.24)
\]

or equivalently:

\[
g \cos \varphi = g_1 \cos \alpha + g_2 \cos \beta \\
g \sin \varphi = g_1 \sin \alpha + g_2 \sin \beta \quad (4.25)
\]

and can be re-written as:

\[
\tan \varphi = \frac{g_1 \sin \alpha + g_2 \sin \beta}{g_1 \cos \alpha + g_2 \cos \beta} \quad (4.26)
\]

Comparatively, in a case where tangent amplitude panning law is utilized, as shown in Figure 4.16, the two loudspeakers are symmetrically positioned at ±θ; and the resulting source location is given by:

\[
\frac{\tan \varphi}{\tan \theta} = \frac{g_1 - g_2}{g_1 + g_2} \quad (4.27)
\]

Comparing the two cases in Figure 4.15 and Figure 4.16, it can be found that, the classical tangent amplitude panning with symmetrically positioned loudspeaker pair is a special case of vector panning where:

\[
\alpha = -\beta = \theta \quad (4.28)
\]

by assuming that α is the positive direction. This is proved by substituting Eq (4.28) into Eq. (4.26), which results in Eq. (4.27).
In order to generalize the tangent amplitude panning law, where $\alpha \neq -\beta$, Eq. (4.25) can be solved to give:

$$g_1 = g \frac{\sin(\phi - \beta)}{\sin(\alpha - \beta)}$$
$$g_2 = g \frac{\sin(\alpha - \phi)}{\sin(\alpha - \beta)}$$

(4.29)

It can be further derived that:

$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{\sin(\phi - \beta) - \sin(\alpha - \phi)}{\sin(\phi - \beta) + \sin(\alpha - \phi)}$$
$$= \frac{\cos\left(\frac{\alpha - \beta}{2}\right)\sin\left(\phi - \frac{\alpha + \beta}{2}\right)}{\sin\left(\frac{\alpha - \beta}{2}\right)\cos\left(\phi - \frac{\alpha + \beta}{2}\right)}$$
$$= \frac{\tan\left(\phi - \frac{\alpha + \beta}{2}\right)}{\tan\left(\frac{\alpha - \beta}{2}\right)}$$

(4.30)
While this is derived from Eq.(4.25), which is the case of vector panning, by assuming $\varphi$ as half of the angular discrimination between $\alpha$ and $\beta$:

$$\frac{\alpha - \beta}{2} = \varphi$$

(4.31)

Eq.(4.30) is re-written as:

$$\frac{g_1 - g_2}{g_1 + g_2} = \tan(\frac{\varphi - \frac{\alpha + \beta}{2}}{\tan \theta})$$

(4.32)

It can be then found that, $\frac{\alpha + \beta}{2}$ is the azimuth rotation required to transform the current coordinate to a new coordinate, in which the two loudspeakers are symmetrically positioned as in Figure 4.16. This makes Eq. (4.32) the general tangent amplitude panning law for an arbitrary loudspeaker arrangement. Hence, the equivalence between tangent panning law and vector based amplitude panning is proved.

### 4.4.2 2D orthogonal analysis for virtual sound source localisation

It is the nature of geometries that any vector analysis in an orthogonal Cartesian coordinates system can be replaced by doing orthogonal decomposition analysis. Here, the analysis in Figure 4.15 can be replaced by decomposing the two loudspeaker signals into x-y orthogonal components. As shown in Figure 4.17, the gain $g$ and the direction $\varphi$ of the resulting source vector $p$ can be derived as:

$$g^2 = (g_1 \sin \alpha + g_2 \sin \beta)^2 + (g_1 \cos \alpha + g_2 \cos \beta)^2$$

(4.33)

$$\tan \varphi = \frac{g_1 \sin \alpha + g_2 \sin \beta}{g_1 \cos \alpha + g_2 \cos \beta}$$

(4.34)

This is consistent with the vector analysis shown in Section 4.4.1. Considering
the equivalency between vector amplitude panning and tangent amplitude panning proven in Section 4.4.1, a conclusion can be drawn that both panning methods can be generalized by orthogonal analysis applied to arbitrary positioned loudspeakers.

Based on this, for a multichannel audio signal with $N$ numbers of channels, while the azimuth for each channel is represented as:

$$\theta_i \quad i = 1, 2, \ldots, N$$  \hspace{1cm} (4.35)

the resulting source amplitude rendered by all channels can be derived as:

$$g^2 = \left[ \sum_{i=1}^{N} g_i \cdot \cos \theta_i \right]^2 + \left[ \sum_{i=1}^{N} g_i \cdot \sin \theta_i \right]^2$$  \hspace{1cm} (4.36)

where $g_i$ is the gain parameter of the $i^{th}$ channel. To avoid source energy cancelation, Eq. 4.36 is re-written as:
Further Analysis and Improvements

\[ g^2 = \left[ \sum_{i=1}^{N} |g_i \cdot \cos \theta_i| \right]^2 + \left[ \sum_{i=1}^{N} |g_i \cdot \sin \theta_i| \right]^2 \] (4.37)

while the azimuth of the resulting source can be derived as:

\[ \tan \varphi = \frac{\sum_{i=1}^{N} g_i \cdot \cos \theta_i}{\sum_{i=1}^{N} g_i \cdot \sin \theta_i} \] (4.38)

When applied in the frequency domain, this analysis effectively results in frequency domain virtual sound source and localisation similar to that derived by the virtual source localisation estimation step in normal S$^3$AC encoding, as presented in Section 3.3.2. The S$^3$AC spatial squeezing and re-panning to downmix can be applied for compressing a multichannel signal to stereo, while the source can also be saved as a mono downmix with accompanying side information based on the derived localisation (presented in Section 4.3).

Based on the fundamental equivalence between the proposed orthogonal analysis and the amplitude panning and vector based panning, proved in Section 4.4.1, the proposed orthogonal analysis provides the same efficiency as the amplitude panning based S$^3$AC algorithm, for preserving localisation accuracy, when compressing a 2D spatial audio signal. In addition, this algorithm is extended to 3D soundfield analysis and coding in Chapter 6, while further evaluation will be performed and presented.

### 4.5 Summary

Following the fundamental principles and implementation of S$^3$AC presented in Chapter 3, this Chapter presents further analysis on S$^3$AC as well as extended schemes. A time-domain window aliasing problem caused by S$^3$AC modification on the fre-
Further Analysis and Improvements

frequency domain coefficients is described. The source localisation distortion caused by this problem is analyzed objectively, although listening test results show that the perceptual impact is minimum. The source localisation resolution in the S³AC squeezed soundfield is also analyzed, while it is shown that the localisation resolution is frequency dependent. The limitation of the S³AC spatial squeezing process, without introducing source localisation distortion, is derived, while listening test results are shown to further evaluate the perceptual impact introduced by using an S³AC squeezed soundfield smaller than standard 60° stereo soundfield. A ‘mono downmixing + cues’ approach is then introduced to S³AC to further reduce the bit-rate for compressing a spatial audio signal. Based on perceptual evaluation results, a psychoacoustical localisation based cue quantisation approach is introduced. When combined with a lossless frame-wise differential coding, the proposed cue quantisation approach provides efficient coding of S³AC localisation cues without introducing significant perceptual localisation distortion. In addition, the S³AC localisation cues can be further exploited for efficient representation of multiple sound sources overlapped in both time and frequency. Finally, the fundamental equivalence between the amplitude panning used in S³AC and the vector based amplitude panning is proved, which is further extended for a source localisation estimation algorithm based on orthogonal analysis in Cartesian coordinates. This orthogonal analysis for sound source localisation has improved flexibility and can be used for estimating sound source localisation rendered by any arbitrary 2D loudspeaker setup. In the next Chapter, the proposed S³AC is extended to several applications, including compression of Ambisonics signals, binaural reproduction and spatialised multi-cite teleconferencing, while details of each application are presented.
Chapter 5

S$^3$AC Extended Applications

5.1 Introduction

The fundamental principles and a typical application for efficient compression of ITU 5.1-channel signals of S$^3$AC has been introduced in Chapter 3. Further analysis and extended S$^3$AC coding schemes have also been addressed in Chapter 4, including ‘mono + cues’ S$^3$AC coding, psychoacoustics based S$^3$AC cue quantisation and the limitation of S$^3$AC spatial soundfield squeezing. This Chapter presents three extended S$^3$AC applications, including efficient and backward compatible compression of Ambisonics signal, binaural reproduction of S$^3$AC coded spatial audio signals and utilizing S$^3$AC in a multi-party teleconferencing scenario.

5.2 S$^3$AC Compression of Ambisonics Signals

5.2.1 Ambisonics and Amplitude Panning

The Ambisonics B-format signal [82] provides efficient recording and representation of a surround sound field. As described in Section 2.6, a typical Ambisonics B-format signal contains one omnidirectional channel with sound pressure information and three directional channels with sound velocity information [14], such that:
where the sound source $S$ is located at azimuth $\mu$ and elevation $\eta$ in a 3D soundfield.

For representing a 2D soundfield, the dimension of the Ambisonics B-format can be reduced to 2D by removing the source elevation information, such that:

\[
W = \frac{\sqrt{2}}{2} S \\
X = \cos \mu \cdot S \\
Y = \sin \mu \cdot S \\
Z = \sin \eta \cdot S
\]  

(5.1)

To reproduce a 2D Ambisonics B-format signal over a loudspeaker system, a set of decoding parameters is derived as [81]:

\[
g_{w,i} = \sqrt{2} \\
g_{x,i} = \cos \theta_i \\
g_{y,i} = \sin \theta_i
\]  

(5.3)

where $\theta_i$ is the azimuth of the $i^{th}$ loudspeaker. The signal for the $i^{th}$ loudspeaker $LS_i$ is then calculated as:

\[
LS_i = 0.5 \cdot [(2 - d) \cdot g_{w,i} \cdot W + d \cdot (g_{x,i} \cdot X + g_{y,i} \cdot Y)]
\]  

(5.4)
where $d$ is a directivity factor.

Considering a source located at azimuth $\mu$, as given in Eq. 5.2, to render this source by a pair of loudspeakers symmetrically place at $\pm \theta$, the loudspeaker signals can be derived by substituting Eq. 5.2 into Eq. 5.4, resulting in:

$$LS_1 = 0.5 \cdot [(2 - d) \cdot S + d \cdot (\cos \theta \cdot \cos \mu \cdot S + \sin \theta \cdot \sin \mu \cdot S)]$$
$$LS_2 = 0.5 \cdot [(2 - d) \cdot S + d \cdot (\cos \theta \cdot \cos \mu \cdot S - \sin \theta \cdot \sin \mu \cdot S)] \quad (5.5)$$

Eq. 5.5 can be also represented in a fractional form such that:

$$\frac{LS_1 - LS_2}{LS_1 + LS_2} = \frac{d \cdot \sin \theta \cdot \sin \mu}{(2 - d) + d \cdot \cos \theta \cdot \cos \mu} \quad (5.6)$$

By assuming that:

$$d = \frac{2 \cdot \cos \theta}{\cos \theta + (1 - 2 \cdot \cos^2 \theta) \cos \mu} \quad (5.7)$$

Eq. 5.6 can be expressed in the form of amplitude panning, such that:

$$\frac{LS_1 - LS_2}{LS_1 + LS_2} = \frac{\tan \mu}{\tan \theta} \quad (5.8)$$

While the value of $d$ is dependent on both speaker layout $\theta$ and source azimuth $\mu$, in the most commonly used loudspeaker layout in Ambisonics, where four loudspeakers are placed symmetrically at $\pm 45^\circ$ and $\pm 135^\circ$ ($\theta = 45^\circ$ in Eq. 5.7), $d$ becomes a constant value of 2 to satisfy Eq. 5.8. This demonstrates that amplitude panning underpins the localization theory in common Ambisonics playback.
5.2.2 Conventional UHJ compression of Ambisonics

Conventionally, a UHJ downmixing approach [84] is used for backward compatibility with stereo for an Ambisonics signal. As presented in Section 2.6.2, in UHJ, the Left(L) and Right(R) stereo channels can be chosen as [84]:

\[
L = (0.4699 - 0.171j)W + (0.0928 + 0.255j)X + 0.3277Y \\
R = (0.4699 + 0.171j)W + (0.0928 - 0.255j)X - 0.3277Y
\] (5.9)

where \( j \) is the imaginary unit. The decoding from UHJ to three-channel Ambisonics B-format is calculated as:

\[
W = 0.491(L + R) + 0.082j(L - R) \\
X = 0.209(L + R) - 0.414j(L - R) \\
Y = 0.381(L - R) + 0.192j(L + R)
\] (5.10)

However, results shown in Section 5.2.4 indicate that, significant distortion in source localisation is introduced by the UHJ approach due to the fixed parameter encoding/decoding approach given in Eq. 5.9 and Eq. 5.10.

5.2.3 The \( S^3\)AC approach to compressing Ambisonics

In this section, an \( S^3\)AC based approach for coding Ambisonics signal is presented. This approach provides not only bit-rate reduction of an Ambisonics signal, but efficient backward compatibility to the conventional stereo audio system as well. The encoding and decoding process of this system is illustrated in Figure 5.1 and Figure 5.2.
Figure 5.1 $S^3$AC Encoding System for Ambisonics Signals
Figure 5.2 $S^3$AC Decoding System for Ambisonics Signals
In the approach, a 2D Ambisonics signal is first transformed to the frequency domain, which can be achieved by STFT similarly to that described in Section 3.3.1. In comparison with the fix-parameter encoding approach in UHJ, $S^3$AC encoding is based on analyzing the source localisation information for each frequency domain virtual sound source. A source localisation estimation process is performed on the frequency domain coefficients of the Ambisonics signal, which is based on inverting Eq. 5.2:

$$S(p, k) = \sqrt{2}W(p, k)$$

$$\mu(p, k) = \begin{cases} \tan^{-1} \frac{Y(p, k)}{X(p, k)} & \text{if } Y(p, k) \cdot S(p, k) \geq 0 \\ \tan^{-1} \frac{Y(p, k)}{X(p, k)} + 180^\circ & \text{if } Y(p, k) \cdot S(p, k) \leq 0 \end{cases}$$

(5.11)

where $p$ and $k$ are frame and frequency indices respectively. This results in a monophonic sound source $S(p, k)$ and its azimuth $\mu(p, k)$ in the 360° surround soundfield, as derived in the typical $S^3$AC approach presented in Section 3.3.2. This is shown in the ‘Mono Downmixing’ block in Figure 5.1. If a stereo downmix signal is desired, a 360°-to-60° spatial squeezing approach as described in Eq. 3.11 can be performed, while the sound source is re-panned to the resulting azimuth in 60° to generate the stereo downmix, as shown in the ‘Stereo Downmixing’ block in Figure 5.1. On the other hand, the monophonic sound source can be directly transformed back to the time domain as a mono downmix, while the derived source azimuth can be quantised as side information based on the cue quantisation approach presented in Section 4.3.3.

This $S^3$AC based encoding scheme provides efficient backward compatibility to Ambisonics signals, as the mono/stereo downmix can be played back by conventional audio systems. During decoding, as illustrated in Figure 5.2, the sound source and localisation information can be retrieved by either analyzing the $S^3$AC mono downmix and side information (as shown in the ‘Mono Decoding’ block) or performing
S$^3$AC source localisation estimation on the stereo downmix and re-mapping the azimuth to the $360^\circ$ surround soundfield (as shown in the ‘Stereo Decoding’ block). Based on the retrieved source and localisation, the original B-format signal can be re-synthesized according to Eq. 5.2, as:

\[
\begin{align*}
\hat{W}(p, k) & = \frac{\sqrt{2}}{2} \hat{S}(p, k) \\
\hat{X}(p, k) & = \cos \hat{\mu}(p, k) \cdot \hat{S}(p, k) \\
\hat{Y}(p, k) & = \sin \hat{\mu}(p, k) \cdot \hat{S}(p, k)
\end{align*}
\] (5.12)

where $\hat{S}(p, k)$ and $\hat{\mu}(p, k)$ are the time-frequency representation of the recovered source and azimuth.

As shown in Figure 5.2, to reproduce the signal over a loudspeakers, an Ambisonics reproduction process, as described in Section 2.6.4, can be used if a regular loudspeaker array is available, e.g. a hexagon array [83]. However, for an irregular loudspeaker system, such as ITU 5.1-channel setup, Ambisonics reproduction is not desired as the recovered source localisation accuracy is reduced [83]. Although alternative Ambisonics reproduction methods are presented in [96] [97], additional complexity is introduced. While the equivalence between Ambisonics reproduction and amplitude panning is proved in Section 5.2.1, the amplitude panning method used in normal S$^3$AC decoding, as addressed in Section 3.3.3, is used for reproducing S$^3$AC coded Ambisonics signal to a multichannel loudspeaker system, typically ITU 5.1-channel system. In this approach, transforming the source $\hat{S}(p, k)$ and azimuth $\hat{\mu}(p, k)$ information derived from the compressed stereo or ‘mono + side information’ signal to Ambisonics B-format (as given in Eq. 5.12) is not required, while normal S$^3$AC decoding approach presented in Section 3.3.3 can be applied to the derived source $\hat{S}(p, k)$ and azimuth $\hat{\mu}(p, k)$ for surround sound reproduction.

While the encoding/decoding system described above provides compression of a
2D Ambisonics signal, the representation of a 3D Ambisonics signal can also be achieved based on the $S^3$AC approach. For a 3D Ambisonics recording, as described in Eq. 5.1 the sound source and localisation (both azimuth $\mu$ and elevation $\eta$) can be derived in the frequency domain as:

\[
S(p, k) = \sqrt{2}W(p, k)
\]
\[
\eta(p, k) = \sin^{-1}\frac{Z(p, k)}{S(p, k)}
\]
\[
\mu(p, k) = \begin{cases} 
\tan^{-1}\frac{Y(p, k)}{X(p, k)} & \text{if } Y(p, k) \cdot S(p, k) \geq 0 \\
\tan^{-1}\frac{Y(p, k)}{X(p, k)} + 180^\circ & \text{if } Y(p, k) \cdot S(p, k) \leq 0
\end{cases}
\]

(5.13)

The resulting source and azimuth/elevation localisation can be either represented as a mono downmix with accompanying localisation side information or squeezed into an $S^3$AC stereo downmix. The approach for exploiting $S^3$AC to compress a 3D soundfield will be further addressed in Chapter 6.

### 5.2.4 Evaluation

The proposed $S^3$AC compression approach to Ambisonics signals is evaluated both objectively and subjectively. The conventional UHJ stereo downmix approach is compared with the $S^3$AC scheme. Three modes of $S^3$AC are evaluated, including:

- $S^3$AC stereo downmix (abbreviated as $S^3$AC SD)
- $S^3$AC mono downmix with un-quantised side information ($S^3$AC MD-UQ)
- $S^3$AC mono downmix with quantized side information ($S^3$AC MD-Q)

The psychoacoustics based approach presented in Section 4.3.3 is utilized for quantising the localisation cues in the $S^3$AC MD-Q mode, using a bit-rate of approximately 10kbps. No further compression is applied on the downmix signals of the
UHJ approach and the three S3AC modes. Eight 2D Ambisonics recordings, including immersive soundfield, live concert recordings and surround rendered music, are used as test signals.

The Kullback-Leibler Spectral Distance measurement [98] is used for objective evaluation of the encodings and is calculated between the original signal and all four coding conditions. For each component signal, the average Kullback-Leibler Spectral Distance for each coefficient in each channel of 2D B-Format is calculated according to:

\[
D_i = \frac{1}{P \cdot K} \sum_{p=1}^{P} \sum_{k=1}^{K} \left[ C_i(p, k) - \hat{C}_i(p, k) \log \frac{C_i(p, k)}{\hat{C}_i(p, k)} \right]
\]

(5.14)

where \( p, k \) are frame and frequency indices; \( P, K \) the total number of frames and frequency bins; \( C_i(p, k), \hat{C}_i(p, k) \) are the original and recovered B-format channels while \( i = W, X \) or \( Y \) for the three channels of B-Format. For all the eight test files, the three S3AC Ambisonics coding modes are used to compress the B-format and decoded for evaluation. The UHJ encoding/decoding approach is also applied on the test files according to Eq. 5.9 and Eq. 5.10 for comparison. This Kullback-Leibler spectral distance is calculated between the original B-format files and coded files for all the eight test files, while the average result for each channel in the B-format and overall average are both given. The result for each S3AC coding mode is also given separately, as shown in Table 5.1.

While distortion in the W channel relates to perceptual quality degradation, distortion in the X and Y channel will result in error in localization [82]. Since only sound pressure information is stored in the W channel, the X and Y channels contain the azimuth-localization information. The results show that, in all three modes, S3AC gives more precise recovery of the spectrum than UHJ for all of the B-format channels. This suggests that, in comparison with UHJ, S3AC will achieve subjective
improvement in terms of both perceptual quality and source localization. The W channel for \( S^3 {\text{AC MD-UQ}} \) and MD-Q is undistorted, as the mono downmix in these two modes is a perfectly scaled version of the original W channel, as described in Section 5.2.2. While the quantization of azimuth side information adds distortion, it is further evaluated by listening tests that no perceptual impact is introduced. In the listening tests, the same test materials and \( S^3 {\text{AC}} \) coding conditions used in Kullback-Leibler Spectral Distance are used, while the UHJ approach is also used for comparison. The original and coded B-Format files are converted into ITU 5.1-channel format, using Eq. 5.3 and Eq. 5.4 based on the standard loudspeaker positions. The MUSHRA methodology [77] is employed for the listening test, using six listeners including both experienced and non-experienced listeners. The original B-format files are decoded to 5-channel signals and used as reference conditions and hidden reference conditions. An un-localized 3.5kHz low-pass filtered version is used as the anchor condition.

The results including mean and 95% confidence intervals are shown in Figure 5.3. Compared with UHJ, all three \( S^3 {\text{AC}} \) approaches show significantly higher scores, with an average 25% improvement in the MUSHRA score. In addition, it should be noted that, while quantization of the \( S^3 {\text{AC}} \) side information objectively increases

<table>
<thead>
<tr>
<th></th>
<th>( W \times 10^{-2} )</th>
<th>( X \times 10^{-1} )</th>
<th>( Y \times 10^{-1} )</th>
<th>( W-X-Y \text{ Average} \times 10^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{UHJ} )</td>
<td>7.59</td>
<td>1.99</td>
<td>1.68</td>
<td>1.47</td>
</tr>
<tr>
<td>( S^3 {\text{AC SD}} )</td>
<td>3.06</td>
<td>1.35</td>
<td>1.38</td>
<td>1.01</td>
</tr>
<tr>
<td>( S^3 {\text{AC MD-UQ}} )</td>
<td>0.00</td>
<td>1.37</td>
<td>1.24</td>
<td>0.87</td>
</tr>
<tr>
<td>( S^3 {\text{AC MD-Q}} )</td>
<td>0.00</td>
<td>1.77</td>
<td>1.53</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table 5.1
W, X, Y Channel and Average Kullback-Leibler Spectral Distance of Eight 2D Ambisonics Recordings
Figure 5.3 Listening Tests Results Comparing $S^3$AC and UHJ Compression of 2D Ambisonics

the spectral distortion, no significant perceptual distortion is detected in the listening tests, when compared with the $S^3$AC MD-UQ mode. This further indicates that the psychoacoustics based spatial cue quantization approach, as described in Section 4.3.3, can be efficiently used to further reduce the bit-rate of $S^3$AC side information without introducing perceptual localization distortion.

5.3 $S^3$AC Binaural Reproduction

5.3.1 System

The $S^3$AC spatial audio coding scheme and extended applications presented so far in this thesis, have been focused on reproducing the encoding surround soundfield information over multi-channel loudspeaker systems. For users without multi-channel loudspeaker systems, virtual surround sound playback over stereo headphone can be achieved by exploiting a HRTF based (see Section 2.3.5) decoding approach. Ex-
isting spatial audio coding approaches, e.g. MPEG Surround, have adopted HRTF based approach for localised headphone playback. However, as discussed in Section 2.5.6, while typical HRTF measurements are performed based on angular source localisation, additional computational complexity is introduced to transform arithmetic-based spatial cues to actual source localisation information in these approaches.

This section describes a solution that further exploits source localisation information in S\textsuperscript{3}AC for efficient binaural reproduction. As presented earlier in this thesis, rather than exploiting inter-channel arithmetical relationships, S\textsuperscript{3}AC fundamentally derives source localisation information, represented as source azimuth in a 2D soundfield scenario, during its encoding and decoding process. Based on this, no additional derivation is required to locate a correct HRTF for re-rendering a sound source with localised perception over headphone signals. For any type of S\textsuperscript{3}AC coded signal, including both S\textsuperscript{3}AC stereo downmix and S\textsuperscript{3}AC mono downmix with cues, the S\textsuperscript{3}AC binaural reproduction system is illustrated in Figure 5.4.

In this system, similar to other S\textsuperscript{3}AC applications, the frequency domain virtual sound source \( S(p, k) \) and localisation \( \mu(p, k) \) can be either derived from an S\textsuperscript{3}AC stereo downmix or recovered from an S\textsuperscript{3}AC mono downmix with localisation cues. For each virtual source \( S(p, k) \), the ear entrance signals can be derived by filtering the virtual source with the HRTF pair measured at the same location as the derived source localisation \( \mu(p, k) \). This can be described as:

\[
\begin{bmatrix}
B_L(p, k) \\
B_R(p, k)
\end{bmatrix}
= S(p, k)
\begin{bmatrix}
H_L(\mu(p, k), k) \\
H_R(\mu(p, k), k)
\end{bmatrix}
\]  

(5.15)

where \( B_L(p, k) \) and \( B_R(p, k) \) are the time-frequency representation of the left and right ear entrance signals, \( H_L(\mu(p, k), k) \) and \( H_R(\mu(p, k), k) \) are the \( k^{th} \) frequency components of the left and right HRTF measured at \( \mu(p, k) \). This method results in one pair of ear entrance signals for each frequency domain virtual source, while
Figure 5.4 $S^3AC$ Binaural Reproduction System
the full bandwidth signal is assembled from the combined signals of all bands. This combined signal is then transformed to the time domain for binaural playback.

### 5.3.2 HRTF Interpolation

Due to the limited measurement resolution of the HRTF databases, HRTF interpolation must be performed to generate a database with adequate resolution, especially for dynamic sound sources. For example, in the KEMAR HRTF database [39], measurements are made for every 5° in the listening plane. Based on the 1° perceptual localisation resolution suggested by the localisation blur theory [36], the HRTF database needs to be interpolated to a resolution of 1° to avoid localisation distortion. While detailed investigation on the efficiency of HRTF interpolation technique is not the scope of this thesis, the bilinear interpolation method [40] is used to extend the KEMAR database for evaluation purpose. For interpolating HRTFs in a 2D horizontal soundfield, the bilinear interpolation algorithm can be expressed as:

\[
\hat{h}(n) = (1 - \frac{C_\mu}{\mu_{\text{grid}}}) h_a(n) + \frac{C_\mu}{\mu_{\text{grid}}} h_b(n) 
\]

(5.16)

where \( \hat{h}(n) \) is the head-related impulse response (HRIR) of a point located in an arbitrary azimuth between two adjacent HRIR measurements \((h_a(n) \text{ and } h_b(n))\) located at positions \(a\) and \(b\), \(C_\mu\) is the related position and \(\mu_{\text{grid}}\) is the original measurement resolution, as illustrated in Figure 5.5.

### 5.3.3 Evaluation

The proposed S\(^3\)AC binaural reproduction system is evaluated by subjective listening tests. Eight test files, including both ITU 5.1-channel signals and Ambisonics signals, are used for the evaluation and described in the following:

- 5.1-channel signal rendering aircraft flying overhead
Figure 5.5 HRTF Interpolation

• 5.1-channel signal rendering a car siren signal moving from left to right in the front region

• 5.1-channel signal rendering a female speech signal spoken consecutively at $\pm 15^\circ$, $\pm 70^\circ$ and $\pm 145^\circ$

• 2D Ambisonics signal rendering a dynamically flying flea with background music

• 5.1-channel signal rendering a mosquito flying a circle

• 2D Ambisonics signal rendering a percussion instrument played around a circle

• 5.1-channel signal rendering a male speech signal spoken consecutively at $\pm 30^\circ$, 0° and $\pm 110^\circ$

• 2D Ambisonics signal rendering a car engine noise with background music

A modified version of MUSHRA is used for subjective evaluation. The original multi-channel surround audio signals (reproduced to 5.1-channel for Ambisonics
signals) are used as a reference for evaluating binaurally reproduced S$^3$AC coded signals. Since there is no perfect transformation from a multi-channel signal set to the binaural signals in terms of both sound quality and localisation, no hidden reference condition is used in this modified MUSHRA. Instead, listeners are instructed to directly compare the audio quality and localisation performance of the headphone playback versions against the original multi-channel sounds. A mark out of 100 was then given to each binaural item according to the MUSHRA guideline described in Section 2.8.3. The KEMAR HRTF database is used for evaluating the proposed S$^3$AC binaural reproduction system. Since the KEMAR HRTF database is recorded without room reflections, this set of listening tests is conducted in an anechoic chamber to ensure reverberant-free playback of the multi-channel references; hence, the multichannel audio and HRTF filtered headphone versions could be compared by listeners.

In these listening tests, three types of S$^3$AC encoding and binaural reproduction conditions are used:

- S$^3$AC mono downmix of original signals with S$^3$AC spatial cues quantised to approximately 6kbps based on the quantisation approach presented in Section 4.3.3, and binaural decoded using the original KEMAR HRTF database (abbreviated as Mono)

- S$^3$AC stereo downmix and binaural decoding using the original KEMAR HRTF database (abbreviated as Stereo)

- S$^3$AC stereo downmix and binaural decoding using the interpolated KEMAR HRTF database (abbreviated as Interpo)

The bilinear HRTF interpolation technique presented in Section 5.3.2 is used to artificially improve the resolution of KEMAR HRTF database to 1°. Three additional
conditions (abbreviated as AAC Mono, AAC Stereo, AAC Interpo) are also evaluated, where mono/stereo downmix signals in the above three conditions are further coded by AAC in 64kbps per channel. An anchor condition is created using a 3.5kHz low-pass filtering of the mono version of each test signal and played-back over headphones. Eight listeners including both experienced and non-experienced listeners participated in the tests.

The results including mean and 95% confidence intervals are shown separately for each test file in Figure 5.6. The results show that, in comparison with multi-channel playback, the proposed S3AC binaural reproduction systems achieve marks over 80 for most of the test files (i.e. very accurate localisation with good quality), and over 60 for the remainder of the test files (i.e. good localisation with satisfactory quality). Ambisonics files show lower average marks compared with other files and larger confidence intervals. It is suggested that this is due to the more complex soundfields represented by these files, which contain multiple sources and fast moving objects, and which led to non-expert listeners finding difficulties in comparison and marking.

Little improvement is introduced by the HRTF interpolations, while no significant degradation is caused by the S3AC quantisation process in the Mono mode. Comparing with the listening test results given in Section 4.3.2 where reduced source localisation resolution introduces perceptual distortion in a multi-channel playback scenario, this indicates a lower localisation resolution requirement in binaural reproductions when compared to loudspeaker playback. AAC compression does not introduce significant perceptual quality loss either. This proves the backward compatibility of the proposed approach as well as the bit-rate efficiency. For example, using the AAC Mono mode can achieve a bit-rate as low as 70kbps (64kbps for mono downmix and 6kbps for S3AC cues) for compressing a spatial audio signal, while the S3AC binaural reproduction method provides virtual surround soundfield with accurate source localisation over headphones.
5.4 S\textsuperscript{3}AC in Multi-Party Teleconferencing

5.4.1 System

Teleconferencing is an efficient and effective technology for connecting geographically distributed participants in meetings for business, education, or for connecting remote communities. Commercial teleconferencing systems currently available, although offering sophisticated video stimulus of the remote participants, commonly employ only mono and stereo audio playback for the user; however, telepresence can be greatly improved by spatialising the audio to assist listeners to distinguish between (concurrent) participating speakers [100] [101]. A system based on binaural techniques is introduced in [102], where online avatars are created to co-locate remote participants in virtual auditory space with spatialised binaural reproduction.
over headphones. In this approach, speaker location cues are applied to monaural speech to create a user-manipulable soundfield that matches the avatar’s position in the binaural virtual space. However, this system is limited to binaural playback. A different approach is introduced in [103], which applies the Directional Audio Coding (DirAC, also see Section 2.5.3) technique to record, efficiently transmit, and render the remote spatial soundfield. However, this approach does not address the spatialisation of multiple remote sites, and requires specific Ambisonics recording hardware, which can be expensive.

To improve the users’ feel of telepresence, an \textit{\textsuperscript{S}3\textsuperscript{AC}} based solution is proposed in this Section as an efficient approach to spatialised teleconferencing, which provides both efficient representation of teleconferencing recording and unambiguous rendering of multiple remote soundfields. For maximum flexibility, the proposed system utilizes a standard 5.1-channel playback system for rendering and does not require specific recording hardware. Rather, only a mono speech stream accompanied by speaker azimuth metadata is required for the representation of spatialised speech soundfield information for each participating site. This system then merges soundfield information from multiple remote sites unambiguously into a 5.1-channel surround setup at the users’ end. A novel algorithm to ‘squeeze’ multiple soundfields together is introduced, based on the fundamental spatial squeezing principles in \textit{\textsuperscript{S}3\textsuperscript{AC}}.

Figure 5.7 illustrates the proposed teleconferencing recording and playback system. With \( N \) geographically distributed sites concurrently participating in the teleconference of Figure 5.7, each site must thus unambiguously spatialise \( N - 1 \) remote sites. The two main components of the proposed system are: spatialised teleconferencing recording and efficient transmission of audio and spatial metadata between sites, e.g. over the Internet, and merging the \( N - 1 \) remote soundfields at each site using the \textit{\textsuperscript{S}3\textsuperscript{AC}} based ‘squeezing’ approach.
5.4.2 Recording and Representation of Localised Speech Signal

Multiparty meetings are generally recorded with multiple omnidirectional microphones, arranged in an array for signal enhancement and speaker location estimation. In the proposed system, the only recording requirement of each participating site is a mono speech stream transmitted with the speaker azimuth metadata. Thus, any recording hardware setup and speaker azimuth estimation algorithm can be employed. Without loss of generality, a four-element array of omnidirectional microphones is employed, as illustrated in Figure 5.8, with the speaker azimuths estimated...
using the Steered Response Power with PHAse Transform (SRP-PHAT) algorithm [104].

In SRP-PHAT, the source location estimation starts from deriving the time-delay estimation (TDE) $\hat{\tau}$ based on a Generalized Cross Correlation PHAse Transform (GCC-PHAT) [105]. For a microphone channel pair between channel $m$ and channel $n$ with TDE $\tau_{mn}$, $\hat{\tau}$ is estimated by:

$$\hat{\tau} = \arg \max_{\tau_{mn}} \left( \int_{-\infty}^{\infty} \frac{X_m(\omega) \cdot X^*_n(\omega)}{|X_m(\omega) \cdot X^*_n(\omega)|} e^{j\omega \tau_{mn}} d\omega \right)$$

(5.17)

where $X_m(\omega)$ denotes the Discrete Fourier Transform (DFT) of the $m^{th}$ microphone channel. SRP-PHAT thus employs GCC-PHAT in a delay-and-sum beamformer to calculate the SRP, $P(\mu)$:
\[ P(\mu) = \left( \sum_{n=1}^{C} \sum_{m=1}^{C} \int_{-\infty}^{+\infty} \frac{X_m(\omega) \cdot X_n^*(\omega)}{|X_m(\omega) \cdot X_n^*(\omega)|} e^{j\omega \Delta_{mn}(\mu)} d\omega \right) \]  

(5.18)

where \( C \) is the total number of microphone channels, \( \Delta_{mn}(\mu) \) is the steering delay between each candidate source location \( \varphi \) of the SRP search space and microphone pair between channels \( m \) and \( n \). The estimated source location \( \hat{\varphi} \) is thus computed as the candidate location \( \mu \) that maximizes \( P(\mu) \):

\[ \hat{\mu} = \arg \max_{\mu} P(\mu) \]  

(5.19)

Such an exhaustive search of all \( \mu \) defined a priori in the SRP search space can be computationally expensive. However, this can be overcome by reducing the search space based on prior knowledge of the microphone array geometry and room dimensions, which can be easily retrieved or calculated [106].

Based on the derived source location, one of the microphone channels or an enhanced speech signal derived from the array can be used as the mono meeting speech signal in the proposed system, with accompanying metadata containing speaker localisation information.

### 5.4.3 S\(^3\)AC Squeezed Soundfield Reproduction

In the proposed system, each participating conference site receives multiple encoded soundfields from each remote site, which is derived based on the algorithm described in Section 5.4.2. It is desired that soundfield information from different remote sites can be perceptually disambiguated easily, while different speakers in one remote site can also be distinguished. For this purpose, the S\(^3\)AC technique is used to reproduce the ‘squeezed’ soundfield representing multiple remote conference sites. Figure 5.9 illustrates a teleconferencing scenario with three participating sites, where two speakers at different sites may be located too close to be disambiguated (e.g.
a speaker located at 121° in site 2 and a speaker located at 115° in site 3) if spatialised with the original speaker azimuths at the third site. Figure 5.10 illustrates a more complicated scenario with five participating sites, which can result in significant perceptual ambiguity if all remote speakers are rendered to their original locations. To enhance discriminated speaker localization between different conference sites, soundfield information transmitted from each remote site containing the full 360° localization information is squeezed into a unique sector for the user. This is achieved by applying a bijective azimuth mapping function, \( \Theta_n \), on the transmitted azimuth of each remote site:

\[
\xi_n = \Theta_n(\mu_n)
\]  

(5.20)

where \( \mu_n \) is the transmitted source azimuth from the \( n^{th} \) remote site, and \( \xi_n \) is the derived squeezed azimuth in the rendering site. This effectively results in using a unique rendering sector for each remote site. For example, in Figure 5.9, the surround soundfield from site 2 is squeezed into the front 180° sector rendering site 1, while site 2 is squeezed into the rear 180° sector. The azimuth mapping function \( \Theta_n \) is adaptively defined depending on the number of sites and number of participants per site to be spatially rendered. For example, while ‘squeezed’ sectors of equal widths are allocated to remote sites in Figure 5.9 and Figure 5.10, the azimuth mapping function can be modified such that remote sites with a large number of speakers can be assigned a larger sector for unambiguous rendering between speakers from this site. In this squeezing process, while speakers from different remote sites are displaced, the spatial relationship between speakers at each site remains intact.

The transmitted mono speech stream from each remote site is then rendered by the \( S^3AC \) amplitude panning process to the squeezed sector, using the two loudspeakers closest to each mapped azimuth. Similarly as the standard \( S^3AC \) process, the loudspeaker signal is processed in the frequency domain, based on the squeezed azimuth.
\[ \xi_n \text{ obtained by Eq. 5.20, as:} \]

\[ LS_1(p, k) = S_n(p, k) \cdot \frac{\tan \theta_{ab}(p, k) + \tan \xi_n(p, k)}{\sqrt{2 \cdot \tan^2 \theta_{ab}(p, k) + 2 \cdot \tan^2 \xi_n(p, k)}} \]

\[ LS_2(p, k) = S_n(p, k) \cdot \frac{\tan \theta_{ab}(p, k) - \tan \xi_n(p, k)}{\sqrt{2 \cdot \tan^2 \theta_{ab}(p, k) + 2 \cdot \tan^2 \xi_n(p, k)}} \]  

(5.21)

where \( p \) and \( k \) are the frame and frequency indices, \( LS_1(p, k) \) and \( LS_2(p, k) \) are the two loudspeaker signals used for reproduction, \( S_n(p, k) \) is the transmitted stereo signal from the \( n^{th} \) site, \( 2 \cdot \theta_{ab}(p, k) \) is the angular separation of the two chosen channels. \( LS_1(p, k) \) and \( LS_2(p, k) \) are then transformed back to the time-domain to form the loudspeaker feed signals.
5.4.4 Evaluation

To illustrate the proposed teleconferencing system, simulations are conducted from the point of view of a teleconference with N-1 remote participating sites. That is, there are N teleconference sites in total: N-1 remote sites plus the user site, referred as Site 1, spatialising the N-1 remote sites. Two simulation scenarios, as shown in Figure 5.9 and Figure 5.10, are thus conducted with this paradigm, described from the point of view of Site 1 as:

- Scenario 1: two remote meeting sites of two participants each
- Scenario 2: four remote meeting sites, two from the first simulation scenario plus two more of three and four participants each

While ground-truth speaker azimuths (as measured from the positive x-axis) are shown underneath each speaker in Figure 5.9 and Figure 5.10. Speakers at each four remote meeting site are placed at similar azimuths to maximally illustrate the advantage of ‘squeezing’ soundfields that would otherwise overlap if remote meeting soundfields are simply re-synthesized using the original speaker azimuths.

Meeting recordings are simulated using reverberant-free speech recordings. A virtual meeting room of dimensions 3m×3m×3m is used to spatialise speech signal at all meeting sites. Reverberation times (RT60) from 0s (anechoic) to 0.5s are modeled using Allen and Berkeley’s image method described in [107]. To record the simulated meeting speech at remote sites, four omnidirectional microphones placed 20cm apart centered around the origin is modeled at each site, with speakers located on the unit circle. This recording setup is shown in Figure 5.8.

A total of seven different speakers are thus required for the two simulated teleconferencing scenarios. Each teleconference site plays out each speaker in turn, without any speaker overlap. Seven anechoic speech sentences from different speakers,
four female and three male, each approximately 5s in duration are sourced from the Australian National Data-base of Spoken Languages (ANDOSL) [108]. Speech sentences are normalized and downsampled from 20kHz to 16kHz, and stored at 16 bits/sample.

For each of the two simulation scenarios, results are presented as graphical plots of the speaker azimuths from all participating teleconference sites as estimated from SRP-PHAT (i.e., original azimuth) and after ‘squeezing’ into Site 1’s soundfield for site and speaker disambiguation. To illustrate the effect of increasing reverberation time, the speaker azimuths are plotted in concentric circles of increasing reverberation time (RT60=0s to 0.5s in 0.1s increments) with increasing circle radius. The simulation results for the two scenarios are described in the following:

**Simulation scenario 1 with three participating sites**

Figure 5.11 illustrates the ground-truth speaker azimuths for the two remote sites, with the azimuths estimated from SRP-PHAT shown in Figure 5.12. Note that the legend from Figure 5.11 also applies to Figure 5.12 and Figure 5.13. It can clearly be seen from Figure 5.12 that, besides the variance of the azimuth estimation from SRP-PHAT due to increasing reverberation, additional ambiguity is introduced since the speakers’ ground-truth azimuths at Site 2 and 3 are too close to be distinguished. Hence simply re-rendering the speakers to the ground-truth azimuths at their original site will cause spatial overlap at Site 1, where the user will not be able to easily disambiguate between speakers 1 or 2 from either site. Figure 5.13 shows the ‘squeezed’ azimuths for distinguishing speakers from different remote site, based on the proposed S³AC based scheme. Site 2 has been squeezed to the top half of the listening circle, whilst Site 3 is squeezed to the bottom half. The speakers within each site and between sites are clearly spatially separated, even in higher reverberation times where the azimuth estimations from SRP-PHAT exhibit greater variance due to the reverberant signal degradation.
Simulation scenario 2 with five participating sites

While the results of the simulation scenario 1 showed that the proposed squeezing approach can spatially disambiguate different remote sites as well as speakers within a remote site, the simulation scenario 2 aims to explore the squeezing approach in a more complicated environment with more remote meeting sites and with more participants at each remote site.

Similarly as the evaluation in scenario 1, Figure 5.14 shows the ground-truth speaker azimuths for all four sites. The legend in Figure 5.14 differentiates between remote sites with different plot point symbols whilst speakers at the same site are differentiated by color. This legend also applies to Figure 5.15, Figure 5.16 and Figure 5.17.

Figure 5.15 shows the speaker azimuths for all remote sites as estimated by SRP-PHAT, and similar to Figure 5.12 in scenario 1, it can clearly be seen that with more participants the spatial separation of speakers between sites is ambiguous.

Figure 5.16 thus shows the re-spatialised speaker azimuths as rendered by the proposed squeezing approach. The four remote sites are squeezed to:

- Site 1 (two people): top right quadrant
- Site 2 (two people): top left quadrant
- Site 3 (four people): bottom left quadrant
- Site 4 (three people): bottom right quadrant

The four quadrants of sites and speakers in Sites 2, 3, and 5 are clearly spatially separated, even with the greater variance in SRP-PHAT azimuth estimates at higher reverberation times. However, the four speakers of Site 4 in the bottom left quadrant
are more ambiguously placed, owing to the larger number of speakers squeezed into the equally-sized site sectors.

A second spatialisation result employing a different squeezing function is illustrated in Figure 5.17, where the squeezed sector sizes are adjusted according to the number of speakers per site. It is shown that allowing for smaller sectors for sites with fewer participants, e.g. Site 2 and Site 3, does not ambiguously reduce speaker spatial separation, whilst a site with more participants, e.g. Site 4, clearly benefits from greater spatial separation of its speakers.

5.5 $S^3AC$ Representation of Multiple Soundfields

5.5.1 System

In Section 4.2.3, it is discussed and evaluated that the $S^3AC$ spatial squeezing can be performed in a more intensive way than the standard 360°-surround to 60°-stereo squeezing. In particular, evaluation results presented in Section 4.2.3 show that an $S^3AC$ 360°-to-10° spatial squeezing process does not introduce significant distortion in perceptual source localisation. Based on this, it is suggested that a stereo down-mix can be utilized to save the localisation information of more than one surround soundfield.

Figure 5.18 illustrates an approach that the localisation information of two distinct surround soundfields is saved in one single stereo downmix by exploiting the $S^3AC$ 360°-to-10° spatial squeezing approach. In this approach, a standard $S^3AC$ virtual source estimation procedure is applied on each surround soundfield, resulting in frequency domain virtual sound source $S(k)$ and surround localisation information $\varphi_{360}(k)$. The standard $S^3AC$ squeezing step defined by:

$$\varphi_{dm}(k) = f(\varphi_{360}(k))$$

(5.22)
**Figure 5.11** Scenario 1: Original Speaker Azimuths

**Figure 5.12** Scenario 1: Estimated Speaker Azimuths from Two Remote Sites with Different Simulated Room Reverberation Time (Note: Legend from Figure 5.11 applies)
**Figure 5.13** Scenario 1: ‘Squeezed’ Speaker Azimuths from Two Remote Sites Rendered at Site 1 (Note: Legend from Figure 5.11 applies)

**Figure 5.14** Scenario 2: Original Speaker Azimuths (Note: microphones are hidden at circle center)
Figure 5.15 Scenario 2: Estimated Speaker Azimuths from Four Remote Sites with Different Simulated Room Reverberation Time (Note: Legend from Figure 5.14 applies)

Figure 5.16 Scenario 2: ‘Squeezed’ Speaker Azimuths from Four Remote Sites Rendered at Site 1 (Note: Legend from Figure 5.11 applies)
Figure 5.17 Scenario 2: Re-spatialised with Non-Uniform ‘Squeezed’ Sector Size Dependent on Number of Speaker for Each Remote Site

is then modified respectively for the two original soundfields such that the squeezing illustrated in Figure 5.18 is achieved, i.e. each original soundfield is squeezed into 10° and the two squeezed soundfields are saved in the $[30^\circ, 20^\circ]$ and $[-20^\circ, -30^\circ]$ regions in the squeezed soundfield, with a 40° empty region for discriminating purpose. During decoding, the two squeezed soundfields with a size of 10° are derived from the stereo downmix and recovered to two 360° surround soundfields by separately applying the inverse-squeezing approach.

For this approach, it is discovered that sound sources with overlapping time-frequency components, described in Section 4.3.4, can cause distortion in localisation and perceptual quality. Compared with the standard $S^3AC$ approach to compressing one spatial audio file, more overlapping time-frequency components can be introduced by the inter-soundfield interference, while two distinct files are compressed into one $S^3AC$ downmix. However, this scheme is proposed as an efficient approach for com-
pressing multiple surround speech scenes, especially for the multi-party teleconferencing application presented in Section 5.4. In a multi-site teleconference scenario, the distortion caused by overlapping time-frequency components is less significant as it only occurs when multiple participants speak concurrently. However, in a meeting scenario, it is more common that only one participant speaks at a time. When adopted with this approach, only a stereo down stream is required for each participating site, which contains information of multiple soundfields from remote sites.

5.5.2 Evaluation

The proposed S³AC two-soundfield-to-one-downmix approach is implemented on two 5.1-channel surround speech signals, simulating a teleconferencing scenario
where two 5.1 surround speech soundfields are encoded and transmitted to a third site. In this evaluation, soundfield 1 contains localised male speech and soundfield 2 contains localised female speech. The two multichannel signals have the same temporal duration, while the inter-soundfield overlapping time-frequency speech components exist but are not significant (approximately 25% of the overall signal duration of 20 seconds contain overlapping). These two signals are encoded with the proposed approach, resulting in a stereo downmix. This stereo downmix is then decoded and reproduced to two 5.1-channel signals for perceptual evaluation.

The MUSHRA methodology is employed, where the two decoded 5.1-channel speech signal are compared with the original surround speech signal. A hidden reference condition and a 3.5kHz non-localised low-pass-filtered anchor condition are added to each test file. Ten listeners took part in the experiment including both experienced and non-experienced listeners, while they were instructed to compare both the perceptual quality and sound source localisation precision between the reference and candidate conditions. The resulting average MUSHRA marks with 95% confidence intervals for these two coded speech files compared with the hidden reference and anchor are shown in Figure 5.19.

It is shown that the MUSHRA marks for the two coded conditions are in the region between 60 and 90, which refers to ‘good’ and ‘fair’ quality according to the MUSHRA recommendation. In the experiment, listeners claimed that, compared with the reference, there is occasional audible distortion in these two coded speech files. This is caused by the overlapping time-frequency components described above. However, listeners also claimed that the quality of these two speech files is suitable for a teleconference application; in particular, high localisation accuracy is preserved when compared with the reference file.

Note that, the S³AC based ‘squeezed’ re-rendering approach presented in Section 5.4 is not employed for this evaluation, as this evaluation only aims at evaluating the ef-
Figure 5.19 Listening test results for S\(^3\)AC two-soundfields-to-one downmix compression of multiple spatial speech scenes

...ficiency of the S\(^3\)AC two-soundfield-to-one-downmix approach. As the stereo downmix containing information from two soundfields can be further compressed by the existing MP3/AAC coders, this results in an approach that uses the same bandwidth as a conventional stereo MP3/AAC bit-stream for transmitting two multi-channel surround sound scenes. If this approach is adopted in the teleconferencing scenario described in Section 5.4, a bandwidth of only 128kbps is required for transmitting localised speech content of the two remote conference sites to each participating site. At each site, speech scenes can be flexibly spatialised to disambiguate different remote sites, as well as different speakers within one remote site in the teleconference.

5.6 Summary

This chapter has focused on introducing several extended applications based on S\(^3\)AC. An efficient approach to representation of Ambisonics B-format signals based on S\(^3\)AC sound source localisation analysis and soundfield squeezing has been pre-
sented, which also provides backward compatibility to Ambisonics B-format signals. While the S$^3$AC algorithm is based on amplitude panning and inverse panning analysis, derivation is given to show the fundamental equivalence between the amplitude panning based spatialisation method and Ambisonics reproduction. Based on this, the S$^3$AC approach to efficient and backward compatible compression of Ambisonics signals is described. Evaluation on this system is also performed, comparing with the conventional UHJ approach, with both objective and subjective results showing that the proposed S$^3$AC approach offers significant improvement, especially in preserving source localisation accuracy.

To provide extended backward compatibility of S$^3$AC encoding spatial audio signals, a binaural reproduction scheme exploiting HRTF techniques is presented. This system derives sound source and localisation information from any S$^3$AC encoded signal and reproduces a virtual surround soundfield by synthesizing binaural signals. It is shown that, since the source localisation is represented as direct azimuth information in S$^3$AC, it can be utilized directly for locating an HRTF pair according to source location without additional computational complexity. This approach is then perceptually evaluated, using both the original KEMAR HRTF dataset and an interpolated version. The results from subjective testing show that, when perceptually compared with the original multi-channel loudspeaker signals, the S$^3$AC binaural reproduced signals offer good perceptual quality and localisation accuracy at a bit rate as low as 70kbps.

The S$^3$AC spatial squeezing algorithm is then exploited in a multi-party teleconferencing application. The soundfield squeezing idea, while used during encoding in other S$^3$AC applications, is utilized to perceptually distinguish speech signals from different remote sites. In particular, soundfields containing localised speech signals from different remote participating meeting sites are squeezed to different ‘sectors’ for re-rendering at the local site. This provides local site users obvious perceptual discrimination for speech information from different remote sites, while relative lo-
localisation information for speakers within one remote site is also maintained. This system is then evaluated for two teleconferencing scenarios, with three and five participating sites respectively.

In addition, by exploiting earlier perceptual experiment results evaluating the $S^3$AC soundfield squeezing limitation, an approach for representing multiple soundfields using one $S^3$AC stereo downmix is introduced. This is achieved by performing a more intensive $S^3$AC soundfield squeezing, so that a stereo downmix soundfield can be used to save multiple surround soundfields. For example, a system is detailed and evaluated, which compresses two surround soundfields with localised speech content into one $S^3$AC stereo downmix. The evaluation results show that accurate localisation is preserved and this approach is suggested for efficient low bit rate representation of localised speech signals for teleconferencing applications.

So far, the propose $S^3$AC scheme has been focused on coding 2D spatial audio signals and related applications. In the next chapter, the $S^3$AC scheme is extended for efficient compression of 3D multichannel spatial audio.
Chapter 6

Compressing the Multichannel Three Dimensional Audio

6.1 Introduction

In this chapter, a novel approach to efficient compression of a multi-channel 3D spatial audio signal is presented. Compared with the existing Ambisonics 3D audio techniques, this $S^3$AC based approach is focused on providing efficient bandwidth reduction for a multi-channel loudspeaker signal rendering a 3D soundfield, without using a microphone recording approach. This overcomes the distortion introduced by non-ideal recording hardware as well as other shortcomings of Ambisonics. For example, a professional artist produces a 3D multi-channel signal in his studio using e.g. 3D panning techniques and would like to share it over the Internet with both professional users (with 3D playback) and normal users (with surround 2D or stereo playback); recording the 3D soundfield using Ambisonics microphone introduces equipment cost, challenging calibration, environmental/hardware noise as well as undesired room reverberation, not to mention that an Ambisonics signal still requires at least 4 channels (first-order) of full bandwidth with limited stereo/mono backward compatibility. The approach addressed in this Chapter is proposed as an efficient solution for this scenario.
In this approach, based on the fundamental ideas of 2D $S^3$AC addressed in previous chapters, sound source localisation is derived from a multi-channel 3D spatial audio signal on a time-frequency basis. By extending the derivation results presented in Section 4.4, a novel algorithm based on 3D Cartesian orthogonal analysis is introduced for the estimation of source localisation from an arbitrary loudspeaker setup. The derived localisation information consists of azimuth and elevation parameter and a novel Spatial Localisation Quantisation Point (SLQP) method is introduced for quantising a source location in a 3D soundfield. This method effectively forms an index in a 3D localisation codebook for each sound source, which can be either efficiently saved as accompanying side information for a mono downmix of the original 3D multi-channel signal, or represented by squeezing into a unique location in an $S^3$AC stereo downmix. As a result, a multi-channel 3D spatial audio signal is compressed into a backward compatible stereo/mono signal.

This chapter is formatted as follows: Section 6.2 extends the $S^3$AC 2D orthogonal analysis given in Section 4.4 to a 3D analysis algorithm for estimating source localisation from an arbitrary 3D loudspeaker playback. Section 6.3 details the approach to the efficient compression and representation of multichannel 3D soundfield using $S^3$AC, including 3D source localisation estimation, 3D source localisation quantisation, SLQP, and the method to represent a 3D soundfield by either a mono downmix + SLQP or a spatially squeezed stereo downmix based on the $S^3$AC spatial squeezing approach. Section 6.4 presents both objective and subjective evaluation results of the proposed technique, where some 16-channel 3D spatial audio signals are used for the evaluation.
6.2 3D orthogonal analysis for virtual sound source localisation

For efficient representation of a multichannel 3D audio signal, an efficient and flexible 3D source localisation estimation algorithm is essential. This section extends to 3D the orthogonal analysis of loudspeakers positioned in 2D. While tangent amplitude panning law is only used for 2D audio panning, and VBAP for 3D audio is only derived for three loudspeakers (as described in 3D VBAP [89] and Section 2.5.3), neither of the methods can be generalized for estimating 3D spatial sound sources rendered by more than three loudspeakers arbitrarily located in three dimensions. In order to develop an efficient sound source estimation algorithm for 3D spatial audio, an assumption is made that, based on the equivalency between 2D orthogonal analysis and the amplitude panning algorithm, described in Section 4.4.2, three dimensional orthogonal analysis of 3D audio for 3D sound source and localisation estimation can be derived as a universal algorithm for any type of loudspeaker array. The efficiency of this approach remains consistent to other amplitude panning approaches. The derivation of the 3D orthogonal sound source estimation algorithm is given in the remainder of this section, while an application of this algorithm to compressing 3D multi-channel spatial audio signals is presented and the performance of such system is evaluated in the following sections.

Considering a loudspeaker positioned at azimuth $\mu_i$, elevation $\eta_i$, as shown in Figure 6.1, a vector representing the loudspeaker signal $p_i$, where $i$ is the loudspeaker index, can be decomposed into x-y-z coordinate components as:

$$p_i = g_i \cdot \begin{bmatrix} \cos \mu_i \cdot \cos \eta_i \\ \sin \mu_i \cdot \cos \eta_i \\ \sin \eta_i \end{bmatrix}$$  \hspace{1cm} (6.1)

where $g_i$ represents the gain of this loudspeaker. In a circumstance that the number
of loudspeakers, $N$, are used for rendering a 3D spatial sound source, the resulting source level $g$ can be calculated by the orthogonally decomposed components over the three axes, such that

$$g^2 = \left[ \sum_{i=1}^{N} g_i \cdot \cos \mu_i \cdot \cos \eta_i \right]^2 + \left[ \sum_{i=1}^{N} g_i \cdot \sin \mu_i \cdot \cos \eta_i \right]^2 + \left[ \sum_{i=1}^{N} g_i \cdot \sin \eta_i \right]^2$$  \hspace{1cm} (6.2)

In order to avoid front/back, left/right energy cancellation, which does not happen in the real case, Eq. 6.2 is re-written as:
\[ g^2 = \left( \sum_{i=1}^{N} g_i \cdot \left| \cos \mu_i \cdot \cos \eta_i \right| \right)^2 + \left( \sum_{i=1}^{N} g_i \cdot \left| \sin \mu_i \cdot \cos \eta_i \right| \right)^2 + \left( \sum_{i=1}^{N} g_i \cdot \left| \sin \eta_i \right| \right)^2 \]  

(6.3)

And the direction of the resulting source, at azimuth \( \mu \) elevation \( \eta \), is estimated by:

\[ \tan \mu = \frac{\sum_{i=1}^{N} g_i \cdot \cos \mu_i \cdot \cos \eta_i}{\sum_{i=1}^{N} g_i \cdot \sin \mu_i \cdot \cos \eta_i} \]  

(6.4)

\[ \tan \eta = \sqrt{\left( \sum_{i=1}^{N} g_i \cdot \cos \mu_i \cdot \cos \eta_i \right)^2 + \left( \sum_{i=1}^{N} g_i \cdot \sin \mu_i \cdot \cos \eta_i \right)^2} \]  

\[ \sum_{i=1}^{N} g_i \cdot \sin \eta_i \]  

(6.5)

### 6.3 S³AC Compression of 3D Multi-Channel Spatial Audio

Based on the derivation for estimating 3D spatial sound source localisation in the previous section, this section presents a methodology for the analysis and compression of 3D multi-channel audio signals, based on the extension of the S³AC spatial squeezing approaches (see Section 3.2.1) and psychoacoustic-based cue quantisation (see Section 4.3.3) to 3D soundfields.
6.3.1 System Overview

The proposed compression scheme is aimed at low bit rate and backward compatible transmission of 3D soundfields rendered by an arbitrary number of loudspeakers, where minimum sound source localisation distortion is desired. In other words, the 3D sound scene and source localisation content recovered at the decoding site is required to have minimum distortion when compared to the original rendering, even if a different hardware setup (loudspeaker array) to the original is used at the decoding site. The Ambisonics encoding approach for representing 3D soundfield content is not desired here, as it requires a spatialised microphone array, which may lead to recording distortion, and its reproduction localisation accuracy is sensitive to loudspeaker setup. The Ambisonics approach is also limited by not being backward compatible with conventional mono/stereo system.

In the proposed encoding system, as shown in Figure 6.2, a frequency domain orthogonal analysis as presented in Section 6.2 is utilized for estimating sound sources and their location rendered by a 3D multi-channel spatial audio signal (see Section 6.2 for a detailed description). This is followed by quantisation of the derived 3D spatial localisation, which exploits perceptual localisation redundancy for bit-rate efficiency as well as a further $S^3AC$ spatial squeezing analysis. The resulting quantised ‘virtual’ 3D sound sources can be saved in an $S^3AC$ stereo downmix, with the stereo downmix soundfield intelligently designed to have a unique mapping between Spatial Localisation Quantisation Points (SLQPs) in the original 3D soundfield and the downmixed localisation points in the downmix stereo soundfield. The quantised virtual 3D sound sources can also be saved as a mono downmix while the SLQP information is saved as side information, which can be further quantised to reduce the bit rate. Based on this encoding approach, an 3D $S^3AC$ decoder, as illustrated in Figure 6.3, can derive a sound source with its direct localisation information in a 3D soundfield from either a 3D $S^3AC$ stereo downmix or a mono downmix with accompanying 3D local-
Compressing the Multichannel Three Dimensional Audio

Figure 6.2 $S^3$AC 3D Encoding System
Figure 6.3 $S^3$AC 3D Decoding System
Compressing the Multichannel Three Dimensional Audio

itemisation side information. This ‘source + 3D localisation’ format provides flexibility in that any major 3D reproduction method can be applied, e.g. VBAP, any order of Ambisonics designed for a particular loudspeaker setup, etc. Hence, a user can choose the same method as used in producing the original multichannel signals ensuring minimum distortion, or the most appropriate rendering solution that meets the application’s requirement.

6.3.2 Time-Frequency Orthogonal Analysis for 3D Source Azimuth-Elevation Information

The proposed method starts with a time-frequency decomposition applied separately to each of the $N$ channels of the input spatial audio signal. Any modern time-frequency decomposition can be used such as a short-time Fourier transform (STFT) (addressed in Section 2.2.3) or a pseudo quadrature-mirror filterbank (PQMF) with further perceptual bank decomposition (as used in MP3/AAC [1] [2]). This resulting time-frequency representation for the signal from the $i^{th}$ loudspeaker is then orthogonally decomposed based on Eq. 6.1, resulting in:

$$ p_i(k, n) = g_i(k, n) \cdot \begin{bmatrix} \cos \mu_i \cdot \cos \eta_i \\ \sin \mu_i \cdot \cos \eta_i \\ \sin \eta_i \end{bmatrix} \quad (6.6) $$

where $k$ and $n$ are frequency and temporal frame indices respectively. Based on this and Eq. 6.3, the overall source level $g_s(k, n)$ rendered by all the loudspeakers is given by:
\[ g^2_s(k, n) = \left[ \sum_{i=1}^{N} g_i(k, n) \cdot |\cos \mu_i \cdot \cos \eta_i| \right]^2 + \left[ \sum_{i=1}^{N} g_i(k, n) \cdot |\sin \mu_i \cdot \cos \eta_i| \right]^2 + \left[ \sum_{i=1}^{N} g_i(k, n) \cdot |\sin \eta_i| \right]^2 \]  \hspace{1cm} (6.7)

and the source signal can be generated by applying the phase information \( e^{\phi_M} \) chosen from the channel with the highest amplitude (defined as the \( M^{th} \) channel) in order to maintain phase consistency:

\[ S(k, n) = \sqrt{g^2_s(k, n)} \cdot e^{\phi_M} \] \hspace{1cm} (6.8)

while the source azimuth \( \mu_s(k, n) \) and elevation \( \eta_s(k, n) \) can be derived using Eq. 6.4 and Eq. 6.5 as:

\[ \tan \mu_s(k, n) = \frac{\sum_{i=1}^{N} g_i(k, n) \cdot \cos \mu_i \cdot \cos \eta_i}{\sum_{i=1}^{N} g_i(k, n) \cdot \sin \mu_i \cdot \cos \eta_i} \] \hspace{1cm} (6.9)

\[ \tan \eta_s(k, n) = \sqrt{\left[ \sum_{i=1}^{N} g_i(k, n) \cdot \cos \mu_i \cdot \cos \eta_i \right]^2 + \left[ \sum_{i=1}^{N} g_i(k, n) \cdot \sin \mu_i \cdot \cos \eta_i \right]^2} \bigg/ \sum_{i=1}^{N} g_i(k, n) \cdot \sin \eta_i \] \hspace{1cm} (6.10)

### 6.3.3 Spatial Localisation Quantisation Points

The 3D orthogonal analysis algorithm presented in the Section 6.3.2 derives the azimuth and elevation localisation information of the 3D sound source rendered by
a number of loudspeakers in a 3D soundfield. The precision of the derived azimuth/elevation is based on the input signal bit-precision, e.g., 16 bits. This can be reduced for higher bit rate efficiency, similar to the cue quantisation approach used in 2D S3AC coding described in Section 4.3.3. Psychoacoustics show that, in the most sensitive listening area, the frontal plane, the human auditory system has approximately a 1° azimuth and 5° to 10° elevation resolution, respectively, for localizing tonal sources with frequency components most sensitive to the human ear [36]. This phenomenon, also referred as localisation blur [36], is exploited in 2D S3AC coding to achieve bit rate efficiency. During the S3AC compression of 3D soundfield, both perceptual azimuth resolution and perceptual elevation resolution is exploited, so as to effectively describe a continuous 3D sphere by discrete numbers of localisation points, called Spatial Localisation Quantisation Points (SLQPs), which is designed to ensure minimum loss in perceptual localisation. Each SLQP consists of localisation information described as source azimuth/elevation, for the considerations of efficiency in post-processing and quantisation, as well as the advantages of describing the source localisation information directly as angular information. Two example SLQP designs are described in the following, which have comparatively higher and lower quantisation precision, as shown in Figure 6.4 and Figure 6.5. Based on the available experimental facilities, both examples are designed for an upper-hemisphere 3D sound scene, and used for evaluation in Section 6.4. The two designs are described as follows:

- Based on perceptual experiments on quantisation of 2D S3AC side information described in Section 4.3.3, an azimuth precision of 2° (3° for low precision) is used for the 0° elevation plane.

- Compared with the location dependent azimuth quantisation precision of S3AC side information described in Section 4.3.3, a uniform azimuth quantisation precision is used here, since in 3D surround sound scenarios, listeners are commonly guided to twist heads or turn bodies to fully exploit the impression of a
**Figure 6.4** $S^3$AC SLQP with High Precision

**Figure 6.5** $S^3$AC SLQP with Low Precision
3D surround audio scene.

- Based on the elevation resolution suggested by psychoacoustics, a 5° (10° for low precision) elevation resolution is utilized. This effectively results in SLQPs with parallel layers of different elevations.

- As the perceptual localisation precision degrades when the elevation of a source increases away from the 0° elevation horizontal plane, the azimuth precision used for quantisation is decreased with increasing elevation. Specifically, in the higher precision example, the azimuth quantisation precision used for the adjacent higher layer is reduced by 10 quantisation points, e.g. while the 0° elevation layer has 180 SQLPs, the 5° elevation layer has 170 SLQPs. In the lower precision example, the quantisation precision reduction between layers is 12 points.

- A 90° elevation SLQP is reserved for both designs and it’s the only SLQP on the 90° elevation layer.

- The resulting numbers of SLQPs are 1729 and 658, for the high precision design and low precision design, respectively. Hence, it requires 10.7 bits and 9.4 bits to encode each derived SLQP, for the high precision and low precision design respectively. Note that, these two SLQP examples are designed based on a linearly decreasing resolution with increasing elevation, for evaluating the propose 3D audio coding approach. In practical application, SLQP design with ‘power of 2’ numbers of SLQPs, e.g. 1024 SLQPs, can be adopted for better bit-rate efficiency.

By exploiting available psychophysical theory and experimental results, the goal of the two SLQP designs described above is to ensure minimum perceptual localisation distortion while efficient bandwidth reduction can be achieved. The performance of these designs will be evaluated and justified, both objectively and subjectively, in Section 6.4.
6.3.4 Stereo downmixing and S$^3$AC spatial squeezing for 3D soundfield

Similarly as the approach to re-pan the virtual source to a unique position in the stereo downmix described in 2D S$^3$AC compression, each SLQP derived in Section 6.3.3 is given a unique mapping into a 60° stereo downmix soundfield. A simple design can be made such that the 60° downmix soundfield is uniformly divided into discrete azimuths according to the number of SLQPs in the 3D soundfield. For instance, to save the high SLQP design described in Section 6.3.3 and Figure 6.4, which has 1729 points, two adjacent downmix localisation points then have approximately a 0.035° azimuth discrimination. This can be expressed mathematically by:

$$\varphi_{dm}(k, n) = f(SLQP(k, n))$$  \hspace{1cm} (6.11)$$

where $\varphi_{dm}(k, n)$ is the assigned azimuth in the 60° downmix soundfield and $f(\cdot)$ represents the 3D SLQP to 2D soundfield mapping methodology, which can be defined by the user. An illustrative example of this 3D-to-2D localisation mapping is given in Figure 6.6 and Figure 6.7. Figure 6.6 shows the low-precision SLQP design as described in Section 6.3.3, while each SLQP is distinguished by color. Figure 6.7 illustrates an example localisation mapping to generate a stereo downmix where each SLQP in 3D soundfield is given a unique azimuth in the 2D 60° soundfield, while Figure 6.8 gives a zoom-in view of the left-hand side downmix area in Figure 6.7.

Similarly as in 2D S$^3$AC coding, the stereo downmix is generated by amplitude panning the derived virtual source from Eq. 6.8 to the location in the stereo downmix derived from Eq. 6.11:
Figure 6.6 $S^3$AC SLQP with Low Precision while each SLQP is distinguished by color

Figure 6.7 $S^3$AC SLQP in a stereo downmix
Compressing the Multichannel Three Dimensional Audio

Figure 6.8 $S^3$AC SLQP in a stereo downmix, zoom-in view of the left-hand side of Figure 6.7

\[
L_{dm}(k,n) = \frac{S(k,n) \cdot [\tan 30^\circ + \tan(\varphi_{dm}(k,n))]}{\sqrt{2 \cdot \tan^2 30^\circ + 2 \cdot \tan^2(\varphi_{dm}(k,n))}}
\]

\[
R_{dm}(k,n) = \frac{S(k,n) \cdot [\tan 30^\circ - \tan(\varphi_{dm}(k,n))]}{\sqrt{2 \cdot \tan^2 30^\circ + 2 \cdot \tan^2(\varphi_{dm}(k,n))}}
\]

This is followed by an inverse time-frequency transform for a backward compatible time domain stereo signal representation.

According to the derivation given in Section 4.2, the azimuth resolution in the $S^3$AC downmix soundfield is defined by the source amplitude. Equivalently, the number of derivable azimuths in the $S^3$AC downmix is:

\[
\| g_s(k,n) \| + 1
\]

where $\| \cdot \|$ stands for rounding to the nearest integer and $g_s(k,n)$ is the amplitude of the derived virtual source, which is given by Eq. 6.7. Although a derived virtual source amplitude higher than 1729 (approximately -25.5dB for a signal with 16-bit
quantisation level and defining the level of $2^{10}$ as 0dB) is required for maintaining adequate downmix azimuth resolution, by building the localisation mapping from 3D to stereo downmix such that adjacent SLQPs in 3D are given adjacent downmix azimuths, it is ensured that, for sources having amplitudes lower than the required amplitude, the recovered localisation in 3D has minimum deviation from the original position. This is further evaluated and justified in Section 6.4.

Based on this, a stereo downmix containing squeezed localisation information of a 3D surround sound field can be further compressed by conventional perceptual audio coders, such as AAC, which results in equivalent bit rates as conventional stereo audio for transmitting a 3D spatial soundfield, e.g. 128kbps, while backward compatibility is also maintained.

### 6.3.5 Mono downmixing and further processing on SLQP

The SLQP derived in Section 6.3.3 can also be saved as accompanying 3D localisation side information for a mono downmix generated by the virtual sound source given by Eq. 6.7. According to the evaluation presented in Section 4.3.3, a frame-wise localisation differential coding based on a code-book representation of the derived source localisation can efficiently reduce the bit rate of spatial side information without introducing any distortion. In this work, similarly to the algorithm described in Section 4.3.3, the SLQP set representing a 3D soundfield is transformed into a codebook representation, where each SLQP has a unique index in the codebook as:

\[ I(k, n) \] (6.14)

The codebook distance between two adjacent frames of a frequency is calculated as:

\[ E(k, n) = I(k, n) - I(k, n - 1) \] (6.15)
To ensure tolerance to transmission errors, the code book index in Eq. 6.14 is recorded every number of frames, similar to the approach described in Section 4.3.3, based on application scenarios, while the following codebook distance information is entropy coded, e.g. using Rice Coding [95], for bit rate efficiency. While this differential coding approach applied in 2D S³AC has been investigated in Section 4.3.3, it is not the objective of this Chapter to further evaluate this approach.

### 6.3.6 Decoding and reproduction

For a received S³AC 3D stereo downmix, following a time-frequency transform, the decoder performs a virtual source localisation in the 60° stereo downmix soundfield by inverting Eq. 6.12, where a virtual source \( \hat{S}(k, n) \) and its azimuth \( \hat{\phi}_{dm}(k, n) \) in the 60° downmix soundfield is recovered as:

\[
\hat{S}(k, n) = \sqrt{L_{dm}^2(k, n) + R_{dm}^2(k, n)} \cdot e^{\phi_{dm}}
\]

\[
\hat{\phi}_{dm}(k, n) = \tan^{-1}\left[\frac{|L_{dm}(k, n)| - |R_{dm}(k, n)|}{|L_{dm}(k, n)| + |R_{dm}(k, n)|} \cdot \tan 30°\right]
\]

where the phase parameter \( e^{\phi_{dm}} \) is chosen as the phase information from the channel with the highest energy of the two stereo channels.

This is followed by inverting Eq. 6.11 for re-mapping the localisation to an SLQP in the 3D soundfield, such that:

\[
SLQP(k, n) = f^{-1}(\hat{\phi}_{dm}(k, n))
\]

In the S³AC 3D mono downmix mode, the compressed 3D side information is decoded to recover the SLQP in the 3D soundfield and the mono downmix is decomposed into a frequency domain representation.
Compressing the Multichannel Three Dimensional Audio

This process effectively results in a sound source and accompanying decoded SLQP containing localisation azimuth/elevation information, which has advantages in flexibility of choosing a reproduction method based on playback scenarios or other requirements. Based on the available reproduction facility (see Section 6.4) with a symmetrical loudspeaker array, higher order Ambisonics reproduction is used for 3D soundfield decoding and reproduction. As described in Section 2.6, to reproduce a source signal using \( m^{th} \) order Ambisonics in an array with \( m^2 \) loudspeakers, the source signal is firstly encoded as:

\[
B(k,n) = s(k,n) \cdot y(\mu(k,n), \eta(k,n)) \tag{6.19}
\]

where \( s(k,n), \mu(k,n) \) and \( \eta(k,n) \) are the time-frequency representation of the source, and its azimuth/elevation in the 3D soundfield, respectively, while \( y(\mu(k,n), \eta(k,n)) \) is defined as a vector containing spherical harmonics functions as:

\[
y(\mu(k,n), \eta(k,n)) = [Y_1(\mu(k,n), \eta(k,n)), Y_2(\mu(k,n), \eta(k,n)), \ldots, Y_{(m+1)^2}(\mu(k,n), \eta(k,n))] \tag{6.20}
\]

where \( Y_n(\mu(k,n), \eta(k,n)) \) is defined in Appendix A for up to \( 4^{th} \) order Ambisonics.

The encoded source signal is transformed into a loudspeaker signal matrix \( LS(k,n) \), defined as:

\[
LS(k,n) = D \cdot B(k,n) \tag{6.21}
\]

where \( D \) is the pseudo-inverse of the re-encoding matrix \( C \) such that:

\[
D = \text{pinv}(C) = C^T \cdot (C \cdot C^T)^{-1} \tag{6.22}
\]
The re-encoding matrix \( C \) is defined based on the configuration of the reproduction loudspeakers system as:

\[
C = [c_1, c_2, \ldots, c_i, \ldots, c_N]
\]

(6.23)

where \( i \) is the speaker index and \( N \) the total number of loudspeakers in the reproduction system, vector \( c_i \) defined as the series of spherical harmonics functions \( Y_n(\mu, \eta) \) in Appendix A based on the azimuth \( \mu_i \) and elevation \( \eta_i \) of the \( i^{th} \) loudspeaker, such that:

\[
c_i = [Y_1(\mu_i, \eta_i), Y_2(\mu_i, \eta_i), \ldots, Y_{(m+1)2}(\mu_i, \eta_i)]
\]

(6.24)

Note that, compared with the sound source azimuth \( \mu(k, n) \) and elevation \( \eta(k, n) \) derived on a time-frequency basis, the azimuth \( \mu_i \) and elevation \( \eta_i \) for the \( i^{th} \) loudspeaker is fixed as long as the loudspeaker configuration remains unchanged.

### 6.4 Evaluations

The proposed S\(^3\)AC 3D spatial audio compression technique, including both the stereo and mono downmix modes, is evaluated both objectively and subjectively in this section. For this purpose, the algorithms and methodologies presented in this chapter are implemented based on a 16-channel hemisphere loudspeaker array designed by G. Potard etc. [83] [109] [110], called the Configurable Hemisphere Environment for Surround Sound (CHESS). This system is rebuilt in an anechoic chamber for the purpose of reverberation-free perceptual listening experiments, while the loudspeaker positioning setup remained unchanged. Figure 6.9 is the photo of the current CHESS system and the loudspeaker positioning configuration is described in Table 6.1. A detailed description of the loudspeaker location setting in CHESS can
be found in [83] [109].

### 6.4.1 16-Channel 3D audio files for evaluation

For evaluation of the proposed $S^3$AC compression of 3D audio, based on the 16-channel CHESS loudspeaker system, eight 16-channel 3D audio signals with different audio content were created based on multiple 3D audio rendering methods. These files are described in the following:

- **File 1.** A clear male speech signal presenting the loudspeaker channel number (from 1 to 16) panned to only that channel. The duration is approximately 22 seconds.

- **File 2.** An airplane moving over-head from the rear right (at -144° azimuth) to the front left (at 36° azimuth). The duration is approximately 17 seconds.

- **File 3.** Background noise recorded at a busy restaurant, which is equally panned to all channels, rendered with focusing male speech localised at multiple locations. Duration is approximately 25 seconds. (Noted that the noise recording used in File 3, 4 and 5 are not identical.)

- **File 4.** Background noise recorded at a busy restaurant, equally panned to all channels without other content. Duration is approximately 30 seconds.

- **File 5.** Background noise recorded at a busy restaurant, reproduced to 16 channels with a moving localisation feature. The result of this signal is perceived as the source is moving around the listener. Duration is approximately 32 seconds.

- **File 6.** A 4th order 16-channel reproduction of an Ambisonics B-format recording, featuring music and localised sound sources. Duration is approximately 20 seconds.
Figure 6.9 16-Channel loudspeaker array CHESS in an anechoic environment

<table>
<thead>
<tr>
<th>Channel Number</th>
<th>Azimuth</th>
<th>Elevation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>144</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>216</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>288</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>108</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>180</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>252</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>324</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>72</td>
<td>60</td>
</tr>
<tr>
<td>13</td>
<td>144</td>
<td>60</td>
</tr>
<tr>
<td>14</td>
<td>216</td>
<td>60</td>
</tr>
<tr>
<td>15</td>
<td>288</td>
<td>60</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 6.1 CHESS loudspeaker configuration
• **File** 7. A 4\textsuperscript{th} order 16-channel reproduction of an Ambisonics B-format recording, featuring localised percussion instrument. Duration is approximately 18 seconds.

• **File** 8. A 4\textsuperscript{th} order 16-channel reproduction of an Ambisonics B-format recording, featuring music and localised sound sources. Duration is approximately 21 seconds.

These files are used for both objective and subjective evaluation presented in the following sections.

### 6.4.2 Objective evaluation

File 2 described in Section 6.4.1 is firstly used for evaluation of localisation accuracy after S\textsuperscript{3}AC 3D coding. The signal utilizes a mono signal with synthesized azimuth/elevation localisation parameters, as illustrated in Figure 6.10. It is reproduced into a 16-channel signal by controlling the gain parameters between each channel to ensure correct 3D rendering of an aero plane flying over-head movement from rear right (-144° azimuth) to front left (36° azimuth). Online tuning of the reproduction parameters was performed to ensure the designed source movement feature is perceptually achieved. This signal is processed using the 3D source localisation orthogonal analysis presented in Section 6.3 based on a 1024-point 50% overlapping STFT transform, to derive the source localisation in the 3D soundfield. Figure 6.11 illustrates the derived source elevation feature on a time-frequency basis. Furthermore, Figure 6.12 shows the derived elevation of a single frequency, the 80\textsuperscript{th} frequency bin, which contains the highest amplitude information of the whole signal. Figure 6.13 plots the derived azimuth/elevation localisation on a 3D sphere of the 80\textsuperscript{th} frequency. Note that, as the reproduction method used for creating this signal utilizes multiple loudspeakers located at variable elevations in all time frames, the resulting source localisation has elevation features higher than 0° and below 90° (approxi-
Compressing the Multichannel Three Dimensional Audio

mately between 10° and 60°), when compared to the designed movement in Figure 6.10. However, the intended key localisation feature, which simulates the aero plane moving over-head from the rear right (-144° azimuth) to the front left (36° azimuth), is perceptually achieved.

The derived source azimuth/elevation localisation information is spatially quantised using the SLQP approach described in Section 6.3, followed by both the stereo downmixing approach presented in Section 6.3.4 and the mono downmix approach presented in Section 6.3.5.

The resulting quantised elevation using high-precision SLQP design is illustrated in Figure 6.14, while the result of using low-precision SLQP design is illustrated in Figure 6.15. In addition, based on the methodology presented in Section 6.3.6, Figure 6.16 gives the elevation feature derived from the stereo downmix, which is synthesized by mapping the high-precision SLQP to the 60° stereo soundfield.

By comparing Figure 6.14 to 6.16 with Figure 6.11, the localisation and source movements in the original signal are estimated and recovered in the S3AC 3D encoding/decoding process. Quantisation distortion is introduced due to the SLQP quantisation process, while the high-precision SLQP designed introduces less quantisation distortion compared with the other two coding approaches.

Further analysis is performed on mathematically evaluating the localisation error caused by the S3AC 3D coding. Three different coding modes presented above, including stereo downmixing, mono downmixing with high precision SLQP design and mono downmixing with low precision SLQP design, are evaluated. For each of the original 16-channel 3D audio signals described in Section 6.4.1, the azimuth/elevation localisation information is derived by using the proposed 3D orthogonal localisation analysis on a per-frame-per-frequency basis as the original localisation condition. The error is calculated as the average absolute angular difference between the original condition and the azimuth/elevation localisation derived at the
Figure 6.10 Designed source movement simulating airplane fly over-head effect

Figure 6.11 Time-frequency-elevation mesh of the original signal
Figure 6.12 Elevation feature of the 80th frequency

Figure 6.13 Azimuth/elevation 3D localisation feature of the 80th frequency
Figure 6.14 Time-frequency-elevation mesh of the signal encoded by $S^3$AC 3D mono down-mix with High-precision SLQP quantisation

Figure 6.15 Time-frequency-elevation mesh of the signal encoded by $S^3$AC 3D mono down-mix with Low-precision SLQP quantisation
Compressing the Multichannel Three Dimensional Audio decoder, which can be described as:

\[
\mu_{\text{error}} = \frac{1}{KM} \sum_{k=1}^{K} \sum_{n=1}^{M} |\mu_{\text{original}}(k, n) - \mu_{\text{decoded}}(k, n)|
\]

\[
\eta_{\text{error}} = \frac{1}{KM} \sum_{k=1}^{K} \sum_{n=1}^{M} |\eta_{\text{original}}(k, n) - \eta_{\text{decoded}}(k, n)|
\]  

(6.25)

where \( K, M \) are the total number of frequency bins and frames respectively. In this calculation, signal samples with energy levels lower than -60dB of the maximum energy, which is approximately lower than 32 for a 16-bit quantised signal, are ignored for calculation, in order to avoid random localisation errors caused by recording noise. The three proposed coding modes, including stereo downmix, mono downmix with high-precision SLQP and mono downmix with low-precision SLQP, are evaluated. The resulting average azimuth/elevation errors for each test file are given in Table 6.2 to 6.4. While the resulting errors are file dependent, it is shown that the stereo coding mode and high-precision SLQP mono coding mode have similar error performance. The low-precision mono coding mode introduces higher error. Average errors of less than 5° is found in most files for both azimuth and elevation, while some increased to higher than 10°. The resulting perceptual impact will be evaluated in the next Section.

6.4.3 Subjective evaluation

The proposed \( S^3 \)AC multi-channel 3D spatial audio compression system is further evaluated by perceptual experiments. The eight 16-channel 3D audio files described in Section 6.4.1 are used. Three proposed types of \( S^3 \)AC 3D multi-channel audio compression approaches were evaluated, including stereo downmix, mono downmix with high-precision SLQP, mono downmix with low-precision. The coding process is based on a 1024-point 50% overlapping STFT.

In the stereo downmix approach, azimuth/elevation localisation information is de-
Figure 6.16 Time-frequency-elevation mesh of the signal encoded by S³AC 3D stereo downmix, SLQP recovered from stereo downmix soundfield

<table>
<thead>
<tr>
<th>File</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azimuth Error</td>
<td>11.84</td>
<td>0.08</td>
<td>15.10</td>
<td>0.16</td>
<td>2.70</td>
<td>7.35</td>
<td>1.19</td>
<td>4.60</td>
</tr>
<tr>
<td>Elevation Error</td>
<td>1.19</td>
<td>0.73</td>
<td>1.30</td>
<td>1.90</td>
<td>3.55</td>
<td>1.26</td>
<td>0.42</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Table 6.2
Average azimuth/elevation error (in degrees) from S³AC 3D mono coding with high-precision SLQP design

<table>
<thead>
<tr>
<th>File</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azimuth Error</td>
<td>12.17</td>
<td>0.70</td>
<td>15.37</td>
<td>0.22</td>
<td>5.36</td>
<td>7.64</td>
<td>1.29</td>
<td>4.89</td>
</tr>
<tr>
<td>Elevation Error</td>
<td>2.27</td>
<td>1.45</td>
<td>2.87</td>
<td>3.88</td>
<td>7.01</td>
<td>2.53</td>
<td>0.82</td>
<td>2.36</td>
</tr>
</tbody>
</table>

Table 6.3
Average azimuth/elevation error (in degrees) from S³AC 3D mono coding with low-precision SLQP design
Compressing the Multichannel Three Dimensional Audio

rived for every frequency for mapping into the 60° downmix soundfield. In the mono
downmix approach, SLQP is also derived for every frequency and further quantised
using either the high precision or low precision design. The high precision design re-
sults in a bit-rate of 474 kbps for the 3D side information, while the low precision de-
sign results in a bit-rate of 413 kbps. This process is used for maximizing the coding
performance in the mono downmix mode for evaluation purposes. However, by fur-
ther exploiting band perception properties of human auditory system, e.g. grouping
the FFT coefficients into 20-band double ERBs (equivalent rectangular bandwidths
[31]), the bandwidth required for transmitting 3D side information can be signifi-
cantly reduced. For example, this results in bit rates of 18kbps and 16kbps, for the
high precision and low precision SLQP design, if one quantised SLQP is used for
each double-ERB band. The downmix signals in all three modes are further coded
by AAC with 64kbps/channel. The 4th order Ambisonics reproduction method pre-

sented in Section 6.3.6 is used for reproducing the signal to 16-channel. Based on
the 16-channel CHESS loudspeaker configuration, the 4th order Ambisonics can en-
sure the highest localisation reproduction precision by fully utilizing the available
16 loudspeakers. A perceptual evaluation methodology based on MUSHRA [77]
was utilized for the listening test. Besides the three proposed S3AC 3D coding con-
ditions, an AAC coding condition is incorporated for comparison purposes, where
each channel in the original signal is coded individually with 64kbps AAC, resulting
in a total bit rate of 1024kbps. All coded conditions are randomly mixed with a hid-
den reference and an anchor signal, which is created by a 3.5kHz low-pass filtered
version of the original signal, followed by mono-mixing to all channels to remove
localisation. A total of 17 listeners took part in the experiment, including 6 expe-
rienced listeners. The results are shown in Figure 6.17 with the average across all
listeners and 95% confidence intervals plotted.

The results show that, in the 3D surround sound scenes where little or no source lo-
calisation feature is reproduced, the three S3AC 3D compression modes reveal little
or no perceptual difference between the original and AAC condition, such as File 1 and File 4. In other 3D surround sound scenarios with a higher mixture of sources or more source localisation features, the performance of the three $S^3$AC 3D compression modes degrades. However, for all scenarios, the proposed $S^3$AC 3D compression algorithm achieves grades above the MUSHRA ‘Good’ grade, while most of the results of the proposed conditions lies on the boundary between the MUSHRA ‘Excellent’ grade and ‘Good’ grade. Considering the bandwidth reduction from 16-channel to 2-channel (stereo downmix) or less (mono downmix + 3D side information), the advantage of the proposed $S^3$AC 3D multi-channel audio compression technique can clearly be seen.

6.5 Summary

A novel approach to the efficient and backward compatible compression of multi-channel 3D spatial audio is presented. Based on a derivation proving the equivalence between multiple amplitude panning methods and general Cartesian orthogonal analysis, a 3D Cartesian orthogonal analysis algorithm is proposed for efficient estimation of 3D sound source azimuth/elevation localisation information from an arbitrary 3D reproduction loudspeaker setup. A complete encoding/decoding system is designed based on this algorithm, where the derived 3D source azimuth/elevation localisation information is spatially quantised by the proposed $S^3$AC SLQP approach. The $S^3$AC spatial squeezing approach is extended so as to perform a unique mapping between the 3D SLQP and the 60° azimuth region represented by a stereo downmix. A ‘mono downmix + 3D side information’ approach is also proposed, where the 3D SLQP can be further quantised with different precisions defined by the user. Different 3D reproduction methods for decoding are discussed, where higher order Ambisonics reproduction can be used in symmetrical loudspeaker arrays and a 3D amplitude panning method based on inverse orthogonal analysis can be used in asymmetrical loudspeaker arrays for better source localisation. Objective evaluations
Table 6.4 Average azimuth/elevation error (in degrees) from $S^3$AC 3D stereo coding

<table>
<thead>
<tr>
<th>File</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Azimuth Error</td>
<td>13.03</td>
<td>0.06</td>
<td>15.04</td>
<td>0.08</td>
<td>1.75</td>
<td>7.56</td>
<td>3.81</td>
<td>3.85</td>
</tr>
<tr>
<td>Elevation Error</td>
<td>1.17</td>
<td>1.23</td>
<td>1.31</td>
<td>1.90</td>
<td>2.02</td>
<td>1.26</td>
<td>1.34</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Figure 6.17 $S^3$AC 3D listening test results
show that, while the proposed 3D amplitude panning method can efficiently synthesize an intended source localisation feature defined by the user, the 3D orthogonal source azimuth/elevation localisation estimation algorithm can efficiently derive the localisation information of a three-dimensionally positioned source. The objective results of the proposed 3D S³AC compression algorithms, including stereo downmixing, mono downmixing with high SLQP precision and mono downmixing with low SLQP precision, are also presented. This 3D multi-channel audio compression technique, including the three different modes, is further evaluated by subjective experiments. The results indicate that, while the bandwidth requirement is significantly reduced from 16-channel to 2-channel (or less), the degradation in perceptual quality remains minimal after decoding.
Chapter 7

Conclusions and Further Work

7.1 Conclusions

The Spatially Squeezing Surround Audio Coding, \( S^3AC \), has been presented in this thesis as an efficient approach for the representation of spatial audio signals. Based on exploiting perceptual localisation irrelevancy, the \( S^3AC \) scheme provides low bit rate compression solution to various types of 2D and 3D spatial audio signals, while the \( S^3AC \) stereo/mono downmix can be played-back by conventional stereo systems if surround loudspeaker systems are not available. The following main conclusions from this thesis are drawn:

**\( S^3AC \) Compression of ITU 5.1-Channel Signals**

As a typical application of \( S^3AC \), an ITU 5.1-channel signal can be represented by an \( S^3AC \) stereo downmix, which contains a ‘squeezed’ version of the original surround soundfield. By exploiting perceptual localisation irrelevancy, source localisation distortion is limited after decoding the \( S^3AC \) stereo downmix to a 5.1-channel signal. Improvement on decoded localisation accuracy is found in \( S^3AC \), when compared with existing compression approaches for 5.1-channel signals. In addition, further bit rate reduction is achieved by introducing mono downmixing approach in \( S^3AC \), while source localisation information is saved as \( S^3AC \) spatial cues.
Psychoacoustics Based $S^3AC$ Spatial Cue Quantisation

While the $S^3AC$ spatial cues directly represent source localisation information, an efficient quantisation solution is developed. Based on perceptual evaluation results, quantisation precision for $S^3AC$ localisation cues is dependent on derived source location. This is exploited for efficiently reducing the transmission bit rate for $S^3AC$ cues, while minimizing perceptual localisation distortion.

Limitation of $S^3AC$ Spatial Squeezing

Localisation resolution derivable from an $S^3AC$ stereo downmix is analyzed and shown to be frequency/source dependent. It is shown that, source with high spectral amplitude, i.e. with high perceptual importance, inheritably has higher derivable localisation resolution in an $S^3AC$ stereo downmix. Based on this, further analysis is shown for evaluating more intensive $S^3AC$ soundfield squeezing algorithms than squeezing to a standard 60° soundfield. This indicates $S^3AC$ spatial squeezing limitation without introducing significant perceptual localisation loss. This results is further exploited, where an $S^3AC$ approach for representing multiple surround soundfields with one stereo downmix is introduced.

$S^3AC$ Compression of Ambisonics Signals

The $S^3AC$ soundfield analysis algorithm and spatial squeezing approach is exploited for compression of Ambisonics B-format soundfield recording signals. This approach not only provides bit rate reduction, but extended backward compatibility to Ambisonics signals. While an Ambisonics recorded surround soundfield can be compressed into an $S^3AC$ stereo downmix, or mono downmix with $S^3AC$ cues, the downmix signal can be used in classical audio system. In addition, the $S^3AC$ compressed signal can be directly decoded into multichannel loudspeaker signal, so that Ambisonics decoding limitation, such as requiring regular loudspeaker system, can be overcome. When compared with conventional Ambisonics backward compatible
Compressions and Further Work

A compression solution, UHJ, $S^3$AC offers significant improvement in terms of both quality and localisation.

**$S^3$AC Binaural Reproduction**

For users without multichannel loudspeaker systems, the $S^3$AC binaural reproduction solution is proposed. For any type of $S^3$AC encoded signal, a binaural signal rendering virtual surround soundfield can be generated based on HRTF based techniques. The $S^3$AC representation of sound source localisation, directly as angular information, results in efficient HRTF look-up algorithm for the proposed approach, which does not introduce additional computational cost as found in other binaural approaches for encoded spatial audio signals.

**$S^3$AC for Spatialised Teleconferencing**

The $S^3$AC soundfield squeezing idea is further exploited for perceptually discriminat- ing soundfields from different remote sites during a multi-site teleconferencing. In this application, received speech soundfields from different remote sites are squeezed into discriminated sectors for re-rendering at the local site. This provides benefits that, users at the local site can perceptually disambiguate speakers from different remote sites, while relative speaker location within one remote site is also maintained.

**Signal Format Independent Source Localisation Estimation**

Based on a derivation showing fundamental equivalence between multiple spatialisation techniques, including amplitude panning, vector panning, Ambisonics reproduction and geometrical analysis, an orthogonal analysis algorithm on Cartesian coordinates is presented for estimating sound source localisation for both 2D and 3D audio signals. This algorithm can be used on any arbitrary loudspeaker format for deriving soundfield and source localisation information, while only loudspeaker location information (azimuth/elevation) is required.
S$^3$AC Compression of Multichannel 3D Audio

The S$^3$AC is finally extended for efficient representation of multichannel 3D audio. The 3D orthogonal estimation algorithm is utilized for deriving source localisation in a 3D soundfield rendered by any arbitrary loudspeaker setup. The derived source localisation is then quantised with the proposed SLQP 3D location quantisation approach, which is designed to minimize perceptual localisation distortion. The 3D sound source and quantised localisation can be represented by squeezing into a stereo downmix or a mono downmix with localisation cues. While no existing technique is found for this application, the proposed S$^3$AC approach compresses multichannel 3D audio using bandwidth no more than stereo, while backward compatibility to conventional audio system is also maintained. Evaluation results show that high efficiency is achieved in terms of both quality and localisation accuracy, when compared the coded signal with the original signal.

7.2 A ‘Sound Source + Localisation’ Framework

Based on the various S$^3$AC schemes and applications presented in this thesis, a ‘sound source + localisation’ framework for representation of spatial audio signals is proposed, and illustrated in Figure 7.1. In this framework, any spatial audio signal containing information of either a 2D or 3D soundfield can be analyzed in this framework, resulting in a universal soundfield representation format including source auditory information and related angular localisation information. This derived source and localisation information is independent on the input spatial audio signal format, and can then be saved as either a mono downmix accompanying localisation side information or a stereo downmix containing ‘squeezed’ soundfield. When decoding, as the information retrieved from the transmitted signal is the universal soundfield information containing source and localisation (i.e. independent on the original signal format), it can be reproduced to any spatial audio signal format based on the
users’ choice or available hardware. Compared with the ‘downmix + cues’ framework discussed in Section 2.5.7, this ‘sound source + localisation’ framework has the following advantages:

- The reproduction system is independent on the original spatial audio signal format, for maximum flexibility. In addition, efficient backward compatibility can be achieved with minimum computation cost by adopting the binaural reproduction scheme presented in this thesis.

- By utilizing the 2D and 3D orthogonal source localisation estimation algorithm presented in this thesis, this framework has minimum dependency on the input spatial audio signal format. In addition, computational complexity increment caused by increased number of input channels is also limited.

- By representing source localisation information directly as angular information, psychoacoustical principles can be efficiently exploited for removing perceptual irrelevancy in the signal, e.g. the psychoacoustics based cue quantisation approached presented in this thesis.

7.3 Further Work

While the S$^3$AC technique has been presented in this thesis for efficient and backward compatible compression of both 2D and 3D spatial audio signals, possible further work based on this thesis is:

- Investigation and further evaluation on the proposed ‘sound source + localisation’ framework, where a comprehensive system can be used for encoding and reproducing spatial audio signals with multiple input and output formats independent to each other.
• Further exploiting the S$^3$AC squeezed downmix soundfield, while additional metadata can be saved in spectral domain without introducing side information, e.g. using frequency coefficients with minimum perceptual importance (such as high frequency components) to save metadata containing important perceptual information.

• Investigation on utilizing S$^3$AC psychoacoustics based spatial cue quantisation approach for quantising spatial cues derived by other spatial audio coders with higher efficiency.

• Investigation on signal format independent reproduction algorithms, for both 2D and 3D soundfield re-rendering, based on extending the proposed Cartesian orthogonal analysis, which can be used for reproducing spatial audio signals over any arbitrary loudspeaker configuration while Ambisonics reproduction is not desired.

• Investigation on efficient algorithm for estimating soundfield diffuseness com-
ponents, as well as algorithms for representing diffuseness components by squeezed soundfield.

- Further exploiting the S$^3$AC scheme for representing multiple soundfields using one downmix soundfield, while more than two encoded soundfields are compressed.

- Further investigation and evaluation of encoding 3D multichannel audio signal with irregular loudspeaker configuration and numbers of channels more than 16.
Bibliography


[72] I. M. Neoran, “Surround Sound Mixing using Rotation, Stereo Width, and Distance Pan-Pots,” in Proc. 109th AES Convention, 2000, Los Angeles, CA, USA.


Appendix A

Higher Order Ambisonics and the Related Spherical Harmonics Functions
<table>
<thead>
<tr>
<th>Order</th>
<th>Channel Name</th>
<th>Channel Number</th>
<th>$Y_n(\mu, \eta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>W</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>X</td>
<td>2</td>
<td>$\cos(\mu) \cos(\eta)$</td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>3</td>
<td>$\sin(\mu) \cos(\eta)$</td>
</tr>
<tr>
<td></td>
<td>Z</td>
<td>4</td>
<td>$\sin(\eta)$</td>
</tr>
<tr>
<td>2</td>
<td>R</td>
<td>5</td>
<td>$\sqrt{\frac{3}{2}} \cos(2\mu) \cos^2(\eta)$</td>
</tr>
<tr>
<td></td>
<td>S</td>
<td>6</td>
<td>$\sqrt{\frac{3}{2}} \sin(2\mu) \cos^2(\eta)$</td>
</tr>
<tr>
<td>3</td>
<td>T</td>
<td>7</td>
<td>$\sqrt{\frac{3}{5}} \cos(\mu) \sin(2\eta)$</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>8</td>
<td>$\sqrt{\frac{3}{5}} \sin(\mu) \sin(2\eta)$</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>9</td>
<td>$\frac{1}{5}(3 \sin^2(\eta) - 1)$</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>$\sqrt{\frac{5}{8}} \cos(3\mu) \cos^3(\eta)$</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>11</td>
<td>$\sqrt{\frac{5}{8}} \sin(3\mu) \cos^3(\eta)$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>12</td>
<td>$\sqrt{\frac{15}{8}} \cos(2\mu) \sin(\eta) \cos^2(\eta)$</td>
</tr>
<tr>
<td>4</td>
<td>N</td>
<td>13</td>
<td>$\sqrt{\frac{15}{8}} \sin(2\mu) \sin(\eta) \cos^2(\eta)$</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>14</td>
<td>$\sqrt{\frac{3}{5}} \cos(\mu) \cos(\eta)(5 \sin^2(\eta) - 1)$</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>15</td>
<td>$\sqrt{\frac{3}{5}} \sin(\mu) \cos(\eta)(5 \sin^2(\eta) - 1)$</td>
</tr>
<tr>
<td></td>
<td>Q</td>
<td>16</td>
<td>$\frac{1}{7} \sin(\eta)(5 \sin^2(\eta) - 3)$</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>17</td>
<td>$\sqrt{\frac{70}{128}} \cos(4\mu) \cos^4(\eta)$</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>18</td>
<td>$\sqrt{\frac{70}{128}} \sin(4\mu) \cos^4(\eta)$</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>19</td>
<td>$\sqrt{\frac{105}{24}} \sin(\eta) \cos^3(\eta) \cos(3\mu)$</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>20</td>
<td>$\sqrt{\frac{105}{24}} \sin(\eta) \cos^3(\eta) \sin(3\mu)$</td>
</tr>
<tr>
<td>4</td>
<td>E</td>
<td>21</td>
<td>$\sqrt{\frac{7}{16}} (7 \sin^2(\eta) - 1) \cos^2(\eta) \cos(2\mu)$</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>22</td>
<td>$\sqrt{\frac{7}{16}} (7 \sin^2(\eta) - 1) \cos^2(\eta) \sin(2\mu)$</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>23</td>
<td>$\sqrt{\frac{7}{8}} (7 \sin(\eta) - 3) \sin(\eta) \cos^2(\mu)$</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>24</td>
<td>$\sqrt{\frac{7}{8}} (7 \sin(\eta) - 3) \sin(\eta) \cos(\mu) \sin(\mu)$</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>25</td>
<td>$\frac{1}{8}(35 \sin^4(\eta) - 30 \sin^2(\mu) + 3)$</td>
</tr>
</tbody>
</table>

**Table A.1**
Higher order Ambisonics up to $4^{th}$ order and the related spherical harmonics functions [83]