The Development of SAALTS:
A Spatial Audio Attribute Listener Training System

by

Rafael Kassier

Submitted for the degree of Doctor of Philosophy

Institute of Sound Recording
Department of Music & Sound Recording
Faculty of Arts & Human Sciences
University of Surrey

December 2007
© Rafael Kassier 2007
ABSTRACT

This Thesis describes the background, motivation, development and testing of a Spatial Audio Attribute Listener Training System (SAALTS).

Previous studies in spatial audio attributes are examined, and Rumsey’s Scene-Based Paradigm is found to be the only one that provides a rigorous approach for describing spatial audio scenes. Issues that would cause complications when implemented in a training system are resolved with the development of the Simplified Scene-Based Paradigm (SSBP) which can be used in the description of a wide range of spatial audio scenes for normative, product evaluation or training investigations.

A pilot study to ascertain the effectiveness of a spatial audio attribute training system based upon the training of ranking tasks, and its transferral to tasks involving the rating of spatial audio attributes is reported. As a result of the pilot study, it is concluded that training naïve listeners in the concept and judgement of spatial audio attributes as outlined in the SSBP is possible. This training, however is only found to have transferred as an increase in the range of the scale used by a (potentially more motivated) sub-set of the trained listeners.

Informed by the pilot study and the literature on transfer and motivation, the Spatial Audio Attribute Listener Training System (SAALTS) detailed in this thesis employs the following elements: a tutorial explaining the SSBP and its importance; active learning using the Spatial Audio Attribute Toolkit (SAAT) and self-guided training drills involving motivation-inspiring elements.

As a result of the research described in the transfer investigation, it can now be concluded that training in spatial listening can improve performance in spatial audio evaluation tasks that transfers to similar tasks with similar and different stimuli. It was also found that the performance of SAALTS was comparable to a repetitive practice regime for the target task.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>II</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>III</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>VIII</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>XII</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>XIII</td>
</tr>
<tr>
<td><strong>1 INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td><strong>2 BACKGROUND THEORY</strong></td>
<td>5</td>
</tr>
<tr>
<td>2.1 Previous Work in Spatial Audio Attributes</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1 Eisler’s 1966 Sound Quality Experiment</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2 Nakayama et al. Study</td>
<td>7</td>
</tr>
<tr>
<td>2.1.3 Gabrielsson’s Sound Quality Tests</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3.1 Monophonic Loudspeaker tests: 1974</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3.2 Mono Loudspeaker and Stereo Headphone tests: 1979</td>
<td>10</td>
</tr>
<tr>
<td>2.1.3.3 Stereophonic Loudspeaker Tests: 1985</td>
<td>11</td>
</tr>
<tr>
<td>2.1.3.4 Monophonic Earphone Tests: 1990</td>
<td>13</td>
</tr>
<tr>
<td>2.1.3.5 Summary of Gabrielsson’s Sound Quality Tests</td>
<td>14</td>
</tr>
<tr>
<td>2.1.4 Toole’s Stereo Loudspeaker Tests</td>
<td>14</td>
</tr>
<tr>
<td>2.1.4.1 Mono/Stereo Series I: Spatial Attributes within an Overall Rating of Sound Quality)</td>
<td>15</td>
</tr>
<tr>
<td>2.1.4.2 Mono/Stereo Series II: Spatial Quality Separated from Sound Quality)</td>
<td>17</td>
</tr>
<tr>
<td>2.1.4.3 Summary of Toole’s work on spatial attributes</td>
<td>20</td>
</tr>
<tr>
<td>2.1.5 Letowski’s MURAL</td>
<td>22</td>
</tr>
<tr>
<td>2.1.6 Zacharov &amp; Koivuniemi studies</td>
<td>24</td>
</tr>
<tr>
<td>2.1.6.1 Test stimuli</td>
<td>24</td>
</tr>
<tr>
<td>2.1.6.2 Naïve preference test</td>
<td>25</td>
</tr>
<tr>
<td>2.1.6.3 Direct attribute elicitation</td>
<td>26</td>
</tr>
<tr>
<td>2.1.6.4 Training stimuli</td>
<td>30</td>
</tr>
<tr>
<td>2.1.6.5 Direct attribute rating and analysis</td>
<td>30</td>
</tr>
<tr>
<td>2.1.6.6 Summary</td>
<td>32</td>
</tr>
<tr>
<td>2.1.7 Berg’s Studies</td>
<td>33</td>
</tr>
<tr>
<td>2.1.7.1 Experiment 1</td>
<td>33</td>
</tr>
<tr>
<td>2.1.7.2 Experiment 2</td>
<td>37</td>
</tr>
<tr>
<td>2.1.7.3 Experiment 3</td>
<td>40</td>
</tr>
<tr>
<td>2.1.7.4 Summary of Berg’s Studies</td>
<td>44</td>
</tr>
<tr>
<td>2.1.8 Rumsey’s Scene-Based Paradigm</td>
<td>44</td>
</tr>
<tr>
<td>2.1.9 Summary of Spatial Audio Attributes Used in Previous Studies</td>
<td>48</td>
</tr>
<tr>
<td><strong>2.2 Previous Work in Listener Training</strong></td>
<td>53</td>
</tr>
<tr>
<td>2.2.1 General Training Studies Involving Audio</td>
<td>53</td>
</tr>
<tr>
<td>2.2.1.1 ITU-R Standards in Subjective Assessment of Audio: ‘Familiarisation and Training’</td>
<td>53</td>
</tr>
</tbody>
</table>
Table of Contents

2.2.1.2 Kirk's Preference Training Experiment ................................................................. 54
2.2.1.3 Bech's Selection & Training Experiment .......................................................... 55
2.2.1.4 Retholtz's Training System for Sonarmen ......................................................... 56
2.2.1.5 Zwistocki et al. Study into Practice and Feedback ........................................... 56
2.2.1.6 Hawkey et al. Initial Learning Experiment ....................................................... 57
2.2.1.7 Drennan & Watson's Extended Training Experiment ......................................... 57
2.2.1.8 Jones et al. Close-Combat Trainer Involving Spatial Audio ................................ 58

2.2.2 Previous Timbral Ear Training Systems .................................................................. 59
2.2.2.1 Warsaw Academy of Music's Timbre Solfège ......................................................... 59
2.2.2.2 Everest's 'Critical Listening' Ear Training Programme ........................................... 60
2.2.2.3 Brixen's Spectral Ear Training Programme ......................................................... 61
2.2.2.4 Quesnel's Timbral Ear Trainer ............................................................................ 61
2.2.2.5 Moulton's 'Golden Ears' training programme ..................................................... 63
2.2.2.6 Olive's 'So you want to be a Harman Listener?' ............................................... 63

2.2.3 Previous Spatial Audio Ear Training Systems ......................................................... 65
2.2.3.1 Zacharov & Koivuniemi's Spatial Ear Training ..................................................... 65
2.2.3.2 Corey's Reverberation Matching Training System .................................................. 65
2.2.3.3 Merimaa & Hess: 'Training of Listeners for Evaluation of Spatial Attributes of Sound'. 66
2.2.3.4 Neher's Spatial Trainer ....................................................................................... 67

2.2.4 Summary of Work in Listener Training ................................................................. 68

2.3 Transfer & Motivation ............................................................................................... 70

2.3.1 Transfer of Training ............................................................................................... 70
2.3.1.1 Definition of Transfer .......................................................................................... 70
2.3.1.2 Transfer of learning: Experiment designs ............................................................ 71
2.3.1.3 Specific factors in transfer .................................................................................. 72
2.3.1.4 General factors in transfer .................................................................................. 73
2.3.1.5 Teaching for transfer ......................................................................................... 74
2.3.1.6 Procedural and Declarative Objectives and Transfer .......................................... 74
2.3.1.7 High and Low Road Transfer ............................................................................ 75
2.3.1.8 Ahissar's Reverse Hierarchy and Transfer ............................................................ 76
2.3.1.9 Sternberg & Frensch's Four Mechanisms of Transfer ......................................... 76
2.3.1.10 Transfer on Trial .............................................................................................. 76
2.3.1.11 Transfer Summary ......................................................................................... 77

2.3.2 Motivation .............................................................................................................. 77
2.3.2.1 Malone & Lepper's Taxonomy of Intrinsic Motivations for Learning .................. 78
2.3.2.2 Keller & Suzuki's ARCS Motivation Model .......................................................... 80
2.3.2.3 Summary of Motivation ................................................................................... 82

2.4 Overall Summary .................................................................................................... 85

3 The Simplified Scene-Based Paradigm (SSBP) ............................................................ 86

3.1 Criteria for the Inclusion of Spatial Audio Attributes in a Training Paradigm ........... 86
3.1.1 Discriminate .......................................................................................................... 87
## Table of Contents

### 3.1 OBJECTIVE

3.1.2 Nonredundant ........................................................................................................... 87  
3.1.3 Relate to consumer acceptance / rejection .............................................................. 88  
3.1.4 Relate to instrumental or physical measurements ................................................... 88  
3.1.5 Singular ..................................................................................................................... 89  
3.1.6 Precise and reliable ................................................................................................. 89  
3.1.7 Consensus on meaning / unambiguous ................................................................. 89  
3.1.8 Reference easy to obtain ....................................................................................... 90  
3.1.9 Communicate .......................................................................................................... 90  
3.1.10 Relate to reality ..................................................................................................... 90  
3.1.11 Summary of criteria .............................................................................................. 91  

### 3.2 CONSTRUCTING A PARADIGM FOR SPATIAL EAR TRAINING

3.2.1 Previous studies into spatial audio .................................................................... 92  
3.2.2 The Scene-Based Paradigm revisited .................................................................... 93  
3.2.3 The Scene Component (SC) Concept .................................................................... 98  
3.2.4 Training System Spatial Audio Attributes .............................................................. 101  

### 3.3 DISCUSSION

3.3.1 Method ..................................................................................................................... 104  
3.3.2 Results ..................................................................................................................... 109  
3.3.3 Selecting Candidates for the Training Programme ................................................ 110  
3.3.4 Summary ............................................................................................................... 115  

### 3.4 SUMMARY

3.4 Summary ...................................................................................................................... 119  

### 4 PILOT INVESTIGATION

4.1 THE COLLECTION OF ‘REAL-WORLD’ STIMULI .................................................... 106  
4.1.1 Method ..................................................................................................................... 107  
4.1.2 Informal Evaluation .............................................................................................. 108  
4.1.3 Summary ............................................................................................................... 109  

4.2 SUBJECTIVE ATTRIBUTE RATING: EXPERIENCED LISTENER PANEL ............. 110  
4.2.1 Method ..................................................................................................................... 110  
4.2.2 Results ..................................................................................................................... 112  
4.2.3 Reducing the Number of Programme Items .......................................................... 115  
4.2.4 Summary ............................................................................................................... 119  

4.3 SUBJECTIVE ATTRIBUTE RATING: ‘PRE-TRAINING’ .............................................. 120  
4.3.1 Method ..................................................................................................................... 121  
4.3.2 Results ..................................................................................................................... 124  
4.3.3 Selecting Candidates for the Training Programme ................................................ 128  
4.3.4 Summary ............................................................................................................... 130  

4.4 NAIVE LISTENER TRAINING PROGRAMME ............................................................. 131  
4.4.1 Method ..................................................................................................................... 133  
4.4.2 The Tutorial Phase ............................................................................................... 135  
4.4.3 The Drill Phase ...................................................................................................... 137  
4.4.4 Drill Performance ................................................................................................. 139  
4.4.5 Pre/Post Test Ranking Performance ..................................................................... 141  
4.4.6 Questionnaire Analysis ....................................................................................... 143

- V -
Table of Contents

4.4.7 Summary ..................................................................................................................147
4.5 SUBJECTIVE ATTRIBUTE RATING: ‘POST-TRAINING’ ........................................149
  4.5.1 Method ....................................................................................................................149
  4.5.2 Results ..................................................................................................................151
        4.5.2.1 Analysis of Overall ‘Error’ and Partial Eta Squared ........................................151
        4.5.2.2 Scatter Plot Analysis ......................................................................................153
        4.5.2.3 Consistency and Sensitivity Change Metric Analysis ......................................157
        4.5.2.4 SSE and Range Analysis .................................................................................161
        4.5.2.5 Analysis of Time Taken for the Task .................................................................169
        4.5.2.6 Post Experiment Questionnaire ......................................................................170
4.6 PILOT INVESTIGATION: OVERALL SUMMARY ....................................................173
5 TRANSFER INVESTIGATION .........................................................................................175
  5.1 TRANSFER INVESTIGATION OUTLINE ...............................................................175
  5.2 STIMULUS CREATION .............................................................................................178
        5.2.1 Ranking/Training Task Stimulus Set Creation ..................................................178
        5.2.2 ‘Real-World’ Stimulus Creation .......................................................................181
  5.3 RANKING SESSIONS ...............................................................................................185
        5.3.1 Familiarisation .................................................................................................185
        5.3.2 Ranking Tasks ................................................................................................186
        5.3.3 Initial Ranking Session (‘Task 1’) ....................................................................187
        5.3.4 Selection of Subject Groups .............................................................................188
  5.4 SPATIAL AUDIO ATTRIBUTE LISTENER TRAINING SYSTEM (SALTS) ..............189
        5.4.1 Tutorial ............................................................................................................190
        5.4.2 Spatial Audio Attribute Toolkit (SAAT) ..............................................................190
        5.4.3 Training Drills ................................................................................................192
  5.5 ‘REAL-WORLD’ STIMULUS RATING SESSIONS ..................................................196
        5.5.1 Familiarisation & Practice ..............................................................................196
        5.5.2 Rating Task .....................................................................................................197
  5.6 ANALYSIS OF RESULTS ........................................................................................198
        5.6.1 Question 1 (Initial Inter-Group Performance Equivalence) ....................... 199
        5.6.2 Question 2 (Intra-Group Near Transfer) .........................................................200
              5.6.2.1 ED Tasks ..................................................................................................200
              5.6.2.2 EW Tasks ..............................................................................................200
        5.6.3 Question 3 (Inter-Group Near Transfer) .........................................................201
              5.6.3.1 Ranking Performance ..........................................................................201
              5.6.3.2 Time Performance ...............................................................................201
        5.6.4 Question 4 (Intra-Group Far Transfer) ............................................................201
              5.6.4.1 ED Tasks ..................................................................................................202
              5.6.4.2 EW Tasks ..............................................................................................202
        5.6.5 Question 5 (Inter-Group Far Transfer) ............................................................202
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.6.5.1 Ranking Performance</td>
<td>202</td>
</tr>
<tr>
<td>5.6.5.2 Time Performance</td>
<td>202</td>
</tr>
<tr>
<td>5.6.6 Question 6 (Intra-Group Improvement: Practice vs. Training)</td>
<td>202</td>
</tr>
<tr>
<td>5.6.7 Question 7 (Inter Group Improvement: Practice vs. Training)</td>
<td>203</td>
</tr>
<tr>
<td>5.6.8 Question 8 (Practice Performance)</td>
<td>203</td>
</tr>
<tr>
<td>5.6.9 Task 10 ('Further' Transfer)</td>
<td>206</td>
</tr>
<tr>
<td>5.7 Transfer Investigation Summary</td>
<td>206</td>
</tr>
<tr>
<td>6 CONCLUSIONS &amp; FURTHER WORK</td>
<td>210</td>
</tr>
<tr>
<td>7 APPENDICES</td>
<td>214</td>
</tr>
<tr>
<td>7.1 Reproduction from (Kassier, Lee, Brookes and Rumsey 2005)</td>
<td>214</td>
</tr>
<tr>
<td>7.1.1 Selected Microphone techniques</td>
<td>214</td>
</tr>
<tr>
<td>7.1.1.1 Front Microphone 'Triplet' Techniques</td>
<td>215</td>
</tr>
<tr>
<td>7.1.1.2 4-Channel 'Rear' Microphone Techniques</td>
<td>218</td>
</tr>
<tr>
<td>7.1.1.3 2-Channel 'Rear' Microphone Techniques</td>
<td>219</td>
</tr>
<tr>
<td>7.1.2 Experimental Set-up</td>
<td>220</td>
</tr>
<tr>
<td>7.2 Training Programme PowerPoint Slides</td>
<td>224</td>
</tr>
<tr>
<td>GLOSSARY</td>
<td>239</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>241</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>243</td>
</tr>
</tbody>
</table>
List of Figures

LIST OF FIGURES

Figure 1: Screenshot of 'So You Want To Be A Harman Listener?' ...................................................... 64
Figure 2: Three Examples of a Scene Component (SC) ......................................................................... 100
Figure 3: Elements of the new paradigm ................................................................................................. 103
Figure 4: Photo of the recording set-up in Studio 1. ............................................................................... 107
Figure 5: Familiarisation screen for experienced listeners ................................................................... 111
Figure 6: Practice page for experienced listeners ................................................................................... 111
Figure 7: Experimental page for the experienced listeners ................................................................... 112
Figure 8: Mean width grades given for each subject ............................................................................ 113
Figure 9: Mean width grades given to each ‘process’ ............................................................................ 113
Figure 10: Mean width grades by front microphone technique .......................................................... 114
Figure 11: Mean width grades by rear microphone technique ............................................................. 115
Figure 12: Mean width grades by programme type ................................................................................ 116
Figure 13: Mean width grades by items not selected for the training experiment ................................. 116
Figure 14: Mean width grades by items selected for the training experiment ....................................... 117
Figure 15: Scatter plot indicating the comparative difficulties of consistent grading against the ..... 119
comparative differences between each of the ‘processes’ in each item .....................................
Figure 16: Relative Experience between Courses. ................................................................................. 122
Figure 17: Comparison of mean width grades by Experienced and Naive listeners to stimuli .... 124
incorporating different front microphone techniques (six selected programme items only). ...
Figure 18: Comparison of mean width grades by Experienced and Naive listeners to stimuli .... 125
incorporating different rear microphone techniques (six selected programme items only). ...
Figure 19: Scatter plots showing relative performances (RMS Error against Partial Eta Squared) of the 127
4 main subgroups of subjects (By Course, and by Gender) .........................................................
Figure 20: Bar graph showing mean time taken for the first two experimental sessions for the 129
candidates selected to be in the trained and untrained groups. ......................................................
Figure 21: Scatter plots showing performance of individual subjects (RMS Error against Partial Eta 129
Squared) in the 4 main subgroups of subjects (By Course, and by Gender).....................................
Figure 22: Bar graphs showing the initial mean ‘error’ of trained and untrained groups ...................... 130
Figure 23: Bar graphs showing the initial mean Partial Eta Squared values of trained and untrained 130
groups ..............................................................................................................................................
Figure 24: Photo of Studio 3 showing the equipment set-up for the main training sessions .............. 134
Figure 25: Photo of Studio 3 showing the equipment set-up during the audio examples phase of the 134
tutorial ..............................................................................................................................................
Figure 26: Visual representation of the narrowest familiarisation stimulus ........................................... 137
Figure 27: Visual representation of the widest familiarisation stimulus .............................................. 137
Figure 28: Screen capture of the ‘performance’ window of the implemented training software. ...... 139
Figure 29: Charts of highest difficulty level (not a linear scale) achieved in each session against 140
session number for each of the trained subjects ............................................................................
Figure 30: Bar graph showing the mean SSED for all trained subjects, for the pre- and post-training 142
tests..................................................................................................................................................
List of Figures

Figure 31: Bar graph showing the SSED of each trained subject, for the pre- and post-training tests. .................................................................142
Figure 32: Bar graph showing the mean time taken (in seconds) for all trained subjects, for the pre- and post-training tests. ...........................143
Figure 33: Bar graph showing the mean time taken to complete the familiarisation phases during pre- and post-training for the different course groups. ..............................................................150
Figure 34: Bar graph showing the mean pre- and post-training RMS error values for each subject for all programme items. .................................................................151
Figure 35: Bar graph showing the results of the diff. grade analysis. .................................................................152
Figure 36: Bar graph showing the mean pre- and post-training Partial Eta Squared values for each subject for all programme items. .................................................................153
Figure 37: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for experienced listeners. .................154
Figure 38: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for trained listeners. Pre-Training Sessions. .................................................................154
Figure 39: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for trained listeners. Post-Training Sessions. .................................................................155
Figure 40: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for untrained listeners. Pre-Training Sessions. .................................................................155
Figure 41: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for untrained listeners. Post-Training Sessions. .................................................................156
Figure 42: Scatter plot graphs showing Consistency Change Metric and Sensitivity Change Metric for Untrained and Trained Subjects. .................................................................158
Figure 43: Scatter plot graphs showing Consistency Change Metric and Sensitivity Change Metric for Untrained and Trained Subjects by Course. .................................................................158
Figure 44: Bar graphs showing the mean Consistency Change Metric for Untrained and Trained Subjects by Course. ..........................159
Figure 45: Bar graphs showing the mean Sensitivity Change Metric for Untrained and Trained Subjects by Course. ..........................160
Figure 46: Bar graph showing the mean Sensitivity Change Metric for Trained Tonmeister Subjects. .................................................................160
Figure 47: Bar graph showing SSED Pre- and Post-Training for each subject for all programme items. .................................................................162
Figure 48: Bar graphs showing mean SSED for Untrained and Trained Naïve Subjects by Course. .................................................................163
Figure 49: Bar graph showing mean Range Pre- and Post-Training for each subject for all programme items. .................................................................164
Figure 50: Bar graphs showing mean Ranges for each naïve subject (numbered above each graph) for all programme items during the pre and post tests. .................................................................164
Figure 51: Scatter plot graphs showing SSED change metric and Range Change Metric for Untrained and Trained Naïve Subjects by Course. .................................................................166
Figure 52: Bar graph showing SSED Change Metric for trained and untrained listeners. .................................................................167
Figure 53: Bar graphs showing SSED Change Metric for trained and untrained listeners by Course. .................................................................167
Figure 54: Bar graphs showing the effect of the training programme on the Range Change Metric. .................................................................168
Figure 55: Bar graphs showing the effect of the training programme on the Range Change Metric by Course. .................................................................168
List of Figures

Figure 56: Bar graph showing the time taken (in minutes) for the various subject groups (untrained and trained naïve subjects, and the experienced listeners) to complete their four nominally 30 minute listening tests. ................................................................. 169

Figure 57: Bar graphs showing the effect of the training programme on the Time Change Metric. .............................................. 170

Figure 58: Histogram showing response for the trained and untrained subjects to the statement “At first I found the task (judging width) difficult”. .................................................................................. 171

Figure 59: Histogram showing response for the trained and untrained subjects to the statement “By the end I found the task (judging width) difficult”. ................................................................. 171

Figure 60: Histogram showing response for the trained and untrained subjects to the statement “I am now able to use the width scale more consistently than when I started (e.g. wide sounds have always been graded wide)”. ................................................................. 172

Figure 61: Histogram showing response for the trained and untrained subjects to the statement “I am now able to detect differences between the widths better than when I started”. ................................................................. 172

Figure 62: Transfer Investigation Overview. ............................................................................................................................... 177

Figure 63: Photograph of the author checking the 3/2 Stereo reproduction system in the listening room. ................................................................. 178

Figure 64: Plan-view of the Ensemble Width (EW) simulation. ............................................................................................................................... 180

Figure 65: Plan-view of the Ensemble Depth (ED) simulation. ............................................................................................................................... 180

Figure 66: Plan View of the Experimental Recording Session. ............................................................................................................................... 182

Figure 67: Photograph of front triplets in the ‘front’ position. ............................................................................................................................... 183

Figure 68: Photograph taken from between the loudspeakers and the front wall, showing the loudspeakers, the microphones and the back wall. ............................................................................. 184

Figure 69: Photograph of the recording session, taken from behind the frontal triplets and showing the four loudspeakers and the front wall. ............................................................................. 184

Figure 70: Familiarisation Screenshot. ............................................................................................................................... 186

Figure 71: Ranking task screenshot. ............................................................................................................................... 187

Figure 72: Initial Session ‘Practice Page’ Screenshot. ............................................................................................................................... 188

Figure 73: Spatial Audio Attribute Toolkit showing SC Depth. ............................................................................................................................... 191

Figure 74: Spatial Audio Attribute Toolkit showing SC Width. ............................................................................................................................... 191

Figure 75: The training administration page. ............................................................................................................................... 192

Figure 76: Screenshot of a discrimination drill. ............................................................................................................................... 193

Figure 77: Screenshot of a pairwise ranking drill (for SC Depth). ............................................................................................................................... 193

Figure 78: A discrimination drill in progress. ............................................................................................................................... 194

Figure 79: Screenshot of the success screen. ............................................................................................................................... 195

Figure 80: Screenshot of the rating task ‘familiarisation and practice’ screen. ............................................................................................................................... 196

Figure 81: Screenshot of the rating task test screen. ............................................................................................................................... 197

Figure 82: Rank Accuracy. ......................................................................................................................................................... 199

Figure 83: Time Taken. ......................................................................................................................................................... 199

Figure 84: ED ARD over the practice sessions. ............................................................................................................................... 204

Figure 85: ED Time over the practice sessions. ............................................................................................................................... 204

Figure 86: EW ARD over the practice sessions. ............................................................................................................................... 205

Figure 87: EW Time over the practice sessions. ............................................................................................................................... 205

- x -
List of Figures

Figure 88: Fukada Tree ........................................................................................................................... 215
Figure 89: OCT Technique ..................................................................................................................... 216
Figure 90: INA-3 Technique ................................................................................................................... 217
Figure 91: Near-Coincident Technique ................................................................................................. 217
Figure 92: IRT-Cross-Inspired Technique ............................................................................................ 218
Figure 93: Hamasaki-Square Technique ............................................................................................... 219
Figure 94: Recording Set-Up ................................................................................................................. 221
Figure 95: ‘Front’ and ‘Rear’ Techniques (with Kassier & Lee) .............................................................. 221
Figure 96: ‘Front’ Microphone Arrays .................................................................................................. 222
Figure 97: ‘Rear’ Microphone Arrays (left spaced cardioid not shown) ................................................. 223
LIST OF TABLES

Table 1: Showing Rating Scales used in (Gabrielsson and Lindström 1985)........................................ 12
Table 2: Showing Rating Scales used in the Mono/Stereo Series II experiment in (Toole 1985)........... 18
Table 3: Letowski’s MURAL, after (Letowski 1989)........................................................................... 23
Table 4: Spatial and Timbral Attributes Elicited, after (Zacharov and Koivuniemi 2001a) ............... 27
Table 5: Grouped attributes and descriptions from (Berg and Rumsey 2001).................................. 38
Table 6: Consistency of attributes shown from (Berg and Rumsey 2001).......................................... 40
Table 7: Spatial attributes included as a result of the preliminary elicitation experiment in (Berg and Rumsey 2002).................................................................................. 41
Table 8: Attributes used in the main experiment in (Berg and Rumsey 2002)................................. 42
Table 9: Spatial Attributes used in the scene-based paradigm from (Rumsey 2002).......................... 45
Table 10: Summary of Dimensional Spatial Attributes from previous studies, placed within Rumsey’s Scene-Based Paradigm .................................................................................. 48
Table 11: Summary of Immersive Spatial Attributes from previous studies, placed within Rumsey’s Scene-Based Paradigm .................................................................................. 49
Table 12: Summary of Miscellaneous Spatial Attributes from previous studies, placed within Rumsey’s Scene-Based Paradigm .................................................................................. 49
Table 13: Showing five particular transfer test types, after (Ellis 1965).............................................. 71
Table 14: Similarities and differences between the Taxonomy and the ARCS Model.......................... 83
Table 15: Sample ANOVA table for the Accordion Programme Item.............................................. 117
Table 16: Sample ANOVA table for the Clarinet Programme Item................................................... 118
Table 17: Sample ANOVA table, Subject 108, Sessions 1 and 2 (‘Pre Training’ Sessions)................. 126
Table 18: RMS Error and Partial Eta Squared values for the first two naïve listener experimental sessions....................................................................................................................... 127
Table 19: Drill Difficulty Level and Description................................................................................. 138
Table 20: Programme Item Listing .................................................................................................... 179
Table 21: Loudspeaker Configurations (for the ‘area’ shown in Figure 66)...................................... 183
ACKNOWLEDGEMENTS

This research project was supported by a studentship grant from the Engineering and Physical Sciences Research Council (EPSRC). Travel funding assistance was provided by the Digital Music Research Network (DMRN).

The author would like to acknowledge the help and support provided by his supervisors, Prof. Dr. Francis Rumsey and Dr. Tim Brookes.
1 INTRODUCTION

The research in this Thesis leads towards and describes the development and testing of a Spatial Audio Attribute Listener Training System (SAALTS).

The Need for Training

Multichannel audio has experienced an increase in popularity (Rumsey 2001), mainly due to the enhanced audio capabilities of the Digital Versatile Disc (DVD) video distribution format over previously available formats. 3/2 Stereo as defined by the International Telecommunications Union (ITU-R 1992-1994) – more commonly known as 5.1 Surround Sound (Holman 1999) – has become a standard for domestic multichannel sound systems, whilst research into Wavefield Synthesis (WFS) (Berkhout, De Vries and Vogel 1993), renewed interest in Ambisonic formats (Craven, Law, Stuart and Wilson 2003) and a proposed ‘22.2 Channel’ system (Hamasaki, Hiyama, Nishiguchi and Ono 2004) have pointed towards future sound systems featuring potentially more compelling spatial sound representations.

Despite this recent interest in multichannel audio and the existence of various timbral ear training systems (detailed in Section 2.2.2), relatively little work has been done regarding the training of listeners in the spatial attributes of reproduced sound (see Section 2.2.3).

Trained listeners have been shown to become more consistent in their judgements. Bech (1992) and Olive (2003) have demonstrated that a very large number of untrained listeners are needed to achieve results with similar statistical validity to those achieved from a small number of trained listeners.

It has also been shown that subjects tend to have a faster response time and greater consistency after training using the Timbral Ear Trainer developed at McGill University (Quesnel and Woszczyk 1994; Quesnel 1996). However, the validation study used just seven subjects and did not have a control group to show what would have happened to subjects who did not take the course.

Listener training could be shown to be beneficial if it could provide naïve listeners with equivalent or better skills to experienced and expert listeners. Relying on identifying and utilising experienced and expert listeners may not be a viable method in commercial situations where listening tests are required. Firstly experienced and expert listeners may well vary in their levels of experience and their expectations. Secondly there may not be sufficient experienced or expert listeners available for any given situation. Through the use of a consistent training system, groups of trained listeners of more-or-less equivalent skill could be assembled from naïve listeners that may be more readily available, and all in a relatively short time period compared with the length of time required for a listener to acquire enough experience to be considered an experienced or expert listener.

A motivating issue for this Thesis was therefore whether a training system could be developed to successfully train spatial audio attribute listening skills.
Introduction

Previous Spatial Trainers

There are few reported training systems that involve spatial aspects of sound reproduction, and most of the reported ones feature no experimental verification (see Section 2.2.3).

Neher (2004) explains that an audio training system would require sets of stimuli that simulated changes in spatial attributes in a perceptually unidimensional manner. Neher explained that a perceptually unidimensional change in a spatial audio attribute may involve the change of a number of different physical factors. It is pointed out here that if spatial audio attributes are defined in an auditory manner using sets of reference audio stimuli (rather than in a verbal or graphical manner) language translation and other subject-dependent understanding issues are expected to be alleviated. Neher created four sets of stimuli that he deemed suitable for training exercises, along with a scheme to validate the perceptual unidimensionality of the various stimuli. He successfully simulated "Source Distance", "Ensemble Width" and "Ensemble Depth", and simulated "Source Width" with difficulty. Neher also conducted a preliminary training exercise using the stimuli that he created, but unfortunately only had time to test five subjects in total (three subjects were trained, two subjects were used as a control group). Results were encouraging, but a larger and possibly more rigorous validation test is needed in order to draw more valid conclusions about the effectiveness of the training, and to understand to what extent the training can be generalised.

Problems with Terminology

Rumsey explains that spatial attributes of reproduced sound are terms “concerned with describing and evaluating the three dimensional characteristics of the components of a spatial audio scene that is reproduced using loudspeakers or headphones” (Rumsey 2002).

The spatial audio attributes used in the majority of previous studies (see Section 2.1) have either been provided by the experimenter for use with specific stimuli and experimental conditions, or elicited from subjects using various stimuli and experimental conditions. They can therefore be described as dependent upon these stimuli and conditions. Spatial audio attributes used in one study may not necessarily apply to other studies. If a system is to be devised that will be able to cope with a wider range of stimuli, a different approach is needed in the selection of included attributes.

Toole comments that “only by creating a relatively fast method of obtaining useful subjective data, may reliable listening tests be used widely” (Toole 1985). A listener training system that is designed to allow for a range of programme materials and listening conditions is expected to help achieve this goal, by allowing a standardised way of providing subjects with spatial audio listening experience which they can take with them into specific listening tests.

It was hypothesised by this author that participation in a listener training programme concerned with the spatial aspects of sound reproduction would also help to train listeners to be more consistent and sensitive when evaluating spatial changes in audio reproduction using a different set of stimuli to those used in training (so called transfer of training). In order to demonstrate its usefulness outside the context of the
Introduction

stimuli used in training, any training scheme would need to show that learned skills were transferable.

In documents such as (ITU-R 1994-1997), the terms training and familiarisation (where the procedures involved in listening tests are explained to, and practised by, the test subjects) are used interchangeably, and in (Bech 1992) training could be better described as practising the task. For this research, training refers to a separate process where skills are taught and practised in a context not necessarily identical to the test conditions.

It may not, however, be entirely clear that terms are being used by subjects in the same way. According to Shaw and Gaines (1989), subjects can use the same terms for the same concepts, the same terms for different concepts, different terms for the same concept, and different terms for different concepts. In order to obtain meaningful data from subjective tests, the experimenter needs to be confident that subjects are using similar terms to describe each phenomenon. In addition, language translation issues, as highlighted by Teunissen (1996) and Martens and Giragama (2002) mean that caution must be advised when the terms used in any descriptive analysis experiment are translated from one language to another. A listener training programme that familiarises subjects with the various spatial attributes of sound and establishes clear relationships between terminology and auditory stimuli could be advantageous.

According to Meilgaard et al. (1991) the ability of subjects to discern and describe a particular sensory characteristic in a "sea" or "fog" of other sensory impressions is more important than sensory acuity. Therefore another motivating issue in this Thesis is the need for a universal language to describe spatial audio attributes for reproduced sound. This is particularly necessary for use in a spatial audio attribute training system.

Requirements for a Spatial Audio Attribute Listener Training System

The first concern addressed in this Thesis was the need for a spatial audio description language that could be used as the framework within which to base the training system. This description language needed to conform to various criteria, such as the need for unambiguous terms that did not overlap conceptually with one another. The resulting Simplified Scene-Based Paradigm (SSBP) is developed in Chapter 3.

Once this framework had been established, spatial audio attribute training systems were implemented and studied.

A pilot experiment (based upon Neher’s experimentation and research) was conducted to examine whether or not participation in a training programme would enhance the consistency, sensitivity and fluency performance of subjects in a spatial audio evaluation task. This is detailed in Chapter 4.

Alessi and Trollip (2001) suggest that a combination of behaviourist, cognitive and constructivist learning techniques be used in training programmes, and explain that effective and efficient learning is facilitated through four phases of learning (presenting information, guiding the learner, practising and assessing learning). Their recommendations were implemented in a new training system developed to utilise the SSBP, and that was optimised for transfer of training (see Section 2.3.1)
Introduction

and motivation (see Section 2.3.2). The resulting Spatial Audio Attribute Listener Training System (SAALTS) and its experimental verification is the subject of the transfer investigation detailed in Chapter 5.

Technical Constraints

Regarding the scope of the training system, it was limited to the popular standard 3/2 Stereo spatial audio reproduction system (ITU-R 1992-1994). This resulted in certain limitations being imposed. Precise stereophonic imaging outside the frontal 60° arc is not expected in the recommendation which dictates which attribute simulations are achievable. There is also the lack of the height dimension in the reproduction. SAALTS was devised with these limitations in mind, but its modular nature means that extensions could be easily implemented for future versions.

In order to provide as large an audience for SAALTS as possible, audiometric screening of subjects was not required. The focus of the system was spatial rather than timbral listening skills and did not involve signals at the threshold of detection.

Aim of the Thesis

This Thesis describes the background, motivation, development and testing of a Spatial Audio Attribute Listener Training System (SAALTS).

Before spatial training could begin, a suitable description language for spatial audio would be required. Once this had been established, the training programme needed to be developed and optimised to demonstrate that it could be useful in a wider context. The research presented in the following chapters seeks to determine: (i) what description language would be appropriate for training, (ii) if training in spatial listening can improve performance in spatial audio evaluation tasks, and to what extent, and (iii) how a generalised spatial audio attribute listener training programme will compare with other established methods.

Structure of the Thesis

Chapter 2 covers background theory in Spatial Audio Attributes, Listener Training, and studies in Transfer of Training and Motivation. Chapter 3 describes the development of the Simplified Scene-Based Paradigm (SSBP), a descriptive language that is based on previous studies but optimised for use in training systems. A training study to ascertain the effectiveness of a spatial audio attribute training system based upon the training of ranking tasks is described in Chapter 4. The development and verification of SAALTS is detailed in Chapter 5, overall conclusions are summarised and further work strategies suggested in Chapter 6.
2 BACKGROUND THEORY

Three broad areas have been identified for study. Initially, a paradigm needs to be established within which the training scheme will operate, hence previously published studies will be investigated in order to assess which spatial audio attributes should be incorporated. Studies involving training relating to reproduced audio (in general, as well as existing timbral ear trainers and studies involving spatial audio) will be investigated in order to establish any important factors to be heeded in the design of a spatial audio attribute training system. Because the effectiveness of training can be judged by the degree to which learned skills transfer to other situations, issues relating the transfer of training will be studied by reviewing the pertinent literature.

2.1 Previous Work in Spatial Audio Attributes

Describing measurement in science, Nunally and Bernstein (1994) point out that one measures the attributes of an object, rather than the object itself, and warned that "attributes should not be confounded with each another" (i.e. they need to be independent of one another). They also proposed that attributes must be carefully studied before they are measured: "an attribute we believe in may not exist in the form proposed" (Nunally and Bernstein 1994). It is thus crucial to examine which spatial attributes have been studied in the past in order to be able to propose a set of attributes to be used in spatial ear training.

To this end, a critical analysis of previously published studies that involve descriptions of the spatial aspects of sound has been undertaken. This section is separated into eight sub-sections, each detailing separate research efforts into spatial audio attributes. The studies are critically reviewed and common ground is found between them in order to derive a summary of previously established spatial attributes and paradigms for further consideration.

2.1.1 Eisler's 1966 Sound Quality Experiment

An early pilot study into the applicability of factor analysis (Nunally and Bernstein 1994) was undertaken by Hannes Eisler at the University of Stockholm. Eisler (1966) warns that the experiment was conducted in a way in which "many of the necessary precautions could not be taken", so care must be taken interpreting his results.

Eisler explained that physical measurements had normally been used to classify loudspeakers, whereas he was interested in the underlying perception of sound quality from a psychometric point of view. His experiment was designed to gather preference data from subjects and see whether factor analysis was a suitable tool to uncover the underlying perceptual factors pertaining to perceived sound quality.
Just four subjects were used in the experiment (Eisler was more concerned about getting a set of data that he could use in his factor analysis tests than in obeying correct experimental procedures). These were all “acoustical engineers” (Eisler 1966) employed because they were expected to be less frightened to use the extremes of the scale when indicating their preferences, as “factor analysis works best with large differences” (Eisler 1966).

The experiment itself was a listening test involving twenty-four stimuli with an incredible diversity of programme types, replayed in a random order over ten different loudspeaker systems (the loudspeakers’ controls were fixed and not adjusted during the test, resulting in occasional “overloading”, which is unfortunate). It is not clear which of the systems were mono and which were stereo as there appear to be have been a mixture of both in the tests. The tests took place in a non-standard listening room with the loudspeakers positioned behind a visually opaque cloth. All four subjects took part in the tests together (testing in groups of four as reported has obvious perils regarding inter-subject collusion and other interactions which could potentially bias their responses), with three sessions (consisting of 80 judgements each) scheduled on separate days. Each item was therefore played once through each loudspeaker system with no repetitions. No training or familiarisation of the test system or stimuli is reported to have been undertaken prior to testing. Subjects were asked to mark their responses on a seven point scale (with one decimal place allowable) ranging from 1 (“worst imaginable quality”) through 4 (“average sound quality”) to 7 (“best imaginable quality”).

The preference data provided by the subjects was analysed using factor analysis. Component analysis was used on the mean of the four subject’s responses, with an unusual implementation where the programme stimuli were used as the “tests” and the loudspeaker systems used as the “testees” (Eisler 1966). The nine resulting factors were tentatively labelled by Eisler with the help of “a musician” and “two acoustical engineers”. By far the most important factor was “sound level”, which can be explained because the items were not loudness equalised. Two additional factors appear to have spatial meanings (factor 3: \textit{environmental information} and factor 8: \textit{disturbing directional effects}). Items showing high \textit{environmental information} were said to “communicate a sort of spatial impression, e.g., the placing of the instruments, or, in the case of bird song, that the bird seems to be singing in a courtyard” (Eisler 1966). \textit{Disturbing directional effects} is a negative attribute that Eisler described as the presence of “false spatial impressions” resulting from replay through highly directional loudspeakers. It is entirely plausible (although impossible to verify without further information than is reported), that some of these spatial effects were due to the use of a combination of mono and stereo replay equipment (the factor does appear strongly in the replay technique data and does not seem to be too dependent on programme material). It is also fair to say that both of these \textit{spatial attributes} are not rigorously defined and are doubtless \textit{multidimensional} in their own natures.

In summary, Eisler himself warned about drawing conclusions from his rather ‘rough and ready’ experimental data, and stated that future work would investigate each of the nine factors separately in detail (although this author has been unable to locate any reported follow-up work by Eisler). For purposes of this study, it is noted that multidimensional “spatial” attributes: “\textit{environmental information}” and “\textit{disturbing directional effects}”, were felt to be relatively important in distinguishing between the
reproduction of a multitude of programme materials through a number of different mono and stereo reproduction systems.

### 2.1.2 Nakayama et al. Study

An early investigation that featured spatial sound reproduction was conducted by Nakayama et al. (Nakayama, Miura, Kosaka, Okamoto and Shiga 1971) and concerned the subjective effect of increasing the number of channels of audio reproduction.

In their experiment, recordings of a “band consisting mostly of brasses” (Nakayama, Miura, Kosaka, Okamoto and Shiga 1971) – although the band appears from the supplied photo to include a piano and probably a double bass - were made in a hall using an array of eight directional microphones used such that the centre of the array was at three different positions in the hall: 8.8m, 16.2m and 23.5m away from the centre point of the ensemble. The recordings would be reproduced using an array of eight loudspeakers in an anechoic chamber, positioned at ±15°, ±30°, ±90° and ±150° from the centre front reference position. The microphone array was arranged such that the microphones were placed in the same physical positions with respect to the centre position as the loudspeakers were with respect to the central listening position (pointing away from the centre of the array). Additionally, an “on-mic” recording was made which was effectively a modification of the closest recorded position (8.8m), with the four frontal channels being supplied instead by a two channel M/S microphone, and a stereo mix of an array of microphones placed close to the instruments. Short, six-second clips of the recorded music (which is described as two “popular” music pieces: Brazilian samba and a Japanese folk song) were reproduced in a random order using one of thirteen different configurations of the loudspeakers, ranging from one to eight channels. The designated configurations in the paper show a slight inconsistency in that there appear to be configurations (numbers 1 and 2), that feature a loudspeaker directly ahead or directly behind the listening position. It is therefore likely that ten microphones and loudspeakers were actually used in the set-up and that this has been omitted in the description of the experiment.

Ten subjects (three male, seven female) described as “college students” took part in the tests (five at a time, all seated close to the centre of the loudspeaker array). No explanation is given as to why the subjects were not individually tested (this should have been possible as there were just ten subjects involved). One has to assume that experimental time constraints precluded individual testing of the subjects. Testing in groups of five as reported has obvious perils regarding inter-subject collusion and other interactions which could potentially bias their responses. Unfortunately, no information is left by Nakayama et al. as to the listening experience of the subjects (whether they were experienced or naïve). There is also no indication as to whether a training programme was conducted before testing, so it is possible that the subjects were naïve and possibly untrained – in which case many more subjects should have been used for reliable results (Bech 1992). Subjects were asked to indicate a preference rating for each of the individual replay configurations, and give a judgement of similarity between pairs of the configurations (omitting the mono replay configuration).
Of importance to this work is the fact that the similarity judgement data was analysed using Multidimensional Scaling (MDS) (Nunally and Bernstein 1994), and it was reported that three dimensions could explain the data, which Nakayama et al. labelled:

**Depth of the image sources** (which appears to correspond to the sources “distance” away from the listeners)

**Fullness** is strongest with all loudspeakers active and importantly has larger values whenever the side loudspeakers - those at ±90° from the centre reference position - are active.

**Clearness** (which Rumsey (2001) has suggested could be a measure of D50 -- *Deutlichheit*) -- a measure of the Direct to Reverberant energy about a 50ms cut off point after the direct sound. This author, however, believes that there are too many anomalies to draw this conclusion from the data. ‘Configuration 9’, for example, features four frontal channels at ±15° and ±30° and just two rear channels at ±150° yet has a negative “Clearness”. ‘Configuration 4’, on the other hand, features two frontal loudspeakers at ±30° and two rear speakers at ±150° but has a positive “Clearness”. It must be noted however, that the “clearness” dimension had very little overall effect on the similarity grades.

Regardless of the labels given to the dimensions, regression analysis showed that the dimensions corresponding to **fullness** and **depth of the image sources** were the most important for preference.

To summarise then, Nakayama et al. varied the number of replay channels of a multichannel audio system and found that subjective similarity judgements for the various configurations could be explained in terms of three separate dimensions which they called **fullness**, **depth of the image sources** and **clearness**. **Depth of the image sources** seems to be better explained as a distance perception, and **fullness** seems to be connected with configurations that use the loudspeakers to the side of the listener. It is unclear what the **clearness** dimension refers to, D50 has been suggested (Rumsey 2001), but there are apparent anomalies with the data, hence it could simply indicate a residual noise in the data due to the varying programme materials, recording system distances, and inconsistencies in the (possibly naïve and untrained) subjects’ responses.

### 2.1.3 Gabrielsson’s Sound Quality Tests

In attempting to gain insight into the various facets of perceived sound quality of sound reproduction systems, Gabrielsson conducted a series of studies that he published with others (Gabrielsson, Rosenberg et al. 1974; Gabrielsson and Sjögren 1979; Gabrielsson and Lindström 1985; Gabrielsson, Hagerman et al. 1990), each is discussed below.

#### 2.1.3.1 Monophonic Loudspeaker tests: 1974

In experiments published in 1974, Gabrielsson et al. attempted to obtain a set of perceptual dimensions that underlie perceived sound quality (Gabrielsson, Rosenberg and Sjögren 1974).
Background Theory – Spatial Attributes

In the first of the reported experiments from 1974, Gabrielsson et al. obtained subjective fidelity grades, verbal descriptions and pairwise similarity judgements for a number of monophonic stimuli (five or three) replayed over five different loudspeakers. Multidimensional scaling of the similarity data revealed two underlying dimensions which Gabrielsson et al. interpreted using subjective verbal descriptions of the stimuli.

They stated that the first dimension related to clarity, transparency and directness, with loudspeakers that failed to have this quality being described as confused, clashing, thick or muddy. They surmised that this dimension was affected by the amount of distortion of each system – which we can term a technical quality attribute, rather than a spatial attribute using the nomenclature in (Rumsey 2002).

The second dimension, which involved “brightness - darkness” and “balance between treble and bass”, was explained as correlating to the frequency response of the systems – and is a “timbral attribute” rather than a “spatial attribute” using the terms found in (Rumsey 2002).

Although no spatial attributes were reported, it is possible that they may have been elicited from the subjects but not included in the report. It is also noted here that by supplying five different loudspeakers of different technical qualities and frequency responses a range of monophonic material, variation was achieved along the two major dimensions that describe the main differences between the systems: namely their technical quality and frequency response.

In the second experiment reported in 1974, the effect of changing the level and frequency response was investigated. This time, a single loudspeaker was used to replay versions of three excerpts with either normal level, or attenuated by 6dB, either with a flat frequency response, or with a high frequency boost or cut of 6dB at 10 kHz (it is unclear whether this was a shelving or notch filter. A notch filter is used in later experiments, so it is presumably used here as well). Dimension analysis of subjective similarity ratings showed, unsurprisingly, that the two dimensions underlying the similarity ratings were “brightness, lack of treble” and “loudness” (these correspond in turn to the two major objective measures that were varied). However, even though the reproduction was monophonic, the reported verbal descriptors of the different stimuli included descriptors pertaining to spatial aspects of the reproduction. These were:

“Perceived distance”, “More distant”, “Far away”, “Nearer” (which all appear to pertain to the distance of the sound sources)

“Space of the reproduction”, “Poor reverberation”, “Shut up” (which all appear to pertain to the dimensions and characteristics of the reproduced environment)

“Fullness”, which Gabrielsson and Sjögren later (1979) equated to the concept of the “volume” of tones which Stevens and Davis (1938) explained as “apparent largeness” or “extensiveness”.

In summary, Gabrielsson et al. in their experiments reported in 1974 obtained dimensions that pertained to the actual variables that they were varying: technical quality, frequency response and loudness, which overwhelmed any finer differences between the loudspeaker systems themselves. Importantly, they also gathered verbal response data which showed that, although similarity ratings were based mainly upon the coarser experimental variables, a number of perceptual attributes that described...
the stimuli and the differences between the stimuli were elicited from the subjects. Surprisingly, even though the reproduction was monophonic, a number of these attributes referred to spatial perceptions.

2.1.3.2 Mono Loudspeaker and Stereo Headphone tests: 1979

Gabrielsson and Sjögren (1979) used a different tactic to attempt to explain the multidimensional nature of perceived sound quality. Beginning with a list of around 60 attributes of reproduced sound (which had been reduced from a longer list of 200, provided by a panel including 40 sound engineers), they attempted to find out which were important for sound quality judgements. Subjects were asked to provide ratings of the various attributes (grading each of the stimuli on each of the 10-point attribute scales); similarity judgements between pairs of stimuli (on a scale of 0-100); and to provide verbal descriptions of the stimuli. Experiments were carefully conducted with a familiarisation phase conducted before testing to allow a degree of listener training. To avoid bias, the order in which the attributes were presented in lists for the subjects was randomised. There appears to have been no loudness equalisation attempted, although Gabrielsson and Lindström (1985) later infer that the form of loudness equalisation used in later experiments was also used in these experiments.

The first set of tests involved monophonic reproduction over loudspeakers. Twenty subjects described as “male hi-fi fans” (who may or may not have been experienced listeners) rated five programmes replayed over 9 different “loudspeakers” (six of which were altered versions of one of the high quality loudspeakers tested – either with added distortion or cut or boosted treble or bass). 55 of the provided attributes were rated.

Factor analysis of the correlation between the adjectives revealed four factors that accounted for 90.6% of variance. The first dimension included various spatial attributes; two other dimensions could be described as predominantly timbral, with the fourth dimension concerned with technical qualities of the loudspeakers. The first dimension was interpreted by Gabrielsson and Sjögren as a quality factor emphasising *cleanness/distinctness*, (and importantly) *feeling of space and nearness* in the reproduction. Other spatial factors on the positive side of the scale include *open, airy, feeling of room, full, near, true-to-nature* and *pleasance*. Negative factors include *closed/shut up, diffuse, narrow, distant and dry*. Gabrielsson and Sjögren also attempt to relate this dimension to physical measures, stating that boosting the treble leads to an increase in the dimension, whereas presence of distortion and narrow bandwidth cause a decrease in this dimension. Interestingly the subjective perception “*true-to-nature*” also appears in the predominantly timbral second dimension of “*sharpness/hardness-softness*” indicating the complex nature of the sense of whether a reproduction is *true-to-nature* or *realistic*. Dimension three concerned the timbral perceptions of *brightness/darkness* as well as the technical quality *noisy/rumbling*. Dimension four appears to be purely concerned with technical quality, incorporating “*noise/hissing*”.

In their 1979 paper Gabrielsson and Sjögren also detail experiments conducted stereophonically with headphones. These were conducted in a similar manner to the loudspeaker tests, except that 20 subjects (described as “musicians” – 14 males and 6 females) used a reduced set of 30 attributes to grade five stereophonic programme
items on eight sets of headphones. Of the resulting five dimensions two are of interest here: dimension two, similar to dimension two of the loudspeaker tests, also contained the true-to-nature attribute as well as feeling of presence, but was otherwise predominantly timbral; and dimension five, which features feeling of space vs. closed/shut up. Whether or not the subjects were told what "sense of presence" meant is not detailed in the report. It might have been explained to them, or possibly they may have just used the term as they felt it should be used. Concepts that have been demonstrated to the subjects, or that they are familiar with would be more reliably identified by the subjects themselves.

In summary then, Gabrielsson and Sjögren's 1979 experiments rated various sound stimuli using provided scales which included adjectives describing the spatial nature of sound reproduction. In monophonic loudspeaker tests, a perceptual dimension pertaining to a "feeling of space" and "nearness" appeared to be very influential. The perception of "realism" indicated by the "true-to-nature" attribute appeared both in predominantly timbral and predominantly spatial dimensions. Although a predominantly timbral attribute, "fullness" was likened to the "volume" or "apparent largeness", suggested by Stevens and Davis (1938). "Feeling of space" also appeared in the stereophonic headphone-based tests, but to a lesser degree. Additionally "true-to-nature" appeared in the same dimension as an unexplained variable called "feeling of presence", which may have to do with realism of the sound reproduction.

2.1.3.3 Stereophonic Loudspeaker Tests: 1985

Gabrielsson conducted further research into the perceived quality of loudspeaker reproduction with Lindström (1985). For these experiments, a reduced set of adjectives/attributes was used to rate a number of stereo loudspeaker systems. Of interest to this work, the second of two preliminary tests investigated the reliability of ratings given by subjects to the same loudspeakers and programme materials. Ratings were collected for six loudspeakers over two iterations of the stimuli, and compared to those taken when the same six loudspeakers were used in different situations. During the first repeat the six loudspeakers were rated alongside an additional three loudspeakers with a single iteration of the stimuli over the same amount of time. The second asked for absolute grades to be given for the six loudspeakers which were presented in pairs with a further reduced number of stimuli. The results showed that if the context of the stimuli or the judgement conditions of the experiment are changed, different ratings can be expected for the same stimuli, at least for certain cases and certain subjects.

The main experiment reported in the 1985 paper, however, used a reduced set of seven rating scales to rate 18 commercially available pairs of loudspeakers (two low quality "anchors" were also included in the test, but not in the results). Care was taken in the experimental set-up with the 18 experienced (male) subjects who took part in the test (subsequently reduced to 16 during testing and after post-screening), all taking part in a familiarisation phase before the main test. Eight programme items were used with a form of loudness equalisation where levels were set by two experienced listeners to be "true-to-nature", and hence not loudness-equalised with one another. However, each loudspeaker setup was loudness equalised with the others for each programme item (a method that Gabrielsson and Lindström claim was used in the 1979 tests). Loudspeakers were placed behind an acoustically
transparent curtain and were tested in four groups by three subjects at a time (as mentioned before, testing in groups could unduly affect the results obtained from the subjects). Each group of loudspeakers included the most expensive speaker pair as a reference.

Table 1 shows a summary of the scales used in the experiment, along with specific discussion about each scale.

<table>
<thead>
<tr>
<th>Rating Scale</th>
<th>Definition / Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clarity (clearness/distinctness)</strong></td>
<td>Earlier Gabrielsson et al. stated that “Clarity” was related to distortion of the system (Gabrielsson, Rosenberg and Sjögren 1974), and hence influenced by technical qualities of the sound system.</td>
</tr>
<tr>
<td><strong>Fullness vs. Thinness</strong></td>
<td>Fullness is possibly equivalent to “tonal volume” or “apparent largeness” in (Stevens and Davis 1938).</td>
</tr>
<tr>
<td><strong>Spaciousness (feeling of space)</strong></td>
<td>Defined in their instructions for listeners as follows: “Spaciousness means that the reproduction is spacious, that it sounds open, has breadth and depth, fills up the room, gives a feeling of presence. The opposite is a reproduction that sounds closed, shut up, narrow, without feeling of presence.” (Gabrielsson and Lindström 1985)</td>
</tr>
<tr>
<td><strong>Brightness vs. Dullness</strong></td>
<td>Considered a timbral attribute</td>
</tr>
<tr>
<td><strong>Softness vs. Sharpness</strong></td>
<td>Considered a timbral attribute</td>
</tr>
<tr>
<td><strong>Absence of extraneous sounds</strong></td>
<td>Considered a technical attribute</td>
</tr>
<tr>
<td><strong>Fidelity</strong></td>
<td>The Swedish name used was ”naturtrohet” which, when translated literally means “nature fidelity”. This could have been interpreted as “naturalness” or “true-to-nature” rather than what is traditionally thought of as ”fidelity” (which is more to do with faithfully reproducing the original recording). Fidelity is defined in the instructions for listeners as: “Fidelity refers to the similarity of the reproduction to the original sound... It can be difficult to judge the fidelity if you have not heard the original music, but you must try to imagine how it really sounded. For assistance, you will be given some information about the music and the rooms; otherwise you must rely on your own experience.” (Gabrielsson and Lindström 1985)</td>
</tr>
<tr>
<td><strong>Stereo Impression / Stereo Image</strong></td>
<td>Not tested, but suggested by subjects in (Gabrielsson and Lindström 1985).</td>
</tr>
</tbody>
</table>
confirms that the perceptions reported by subjects are likely to be dependent on the programme material used, and the method of reproduction.

For purposes of this work it is simply noted that subjects appeared to need to use more descriptors for the spatial aspects of the reproduced sound than were being provided by Gabrielsson and Lindström.

In summary, the 1985 experiments showed that rating scale grades could be affected by the context of the stimuli and judgement conditions – sending a cautionary note to future experimental designers involved with subjective attributes of sound quality. They also provided a list of timbral, technical and spatial rating scales that appeared to be an incomplete spatial description of the subjective experience.

### 2.1.3.4 Monophonic Earphone Tests: 1990

Gabrielsson *et al.* (1990) also investigated the effect of altering the level and frequency response of stimuli on various subjective rating scales, this time including *nearness* as well as *fullness/thinness*, *spaciousness* and *fidelity*.

Fourteen subjects (seven male, seven female) were asked to grade three programmes (pink noise, anechoic female speech and Jazz) presented monophonically via stereo earphones, using a total of eight grading scales (*loudness*, *fullness*, *brightness*, *softness/gentleness*, *nearness*, *spaciousness*, *clarity*, *fidelity*). 24 stimuli were creating using four “filters” (flat, +9dB below 200 Hz, +9dB at ~1 kHz, and +9dB at ~4 kHz) with two replay levels (“natural” level and -10dB). Subjects were given twelve practice trials prior to testing.

Although the results are too crude for in-depth analysis, they did give some pointers as to the relationship between certain rating scales and physical measurements of sound. For example, boosting lower frequencies reduced *spaciousness* for all replay levels relative to “flat” (except the jazz at low level – Gabrielsson *et al.* (1990) point to the possibility of the boost in overall level of the filter counteracting the natural tendency for reduction in *spaciousness* with boosts at low frequencies – although further experimentation would be needed to prove this). Regarding *nearness* – apparently all filters made the jazz and noise sound nearer (again, maybe because of sound level boosts due to the filters – but again, further experimentation would be needed to prove this), although the degradations used were too crude to allow for an accurate simulation of source distance cues. Boosting the overall replay level was seen to increase all of the *spatial* ratings scales (*nearness* as well as *fullness* and *spaciousness*).

In summary, the fact that the experiment featured relatively crude and unrealistic degradations made it difficult to deduce meaningful results with regard to how spatial attributes of sound correspond to physical measures. *Spaciousness* can be seen to decrease when elements of the frequency spectrum are boosted, and boosting the overall level seems to increase *nearness* as well as *fullness* and *spaciousness*. It must be expected though, that monophonic earphone reproduction is unlikely to be as capable at creating subjective spatial impressions as multichannel stereo over loudspeakers.
2.1.3.5 Summary of Gabrielsson’s Sound Quality Tests

Gabrielsson and his colleagues conducted a series of experiments into the multidimensional in nature of sound quality, and specified a number of spatial attributes.

The Gabrielsson studies contain warnings about the choice of programme material for listening tests: whatever is used, and, more importantly, how it is used will affect the results obtained, or indeed any attributes elicited from experimentation. Gabrielsson and Sjögren (1979) also warn that their list of attributes may not be exhaustive, or may contain redundancy due to co-varying attributes.

That said, with the use of selected provided scales, sound quality judgements were successfully gathered from a range of subjects on a variety of different programme materials and reproduction systems. A number of what can be deemed “spatial” ratings scales were found to be usable by the subjects including:

- Spaciousness (feeling of space)
- Perceived Distance
- Feeling of Presence / True-to-Nature / Fidelity
- (True-to-nature and fidelity also appear alongside specifically timbral attributes, indicating that these descriptors, along with Feeling of Presence are multidimensional in themselves, containing both timbral and spatial components)
- Fullness (apparent largeness)

Additionally, subjects indicated that a subjective scale for stereo image would be useful to grade stereo reproduction, confirming that the list of attributes suggested was an incomplete description of the spatial scene.

2.1.4 Toole’s Stereo Loudspeaker Tests

Floyd Toole (1985) reported some of the findings of an examination of a series of loudspeakers that included the spatial aspects of sound reproduction, especially pertaining to the differences between mono and stereo reproduction.

Toole had previously stated (1982) that monophonic listening reproduction is capable of evoking subjective quality assessments including presence, distance, openness and spaciousness (these descriptors were mentioned in passing and lacked any pertinent references or definitions). He went on to argue that stereophonic reproduction is in fact “an embellishment of, not a substitute for, accurate sound reproduction” but he did give examples of cases where testing of stereophonic reproduction would be necessary. Although Toole has not published the results of the analysis of the individual spatial attributes that he studied (just overall spatial quality rating data), a study of his experiments and the attributes that he defined serves as a useful background to this author’s work, and is presented below.

The tests consisted of two separate series of experiments, one where scales similar to Gabrielsson’s (see Section 2.1.3) were used to rate various aspects of sound quality. For the second series of tests, Toole separated the concepts of “sound quality” and
"spatial quality". Sound quality scales were a reduced set of the predominantly non-
spatial attributes used in the first series. The spatial quality assessment involved the
subjective rating of a number of spatial attributes that Toole admitted were not
defined as rigorously as those for sound quality, having been suggested as the result
of a series of pilot tests.

The full series of loudspeaker tests involved 42 listeners, evaluated 37 loudspeakers
and took over two years to complete. The main focus of the experiment was to
attempt to control the factors contributing to the personal opinions expressed in
listening tests, and there is evidence of several steps taken in order to achieve this. In
particular, the relative loudness of each loudspeaker (or pair of loudspeakers) was
equalised using pink noise. Scaled rather than rank order data was collected to
attempt to ascertain by “how much” the loudspeakers varied with one another.

The initial series of experiments reported by Toole (1985) involved monophonic
replay (with four loudspeakers being evaluated side-by-side, hidden behind an
“acoustically transparent but visually opaque screen”). The final two test series
involved stereophonic listening which created a specific problem in the test
methodology. Toole noted that the difference introduced when comparing
loudspeaker pairs side-by-side (due to the shifting of the stereo image between replay
pairs) was sometimes greater than was evident due to the actual differences between
the loudspeakers themselves. For this reason, loudspeakers for the stereo tests were
placed on rotating tables that allowed each of the pairs to be positioned in the same
location in the room for evaluation.

2.1.4.1 Mono/Stereo Series I: Spatial Attributes
within an Overall Rating of Sound Quality)

The first of the experiments involving stereo reproduction required subjects to rate
four different types of loudspeaker using ten scales derived from Gabrielsson’s (see
Section 2.1.3.3):

- Clarity / Definition
- Softness
- Fullness
- Brightness
- Spaciousness / Openness
- Nearness / Presence
- Hiss, Noise, Distortions
- Loudness
- Pleasantness
- Fidelity

The scales had defined endpoints and a midpoint that read “midway”, with no other
markings, except the pleasantness and fidelity scales that had eleven divisions,
numbered 0-10. The fidelity scale was labelled “Bad/Poor/Fair/Good/Excellent” at
points 1, 3, 5, 7 and 9 respectively. Two of the scales are spatial in nature:
Spaciousness/Openness and Nearness/Presence. It is interesting to note that
Nearness and Presence were placed in the same scale. Toole seems to use the term
Presence in the way in which it is used in audio engineering, that is, something to do
with the proximity of the sources. The two overall scales of pleasantness and fidelity
may possibly also contain spatial elements, as spatial properties of the reproduction may well account for a more subjectively pleasurable reproduction, or one that appears to be more faithful to the real or imagined “original” sound field.

The stimuli used in Mono/Stereo Series I were eight different three-minute excerpts from commercial recordings featuring a range of musical styles (chamber, symphonic, choral, jazz and rock/pop styles). A description of the experimental methodology includes the fact the items were randomised but repeated an equal number of times on each loudspeaker. The loudspeakers were changed in a randomised order every 5-15 seconds. Results were gathered during monophonic listening conditions (with side-by-side evaluation similar to the initial series of experiments) and in the stereo condition described above.

Unfortunately, the results published only show the results of the “fidelity” grades, not the other attribute scales. Of the four loudspeakers tested, there appear to be three higher quality loudspeakers and one lower quality loudspeaker. The results show that the three higher quality loudspeakers were given similar “fidelity” ratings in the monophonic and stereophonic tests (Toole normalised his data to allow for the differences in the ways in which the subjects used the grading scales to be accounted for). Lower quality loudspeakers, however, increased in “fidelity” dramatically in the stereo ranking, but the variability of grades also increased. Toole argues that this difference in the grades between mono and stereo must be due to the stereo reproduction of the loudspeaker itself, rather than any anomalies due to positional effects during the mono test (because all four loudspeakers were evaluated in the all four positions). Presumably stereo listening was able to ameliorate some of the negative sound qualities of the lower quality loudspeaker. Toole (2004) agreed with an explanation advanced by this author:

“Spatial unmasking ... is a very real part of this. When the components of a spatial 'scene' are spread out, as in stereo compared to mono, some that would have been masked may now be more clearly audible. This is the directionally and spatially rich soundstage compensate(s) for mediocre sound. I think we see the same thing, to an even greater degree, in multichannel presentations - how else to explain the declining interest in truly excellent loudspeakers, and the fascination with cheap and bad 'home theater in a box' systems." (Toole 2004)

Toole’s personal opinions aside, the above is to say that the spatial segregation of elements panned across a stereo image improves for example, the “clarity” of the otherwise lower quality system, whereas the higher quality systems may potentially benefit relatively little from this phenomenon as they already have, for example, relatively good “clarity”. Although this effect would have to be studied further, the above discussion could indicate a co-dependency between certain technical or timbral aspects of perceived sound quality, with spatial modes of sound reproduction. This current project uses the 3/2 stereo reproduction system exclusively, but the issues regarding monophonic reproduction versus “stereophonic” reproduction are expected to have ramifications, as sounds within the 3/2 reproduction system can either be reproduced using a single loudspeaker, or a combination of different loudspeakers.
Toole’s general conclusion from the Mono/Stereo Series I experiments were that the variability of listeners’ grades increased in stereo listening tests versus the mono tests (this author would point out that since monophonic listening tests are more controlled, they would therefore be naturally subject to less variability in their results). Elsewhere in the paper, Toole showed that hearing deficiencies caused an additional increase in variability. He therefore suggested that listeners involved in stereophonic listening should be selected to have especially “good” hearing.

### 2.1.4.2 Mono/Stereo Series II: Spatial Quality (Separated from Sound Quality)

In the second series of experiments involving loudspeaker reproduction, Toole separated “*spatial quality*” from “*sound quality*” into a series of separate scales which were designed to embrace comments from listeners during pilot tests. As an indicator of their success at adequately characterising spatial quality, Toole referred to the lack of additional written comments left by users rating spatial quality (subjects had provided comments on the other sound quality scales, which Toole had used to refine said scales). This author would argue, however, that subjects who lack experience of using a descriptive language for spatial audio reproduction may be less inclined to leave comments or suggestions regarding an area in which they have less expertise. Without further research, answers to questions on the scales’ validity are speculative.

In the experiments, three loudspeaker pairs (two that were considered to be of high quality, one of average quality) were evaluated using repeated assessments of the loudspeakers for particular features. Monophonic and stereophonic tests were conducted, with similar procedures for each (monophonic tests were conducted using the left speaker of the stereo pair to avoid loudspeaker position becoming a variable between the two conditions). The ten subjects employed in the test were described as “audiophiles” with “essentially normal hearing” (Toole 1985). Just four programme materials were used recorded in various ways: a choir recorded using multiple microphones in a concert hall, chamber music recorded using a coincident stereo pair, a small, jazz ensemble recorded with multiple close microphones, and a popular music track which had been “given the full treatment of signal processing for special spatial and spectral effects” (although no specific information was supplied). High quality recordings were considered essential for these experiments, as Toole (1985) considered that it was possible for “background hiss to be associated with ambiance”.

The scales that Toole used in these experiments were grouped together as *sound quality* and *spatial quality*. The *sound quality* group contained many of the scales used in the previous tests, derived from Gabrielsson’s work (see Section 2.1.3.3), however *spaciousness/openness* and *nearness/presence* were removed, along with *loudness*. It is unclear why *loudness* was removed – presumably the loudness equalisation strategy employed in the first stereo experiments helped to reduce loudness variation sufficiently for its inclusion as a test variable to be deemed unnecessary for the second series. Toole had presumably classified *spaciousness/openness* and *nearness/presence* as spatial attributes, justifying their removal from the “*sound quality*” scale set.

It is interesting to note that the term *fullness*, which Gabrielsson (see Section 2.1.3.2) had indicated may have incorporated the concept of *spatial volume*, is also used here.
Toole, however, defines fullness in a way that avoids references to any spatial aspects of the term: “Fullness: Refers to the quantity of low-frequency sounds and their balance with respect to the middle- and high-frequency sounds. Good sound should be neither too full nor too thin” (Toole 1985). The possibility that changes in the content of mid- and high-frequencies could result in the perception of differences in source size is possible, but this could have resulted in a correlation between fullness and a potential “source size” scale (the closest spatial scale that Toole used was width of the sound stage).

As well as an overall spatial rating scale, Toole (1985) included spatial quality scales shown along with the full definitions shown in Table 2.

| Table 2: Showing Rating Scales used in the Mono/Stereo Series II experiment in (Toole 1985) |
|-----------------------------------------------|-----------------------------------------------|
| **Spatial Quality Scale** | **Definition / Explanation** |
| Definition of the sound images | Refers to the extent that different sources of sound are spatially separated and positionally defined. Images should not move as the pitch of the music rises and falls. The size of the image should be appropriate to the source of the sound. |
| Continuity of the sound stage | Is the display of sound images continuous, left to right, or are there illogical groupings of images, with large gaps in between? Is the reverberation uniformly displayed or is it concentrated in strange places? |
| Width of the sound stage | Refers to the left-right display of sound images. The response scale represents the one in front of you in this room. Mark on it the left and right limits or boundaries of the sounds you hear. Do not include vague reverberant sounds, only those of the orchestra. |
| Impression of distance or depth | Should be judged on the basis of a satisfactory impression of instruments at various distances. An unsatisfactory reproduction would have all of the instruments at one distance (two-dimensional), or some of them too close or too far, and so on. |
| Abnormal effects | Refer to spatial sensations that do not occur in common experience. For example, it is possible for some sounds to appear to stretch between you and the screen, perhaps even some of the sounds will appear inside your head. Other sounds may appear to have no location, when you know the instrument should be precisely localized. |
| Reproduction of Ambiance, Spaciousness and Reverberation | Not defined by Toole. |
| Perspective | Refers to your general impressions of the experience. A good reproduction of a good recording with natural room or hall acoustics should suggest that “you are there” at the performance, complete with a sense of the enveloping ambient sound. A less perfect reproduction could separate you from the performance, giving the impression that you are “close, but still looking on.” In a still worse reproduction it may seem that you are listening through an opening between the loudspeakers. It is as though you were “outside looking in” - there is no impression of being within the ambient sound. Other recordings may appear to transport the musicians to the listening room, “they are here.” The ambiance is that of the listening room, and the instruments sound close. Still other recordings are created as abstract special effects, with no attempt to simulate a realistic experience. |
It can be argued that every one of these scales is multidimensional (except perhaps width of sound stage), in that they encompass more than one perceptual or logical notion. Examples of this multidimensionality are relatively easy to spot: Definition of the sound images for example, contains descriptors about the stability of sources as well as the congruence of their size within the image. Continuity of the sound stage pertains not only to the distribution of sources across the soundstage, but also the dispersal of reverberation. Impression of distance or depth not only contains the terms distance and depth without defining how they differ from one another, but also allows for the potentially problematic situation where instruments are not at the same distance from the listener (i.e.: not “two-dimensional”, which is a “good” thing as far as impression of distance or depth is concerned), yet some of the instruments could also be deemed too close and/or too far away (which is a "bad" thing as far as impression of distance or depth is concerned). The definition of abnormal effects contains a number of different examples of these effects which may prove difficult to separate from one another in the analysis as the scale simply asks for a judgement as to the frequency of any such effects, once the data is collected. The spatial quality scale perspective is certainly multidimensional. On the one hand it seems to be an overall rating similar to the overall spatial rating scale (which is itself supposed to be the sum of the spatial quality scales), with subjects seem to be required to provide a rating of how they perceive the spatial fidelity of the reproduction to be. Spatial fidelity is a term which this author has used as a convenient term to describe the various changes that occur in widely varying stimuli and replay configurations (Zieliński, Rumsey, Kassier and Bech 2005). On the other hand it appears to have strong connotations of something to do with how real the illusion that is projected by the loudspeakers is. Looking at it yet another way, perspective seems to imply elements of the distance of the reproduction from the listener, or whether the listener is surrounded by the sound and feels involved. The perspective scale is labelled in such a way to allow a sliding scale with three labels: “you are there”, “close, but still looking on”, and “outside looking in”. It also featured three separate response categories: “they are here”, “artificial, contrived” and “other, describe”. The use of the three labels on the scale seems to recall a description of the lack of spatial responsiveness reported as “the feeling of looking at the music” in (Marshall 1967). This was later reworded in (Barron and Marshall 1981) to read “the sensation of spatial impression corresponds to the difference between feeling inside the music and looking at it, as through a window”, which recalls an earlier article by Voigt (1950) who also uses the imagery of observing reproduced audio as if through an open window.

It would have been interesting to see the correlation between perspective and overall spatial rating to see whether “you are there” corresponded to “excellent” overall spatial grades and to see what low overall spatial ratings corresponded to in terms of perspective. It is likely that these multiple, potentially confounding dimensions within each of the above scales could cause perplexing results for any experimenter attempting to ascertain how the individual perceptual dimensions are affected by external influences.
2.1.4.3 Summary of Toole’s work on spatial attributes

A general summary of the spatial attributes used by Toole in his investigations into the spatial qualities of stereo loudspeaker systems could indicate that he was concerned with the definition and localisability of sources, the homogeneity, stability and width of the stereo image and the illusion of auditory perspective (regarding sources appearing closer or further away from others). There is also an attempt to define a realistic reproduction, including rating schemes that allow subjects to judge the degree of reproduced reverberation and their feeling of involvement in the reproduction.

That Toole did not present any results from the individual spatial attributes that he studied is unfortunate. He was mainly interested in loudspeaker performance at the time and he notes that “it became clear early on that (spatial quality) was a factor much more controlled by the recording than it was by the speaker” (Toole 2004). Additionally, multivariate analysis of the attribute data (without the aid of computers at that time), was considered to not be cost effective.

The results presented for Mono/Stereo Series II are based around the “overall spatial rating”. It was found that the fidelity ratings (which Toole labelled as being equivalent to sound quality ratings) of the two high quality loudspeakers were similar in both monophonic and stereophonic tests. The fidelity of the poorer quality loudspeaker jumped from around 5 to around 7 on the 0-10 scale, just 0.2 rating points lower than the higher quality loudspeakers (Toole published normalised data). This is similar to the behaviour of the anchor in the Mono/Stereo Series I experiment. There were, however, remarkable improvements in the overall spatial rating scales for stereophonic reproduction when compared with monophonic reproduction. During monophonic tests, the overall spatial rating for the higher quality loudspeakers was rated on the 0-10 scale at about 6.4 and 6 respectively, with the poorer quality loudspeaker achieving around 4.7. The overall spatial rating of all loudspeakers was rated equivalently on average, at around 7 on the 0-10 scale. Regarding the individual scales, Toole reported that subjects seemed to have little difficulty in using the width of sound stage, impression of distance and depth, abnormal effects and reproduction of ambiance, spaciousness and reverberation scales (which, this author would point out, is not to say that they were using them correctly as no evidence was given that training was provided on how to detect and discriminate between the spatial phenomena that pertain to the scales).

Toole argues that “fidelity” ratings are a good predictor of spatial quality, as they were not significantly different from the “overall spatial rating” grades, pointing out that loudspeakers of high sound quality seem to also have good spatial quality. Stereo listening can, however, also improve the perceived spatial and sound quality of otherwise lower quality items. This author would argue that similarity between “fidelity” (supposedly a non-spatial scale) and “overall spatial rating” could just as easily indicate that “fidelity” ratings include a large “spatial” component. Again, this cannot be proven without the individual scale data.

A concern that this author has about the experimental technique used in these experiments is that there were only three pairs of loudspeakers tested. It was Toole
himself that suggested that “it may be advisable to ensure that listeners are exposed to a somewhat standardized range of sound qualities” (Toole 1982). He goes on to suggest that loudspeakers of known quality be included in tests, as these will define “anchor” points on subjective scales “a group of ‘good’ test products may need some ‘poor’ anchors, and vice versa” (Toole 1982). This author believes that the range of test items included in Mono/Stereo Series II was not wide enough, and even Toole comments that, for the correlation of overall spatial rating with fidelity, “the data are so closely clustered” (Toole 1985) that trends may be misleading.

Toole (1985) presented charts of the averaged data for each of the spatial scales averaged across all program materials. He commented that only the abnormal effects scale seemed to be rated in significantly different ways between loudspeakers. Toole (1985) also reported that subjects noticed a distinctive abnormality of the lower quality loudspeaker system which “mainly consisted of the illusion of sounds originating close to the listener’s head or center images far forward of the remainder of the sound field”. If this really was the case, would it not seem logical to also see the effect of this anomaly in the depth of sound stage rating? From the mean of the averaged data presented in the charts for the depth of sound stage scale, this did not appear to be the case, although there are more “poor” ratings for the lower quality loudspeaker than for the other two. In a later paper Toole (1986) does comment that this loudspeaker was occasionally described as “presenting a narrower sound stage ... and a less satisfactory rendering of depth and ambience” (Toole 1986), but this did not show up in the results presented in (Toole 1985). Toole does show that the lower abnormal effects rating achieved by the lower quality loudspeaker is programme dependent. Overall spatial rating results averaged for all subjects showed that the lower quality loudspeaker was rated consistently low during the pop music item, which corresponds to the monophonic fidelity results for that loudspeaker. Toole hypothesises that this could be due to problems with the loudspeaker systems reproduction of the panned mono sources in the pop recording equating to the performance of the loudspeaker in mono. In a later paper, Toole (1986) points to the large tweeter unit of the loudspeaker being responsible for an increase in directionality of the loudspeaker at high frequencies which could have caused these effects. An ANalysis Of VAriance (ANOVA) test may well have uncovered this and other interactions between programme type and loudspeaker system, but this test was not performed, and the individual data were not presented.

In his summary, Toole postulates that stereophonic reproduction was rated higher than monophonic reproduction due to the fact that decorrelated noise and distortion, which in stereo is perceived as a large ‘noise ‘image’ resembling, in some ways, well-recorded reverberation” (Toole 1985), is perceived as more objectionable when summed together in mono.

To summarise, it is likely that the use of scales that contained many potentially confounding dimensions is likely to have created results that are difficult to analyse. The spatial scales used provide a useful record of what Toole and his subjects considered to be the salient spatial qualities of monophonic and stereophonic sound reproduction. There is also no evidence that Toole’s subjects were trained to use the scales using examples that exemplify aspects of the qualities in question, and this may have led them to rate the various spatial scales in a similar manner to the overall spatial quality rating scale.
2.1.5 Letowski’s MURAL

The MURAL (MUltilevel auditoRy Assessment Language) published by Letowski (1989) was a systematic attempt at defining the constituent parts of what had, until then, been loosely termed “sound quality”.

In a summary of the various scales by which sound quality was judged, Letowski pointed out that various terms were being used in different reports, and that different subsets of these terms were being used to mean the same thing by different authors.

Letowski pointed out that others have referred to sound quality interchangeably as timbre, tone color, sound color, timbral color, spectral color and spectral timbre. He argued that it was misleading to equate an overall quality judgement (which implies an emotional grading) with timbre or other spectral scales (which imply a neutral judgement of a characteristic of the sound). He would later show that, so long as the entire scheme is consistent, the various characteristics of sound can be rated either qualitatively (along hedonistic or preference scales), or quantitatively (along dominance or similarity scales).

Another problem that Letowski highlights with the concept of equating sound quality to timbre is that “an impression of spaciousness” is also part of sound quality. He suggests that, as a first approximation, sound quality could be described as being the combination of two multidimensional sensations: Timbral Quality and Spatial Quality.

Interestingly, Letowski also uses the term naturalness in his argument – naturalness was also used by Gabrielsson (see Section 2.1.3.3). He suggests that this is used as a scale to rate overall sound quality, which would therefore mean that naturalness incorporated both timbral and spatial aspects. He suggests, however, that unfamiliar signals should be rated on pleasantness not naturalness scales, as the naturalness scale requires knowledge of an external reference (one that is intended to be the ideal reference against which judgements of naturalness are made), whereas pleasantness is a hedonic scale that could be graded against a number of internal standards (internal references would presumably be defined by the subject and held in memory. Hedonic (“liking”) ratings of pleasantness could then be made for sound images against these references). This infers that naturalness could be considered to be a judgement (because of the reference to an external standard), whereas a rating of pleasantness is likely to be predominantly an emotional response to the stimulus.

Letowski (1989) states that “sound quality cannot be sufficiently well described by a global assessment alone”. Because of its multidimensional nature, he proposed that assessments of the multiple dimensions (what he terms parametric assessments) are required to describe sound quality. Although several assessment parameters had been proposed, Letowski felt that the various systems of sound quality assessment were not compatible with one another, and that they typically involved a mixture of qualitative and quantitative scales.

The MURAL is a set of timbral and spatial characteristics that Letowski arranged within a hierarchical system. Different levels of the MURAL were linked in various ways, with characteristics falling ultimately within (or sometimes between) timbral or spatial subsets. Characteristics that were displayed on the same level, Letowski claimed, were “fairly independent and complimentary”. By this, Letowski presumably meant that the characteristics within a specific level described separate
perceptual concepts that together formed a more-or-less complete description of the overall sound quality.

Table 3 shows a version of the MURAL displaying the various characteristics of sound quality that Letowski suggested as a result of “earlier research on sound quality assessment”. Sound characteristics within the same hierarchical band are displayed in the columns. Characteristics that overlap vertically between different columns show a hierarchical link.

<table>
<thead>
<tr>
<th>Auditory Image</th>
<th>Table 3: Letowski’s MURAL, after (Letowski 1989)</th>
</tr>
</thead>
</table>
| **Timbre**    | **Richness**
|               | **...Presence (Nearness)**
|               | **Powerfulness**
|               | **Sharpness**
| **Timbre Balance** | **Brightness**
|               | **Coloration**
| **Distinctness** | **Blend (Compactness)**
|               | **Clarity of Sound Texture**
|               | **Directional Selectivity**
| **Stereophonic Impression** | **Panorama**
| **Reverberance (Liveness)** | **Ambience (Feeling of Air)**
|               | **Perspective (Depth)**
|               | **Presence (Nearness)**...

Letowski did not leave much in the way of justification for his selection of scales, or indeed any definitions for the scales that were included. It is possible though, to examine the many links that exist between the hierarchies to gain further insight into their meaning. For example the lowest level characteristic presence (nearness) is unique in being a part of a timbral characteristic richness and a spaciousness characteristic reverberance (liveness).

Examining the spaciousness hierarchy of the MURAL, various similarities with previously quoted studies become apparent:

**Distinctness**

*Directional Selectivity*: presumably pertaining to the ease with which subjects can laterally locate the source – similar to Toole’s *definition of the sound images* quality scale (see Section 2.1.4).

**Stereophonic Impression**

*Panorama*: similar to Toole’s *continuity of the sound stage* quality scale (see Section 2.1.4).

*Ambience (Feeling of Air)*: similar to Gabrielsson’s *spaciousness (feeling of space)* (see Section 2.1.3) and Toole’s *Reproduction of Ambiance, Spaciousness and Reverberation* quality scale (see Section 2.1.4).
Background Theory – Spatial Attributes

Reverberance (Liveness)
Ambience (Feeling of Air) (see above)

Perspective (Depth): similar to “depth of image sources” from Nakayama et al. (see Section 2.1.2) and Toole’s “Impression of distance or depth” quality scale (see Section 2.1.4).

Presence (Nearness): similar Toole’s “Impression of distance or depth” quality scale (see Section 2.1.4).

It is interesting that Letowski separates “Presence (Nearness)” which presumably refers to the sound image’s distance from the listener, from “Perspective (Depth)” which probably pertains to the perception that different sources are at different distances from the listener. It is unclear why the “Presence (Nearness)” characteristic overlaps with the timbre characteristic “Richness”.

In the paper, Letowski reported that a weighting system that would assign degrees of importance to the various scales for various applications was “under development”. In his summary, he stated that “The future success of this system depends on the development of clear and univocal definitions of all component sensations and on the identification of reliable perceptual weights for selected applications” (Letowski 1989). Unfortunately Letowski became involved in other projects, so did not pursue the weighting scheme for the MURAL “much further”. Regrettably, no other work on the weighting scheme has been undertaken by anyone else (Letowski 1989).

2.1.6 Zacharov & Koivuniemi studies

A large-scale study into the perceptual attributes that influence the subjective preference of spatial sound reproduction systems was undertaken by Zacharov & Koivuniemi (2001a,b,c) and Koivuniemi & Zacharov (2001). They proposed that an understanding of the salient perceptual attributes of spatial sound reproduction systems, including how they relate to preference judgements, would be useful for the improvement or perceptual optimisation of such systems (Zacharov and Koivuniemi 2001c).

Zacharov & Koivuniemi explained that previous studies into spatial sound attributes showed similarities but also differences due to individual factors within each study. Their investigation would study the impact of sound event, source and room acoustics, microphone technique and sound reproduction technique on spatial sound perception.

The study consisted of a number of phases: stimuli creation; naïve subject preference collection; direct attribute elicitation; direct attribute rating and the mapping of naïve preference to direct attribute ratings.

2.1.6.1 Test stimuli

A wide variety of stimuli “representative of a wide range of acoustics” (Zacharov and Koivuniemi 2001c) were recorded using an ingenious multiple microphone setup allowing simultaneous capture of the same acoustic events in a number of different recording formats. Thirteen different spatial scenes were recorded, including six stimuli created by replaying male speech and guitar samples over a loudspeaker in three different environments (anechoic chamber, standard listening room and a
Background Theory – Spatial Attributes

The other seven stimuli were recorded live and in situ at a number of locations, including outdoor and indoor scenes from daily life, and music concerts. The scenes were described as including motion (both horizontal and vertical), reverberation, envelopment, distance cues, indoor and outdoor acoustics, and spatially distributed directional sources (Zacharov and Koivuniemi 2001c).

The stimuli would be replayed in a standard listening room (ITU-R 1994-1997), using one or more of fourteen loudspeakers arranged in various formats around the listening position. Replay formats included a 3-2 stereo set up, an eight-channel periphonic cube, various uses of the standard two-channel stereo reproduction at ±30° (using two of the loudspeakers from the 3-2 stereo array), and Monophonic (at 0° and 90° to the right of the centre front reference position, utilising the centre loudspeaker of the 3-2 stereo rig, plus an additional loudspeaker at 90°). Each of the replay formats had used some combination of direct and/or synthesised microphone signals from the multiple microphone recordings. One of the stimuli was chosen as a reference, which was a simulation of a stereo microphone system consisting of crossed dipole directivity pattern microphones. This reference was chosen because it “was considered to provide a neutral reference, both in terms (of) spatial and timbral qualities” (Zacharov and Koivuniemi 2001c). It is worth noting that Zacharov & Koivuniemi also describe sound quality as consisting of timbral and spatial aspects.

All stimuli were loudness equalised using a loudness meter that took account of directional loudness by employing a binaural recording system at the listening position (Tuomi and Zacharov 2000).

There is no reported evidence that loudspeakers were hidden from the view of the subjects as recommended by Gabrielsson (see Section 2.1.3) and Toole (see Section 2.1.4). Zacharov confessed that the ideal would have been for the speakers to be hidden, but that this had not been done because it was found to be “tricky to hide all of the speakers” (Zacharov 2004), explaining that because of the multitude of loudspeakers in the listening room, it would have been difficult for the subjects to guess which loudspeakers were active (Zacharov 2004). This strategy led to problems with some of the definitions used later on in the experiment. For example, the definition of the attribute sense of depth defined in (Koivuniemi and Zacharov 2001) uses the position of the transducers as a reference, which implies that the loudspeakers needed to be visible to the subjects. Subjects would therefore either need to be told which loudspeakers were active for any given stimulus, or they would need to be relied upon to guess which ones were active when rating the stimuli. Whichever course of action was used may have introduced additional variance into the resultant data. A way of clarifying this situation would have been to loop an acoustically transparent curtain around the listening position, visually isolating it from the loudspeakers for the entire experiment. It is unlikely that the visual position or orientation of the loudspeakers would have entered the descriptive language elicited in the experiment, as this information would not have been available to the subjects during the elicitation phase.

2.1.6.2 Naïve preference test

The first experiment consisted of a test to collect preference data for each stimulus from subjects who are considered naïve: “subjects must be naïve and untrained in order for them to be able to provide integrative grades (e.g. overall quality or
preference scores, where all the [perceived] characteristics are integrated into a single number grade)” (Zacharov 2004). In the experiment, sixteen subjects (no evidence is provided by Zacharov & Koivuniemi to suggest that sixteen subjects would be sufficient for the type of experiment conducted) were given the task of rating their preference between each stimulus and the reference on a ±10 point scale. Although they were described as naive, they were trained to use the interface, and familiarised with all the samples before the test. Subjects were asked to consider any timbral features of importance, but to focus on the spatial aspects of each stimulus.

2.1.6.3 Direct attribute elicitation

The next stage of the experimentation was to elicit a set of attributes of the data from a set of “selected and trained” subjects (Zacharov and Koivuniemi 2001c). It is not made clear in the report, but Zacharov later revealed (2004) that in fact the same sixteen subjects were used in this phase of the experiment that were used in the previous naive preference test. Zacharov (2004) explained that ideally the two sets of subjects should not overlap, with expert subjects chosen for their “expertise and capability” used in the elicitation stage. Whilst it may have been difficult to engage sufficient numbers of subjects for these experiments, the cross-contamination of these subject populations is not ideal and the conclusions of the experiment need to be treated with caution. It was Zacharov’s hope that the naive subjects would gain experience from listening to the stimuli a number of times over tens of hours. The group was reduced from sixteen to twelve members based upon their ability to describe differences between visual images (visual imagery was used in the selection process so that the audio elicitation phase was not pre-biased).

The language development stage was supervised by Koivuniemi who was kept as blind to the experimental objectives as possible, so as to help reduce any personal biases that the experimenters might introduce. In a programme that lasted a number of weeks, subjects listened initially individually, then in groups of four to all the stimuli. Crucially, they were asked to “write down every spatial sentiment” (Koivuniemi and Zacharov 2001) that they perceive when listening to the stimuli. This emphasis on sentiments rather than judgements means that the elicited constructs are more likely to be influenced by personal biases (Nunally and Bernstein 1994) than if the subjects had been asked to provide judgement-orientated constructs. This author believes that the attribute scales elicited from this experiment reflect this initial emphasis on spatial sentiments.

The full elicitation experiment used both absolute and differential elicitation with single and multiple stimulus comparison paradigms respectively. Subjects were encouraged to use both absolute and relative differences to formulate a set of adjectives and synonyms to describe the stimuli. The resulting 1400 attributes (in the Finnish language) were computationally reduced to 532 attributes which were further reduced to twelve attributes after a series of the group discussions. These attributes along with the defined end points of the scales are shown in Table 4.

Koivuniemi & Zacharov (2001) did not make great claims as to the set of attribute scales elicited, describing it as a “first step” and that “new rounds of language development, and a new set of sound samples are necessary to build the attribute list”. It is worth noting that Koivuniemi & Zacharov refer to the dependence of the elicited attributes to the actual sound samples used. Many of the previously
summarised studies involved stationary sound sources, and an immediately noticeable difference with the Zacharov & Koivuniemi study is that there is a unique attribute involving sense of movement. This can be explained by the inclusion of dynamic stimuli in the experiment.

<p>| Table 4: Spatial and Timbral Attributes Elicited, after (Zacharov and Koivuniemi 2001a) |
|---------------------------------|---------------------------------|---------------------------------|---------------------------------|</p>
<table>
<thead>
<tr>
<th>Type</th>
<th>Attribute</th>
<th>Negative Anchor</th>
<th>Positive Anchor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial Attributes</td>
<td>Sense of direction</td>
<td>Ill-defined</td>
<td>Well defined</td>
</tr>
<tr>
<td>Sense of depth</td>
<td>Ill-defined</td>
<td>Well defined</td>
<td></td>
</tr>
<tr>
<td>Sense of space</td>
<td>Ill-defined</td>
<td>Well defined</td>
<td></td>
</tr>
<tr>
<td>Sense of movement</td>
<td>Ill-defined</td>
<td>Well defined</td>
<td></td>
</tr>
<tr>
<td>Penetration</td>
<td>Non-Penetrating</td>
<td>Penetrating</td>
<td></td>
</tr>
<tr>
<td>Distance to events</td>
<td>Close</td>
<td>Distant</td>
<td></td>
</tr>
<tr>
<td>Broadness</td>
<td>Narrow</td>
<td>Broad</td>
<td></td>
</tr>
<tr>
<td>Naturalness</td>
<td>Unnatural</td>
<td>Natural</td>
<td></td>
</tr>
<tr>
<td>Timbral Attributes</td>
<td>Richness</td>
<td>Thin</td>
<td>Rich</td>
</tr>
<tr>
<td>Hardness</td>
<td>Soft</td>
<td>Hard</td>
<td></td>
</tr>
<tr>
<td>Emphasis</td>
<td>Neutral</td>
<td>Emphasised</td>
<td></td>
</tr>
<tr>
<td>Tone Colour</td>
<td>Dark</td>
<td>Bright</td>
<td></td>
</tr>
</tbody>
</table>

Spatial attribute scale descriptions (translated from the original Finnish) from (Koivuniemi & Zacharov 2001) follow, along with a short discussion of each one with reference to this work:

**Sense of direction:**

"Describes how easily the locations of events can be discriminated. This also measures whether several sound sources can be distinguished. A negative value of this attribute implies that the location of a sound event is ill-defined or enveloping" (Koivuniemi & Zacharov 2001).

The use of the term *enveloping* is used here to imply that an *enveloping* sound does not have a clear location or direction. *Sense of direction* seems to apply more to the ability of subjects to be able to locate a sound at a specific location (its *locatedness*?), rather than to define any positional changes between stimuli (which could probably have been termed *source direction* rather than *sense of direction*). An ambiguity arises in this definition when considering monophonic reproduction: it could be that given sources are positioned at a coincident point in space (a reproduction loudspeaker for example), and it is easy to "locate" them there, but perhaps difficult to discern between them because they are spatially coincident. Classification of the scale in terms of a *sensation* rather than a *judgement* has led to this rather vague definition.

**Sense of depth:**

"Describes how strongly the sensation of distance is perceived, or how ambiguous the sensation of distance is. Once again this assesses whether several sound events can be discriminated in terms of distance. A negative value could mean that distances for all events are ambiguous except those originating from the transducer's position" (Koivuniemi & Zacharov 2001).
The *sense of depth* attribute is defined in terms of the perception of a difference in the distances to the various sound events. It is interesting to note that *distance to events* also appears as a spatial attribute in this study. This shows that the subjects considered not only the distance of the sound sources to the listener as being important, but also the perceived distance between sources. The difference between these concepts was highlighted in (Loomis 1995), where he describes the terms *egocentric distance* (a distance to an event measured radially outward from the listener) and *exocentric distance* (a distance between two different sound events). Using Loomis’ terminology, *sense of depth* implies that *exocentric* distances exist between the sources, because the sources are perceived at different *egocentric* distances. As mentioned above, the position of the transducers is problematically incorporated into the definition of *sense of depth*.

**Sense of space:**

“This attribute scales how well the space where the recording was made is perceived. A positive value could mean a strong sensation of being in a certain kind of environment, e.g. in a room” (Koivuniemi and Zacharov 2001).

The way this definition is constructed implies that it is more about the realism of the illusion provided by the system under test than the perception of changes in the *space* of the reproduced environment. This definition is likely to have resulted from the direct elicitation of *sensations*, rather than the description of differences between the two reproduction environments. This author proposes that attribute definitions based on the ability to quantitatively judge differences between stimuli would be of more use when analysing how to optimise the spatial performance of sound systems.

**Sense of movement:**

“This describes whether a sound source is perceived to actually move in the sound space. A negative value could indicate a sound source simply disappearing from its original location and reappearing in another without moving through any intermediate position” (Koivuniemi and Zacharov 2001).

Unique to this study, the inclusion of *sense of movement* is most likely to be the result of the inclusion of dynamic sources within the stimuli. The definition provides a negative example where the movement of the source is not smooth, but does not give other examples. If dynamic sources are to be adequately described and discriminated between, what about descriptors for their trajectory and speed? Do stationary sources have a negative *sense of movement*? Whilst movement of sound sources is a part of the experience of naturally occurring sound fields, this author believes that the inclusion of rigorously defined moving sources will increase the complexity of the task unduly. It would therefore be wise to concentrate on stationary scenes and sources for any initial study, and include moving sources as an extension at a future date.

**Penetration:**

“Describes the sensation often found in cross talk cancelled binaural reproduction. A positive value means that spatial information in the sample seems artificial. The sounds sometimes seem to originate very close to, or even inside, ones head” (Koivuniemi and Zacharov 2001).
The definition for *penetration* is another that seems to have been influenced to a great extent by the nature of the stimuli, referring to the specific technique that produces this sensation. The terms that appear in the definition describe distance (*close to, or even inside one's head*) and a sense of naturalness or realism (*artificial*). It seems to this author that the sensation of *penetration* could be described in terms of two of the other attributes: *distance to events* and *naturalness*. As the term *penetration* seems to apply mostly to just one of the reproduction techniques, it is difficult to justify its inclusion in a generalised scheme.

**Distance to events:**

“This attribute simply describes the actual distance where the sound events appear to originate. A positive value implies that the sound sources are sensed to be far from the listening point” (Koivuniemi and Zacharov 2001).

This is not the first time that the proximity of sound sources features as a salient perception of reproduced sound. Using the terminology proposed in (Loomis 1995) *distance to events* denotes *egocentric* distances. It is unclear with this definition exactly which events are to be used in grading. For example, what about a relatively close conversation between two people observing an event such as an outdoor fireworks display – would *distance to events* be graded low (for the conversation) or high (for the “main event” – the distant fireworks)? Because the stimuli varied so widely it might be difficult to compare *distance to events* grades across the various stimuli because they all featured different stimuli at different distances. However, if *distance to events* for each stimulus type was graded consistently it might allow the different replay formats to be compared. This author expects that controlled stimuli that use consistent numbers of sources could be used to exemplify source distance in a training regime.

**Broadness:**

“This attribute describes how wide an area the perceived sound event seems to have. A strong positive value would mean that sounds are coming from all around the listener i.e. envelope the listener” (Koivuniemi and Zacharov 2001).

The definition of *broadness* also includes the concept of *envelopment*, implying that enveloping stimuli are *angularly wide*, subtending an arc about the listening position. Here no distinction is made between the concepts of *source width* and *envelopment*, terms used to describe separate perceptual phenomena in concert hall acoustics (Beranek 1992).

**Naturalness:**

“This naturalness describes how well the perceived events conform to what the subjects consider as realism. Perception of something that isn’t possible in reality yields a negative value, e.g. a train rising straight up” (Koivuniemi and Zacharov 2001).

Koivuniemi & Zacharov’s definition of *naturalness* includes the term *realism* which implies the degree of success of the illusion of the intended spatial sound field. The example given, however, implies that a *natural* and *realistic* sound should *conform to reality* (which in turn implies that the intended sound field must be able to exist in reality). This author believes that the two concepts *realistic* and *conforming to reality* are subtly different. For example, it is possible to conceive of a totally
Background Theory – Spatial Attributes

convincing auditory rendition of a mythical creature (such as a dragon), or an extinct species (such as a dinosaur), which would not exist in the reality of the subjects, but could nevertheless be a convincing illusion. This author would therefore suggest removing the concept of conforming to reality from any definition of an attribute pertaining to naturalness or realism.

2.1.6.4 Training stimuli

Zacharov & Koivuniemi next decided to create a set of stimuli to train and familiarise the subjects in the use of the rating scales. They felt that this was necessary in order to train subjects to listen for specific attributes, as well as serving as a useful method of training new listeners or refreshing the memories of existing subjects. Koivuniemi & Zacharov (2001) stopped short of creating definitive scale anchors, creating instead a set of stimuli that “indicate” the attribute and polarity of the scale, supplemented with a verbal description. They explained that it is difficult to isolate and scale a single attribute due to “co-linearities” that exist. The example given was that of loudness: “As loudness increases, so the pitch of a sample may shift” (Koivuniemi and Zacharov 2001). It has been pointed out in (Neher 2004) that this is not ideal as subjects might become confused when presented with examples that contain a number of different spatial attributes.

The development of the training samples was described as “an ongoing process of improvement” but no additional research has been published to date on its effectiveness.

2.1.6.5 Direct attribute rating and analysis

The remainder of the reports detail an experiment in the rating of the stimuli along with analysis of the data.

The attributes were rated for each stimulus by the same twelve subjects that elicited the scales. The tests used the single stimulus test paradigm, with two direct attributes being evaluated at the same time.

Analysis of the direct attribute data of importance to this work is summarised below.

For the attribute sense of direction, Zacharov & Koivuniemi (2001b) comment that the subjects appear to have redefined the scales, or did not use them correctly, as the mono replays both show good sense of direction. The scale was supposed to gauge ease of determination of direction of source, not loudspeaker, whereas clearly the monophonic replay (especially as there was replay at 0° and 90°), could not have shown a good sense of direction of the originally recorded sound events. This author would argue that this case demonstrates the difficulty for rating scales that attempt to describe individual attributes of various scenes across various replay methods. In the case of sense of direction, there was confusion in the mind of the subjects when well defined sounds appeared through a single loudspeaker at a certain point in space, as no reference was given to describe where the sound should appear. The sounds were rated as having a good sense of direction, indicating that they were perceived to be well defined (possibly narrow?) sources. A negative correlation between sense of direction and broadness could also be seen, meaning that non-broad (i.e. narrow) sounds result in a greater sense of direction. Zacharov & Koivuniemi suggest a redefinition of the scale, but this author would suggest that rather than a single
stimulus paradigm, a reference is used, as this would provide something against which to rate the sense of direction of the other systems. The problem for Zacharov & Koivuniemi’s study would be which stimulus to take as a reference. If the dipole stereo technique that was used as a reference for the paired comparison tests was used as a reference, one would expect that this would be discriminatory against replay methods that are able to reproduce sound from positions outside the standard stereo arc (such as 3/2 stereo or the periphonic cube).

Another instance where a scale is apparently not sufficiently well defined is naturalness. According to Zacharov & Koivuniemi (2001b), “refinement of the definition of attribute definition would be required” if, as it seems, the subjects were using the scale to rate “natural or familiar soundscapes”. This was because those that were rated high on this scale tended to be ambient/location recordings from real life scenes. For this author, this demonstrates the problem with a rating of overall realism or naturalness. It is hypothesised that an understanding of what constitutes a natural or realistic stimulus will only be achieved once all the perceptual cues that constitute an accurate simulation of actual auditory events and soundfields are discovered, analysed, understood and correctly simulated. Zacharov & Koivuniemi (2001c) themselves state that, despite decades of development, “there is still no clear and definitive understanding of what is required to recreate a perceptually correct spatial sound field”.

For the distance to events attribute, the stimuli created using the same source in different environments showed increases in perceived distance to events in environments that were more reverberant, even though they were recorded at the same distance. The subjects were not reported to have problems using the scale, even though it may not have been clear which sources were used for the rating of distance in the more complex stimuli (those involving multiple sources). It would be reasonable to assume that the most salient direct sounds were used by the subjects in each case, but it is not apparent from the report if such sounds existed in every track. When using a broad range of widely differing stimuli, it is difficult to find suitable adjectives for single rating scales that can be applied equivalently across the stimuli. This author would suggest using a group of similar sound sources that could be identified as equivalent in any direct attribute analysis. For example, if all stimuli involved a single salient source (voice or musical instrument or other sound source), then, for example, distance to events could be specified as pertaining to this particular source. An additional control could be to record the same source (loudspeaker reproduction of speech for example) in every environment, including the realistic “location” – as Zacharov & Koivuniemi (2001c) did for six of their stimuli. This would ensure that each one of the subjects was focussing on the same percept during attribute rating experiments, and therefore the results could be examined across the stimuli.

Correlation analysis revealed a few, rather weak correlations in the resulting attribute data, which Zacharov & Koivuniemi (2001b) argue is potential evidence that the scales used were independent of one another (although they do say that the small amount of data collected voids this conclusion for the time being). The correlation analysis did, however, show that broadness was somewhat correlated with both sense of direction and with sense of depth, a fact that they explained had something to do with the level of reverberation in the recordings, explaining that “increasing the
reverberation" would increase the broadness and may increase the depth of the sound (Zacharov and Koivuniemi 2001b).

The conclusions of the study relate to the model developed to explain the preference data from the "naive and untrained" in terms of the direct attribute data from the "selected and trained" subjects. The underlying principle components are described by Zacharov & Koivuniemi (2001b) as providing "valuable insight into the salient aspects of this experiment", but actually appear to be rather confusing to interpret, composing of multiple attributes and interactions. Zacharov & Koivuniemi point to the fact that the sense of movement attribute is the largest contributor to preference. Unfortunately graphs showing the rating of sense of movement and preference for the various stimuli and the various sound reproduction formats are not presented in the paper. From the description of the stimuli and reproduction systems, it is possible to imagine that certain replay systems (mono, for example) may not have been able to reproduce sound source movement, whereas certain stimuli (male speech or guitar) may not have involved any original source movement. These replay systems or stimuli may not have been preferred for a number of reasons that did not involve sense of movement, but the lack of movement or movement replay ability could have simply coincided with these cases. One cannot rule out the possibility that items involving moving, dynamic sources or reproduction systems that can recreate these cues, are more involving and potentially preferable to listen to. To investigate this further, the direct attribute rating experiment could be repeated using new dynamic recordings of the previously static material (moving speech and guitar sources for example), and static recordings of the previously dynamic material (a stationary rather than moving train at the platform, or stationary traffic at the bus stop for example). The direction, rate and manner of the motion are not specifiable using the sense of movement scale, so this could also potentially add variation to the data. For purposes of this work, it is noted that a unique attribute sense of movement was discovered in the Zacharov & Koivuniemi study, potentially as a result of the programme material selected for the experiment, but that this attribute was not clearly defined and a source of much potential variability.

2.1.6.6 Summary

A lot of the good work and rigour in the planning seems to have been at least partially undermined by the use of just sixteen naive subjects who were subsequently used again as if they were experienced subjects for the direct attribute elicitation and rating experiments. Zacharov & Koivuniemi also failed to hide the replay loudspeakers from the view of the subjects.

Although direct attribute elicitation through group discussion yielded a number of direct attribute scales for spatial audio, these scales tended to be orientated towards sentiments rather than judgements (Nunally and Bernstein 1994), which may have been as a result of using a non-expert elicitation panel (experts may have been able to provide a series of scales by which judgements could be made as to the spatial quality of the stimuli). This has led to scales which are sometimes difficult to interpret and seem to have been used in differing ways by different subjects on a number of occasions. The scales used do, however, provide yet another set of descriptive characteristics for audio stimuli and reproduction systems. The sense of movement scale unique to this study was elicited due to the use of dynamic stimuli, and although this was found to be an important aspect of listener preference, it is
regarded by this author not to have been sufficiently defined to be used consistently and in a way that is comparable across various samples. The lack of detailed reported results for the *sense of movement* scale is evidence that this attribute may not have been used in a way that is fully interpretable.

### 2.1.7 Berg’s Studies

Jan Berg, presented a series of studies with Francis Rumsey (Berg and Rumsey 1999, 2000a,b, 2001, 2002) culminating in his Doctoral Thesis (Berg 2002). Berg set out to obtain a set of spatial audio attributes with a common meaning between subjects by eliciting attributes from a group of listeners using various multichannel stimuli. Like Zacharov & Koivuniemi (see Section 2.1.6), Berg attempted to elicit attributes directly from subjects using audio stimuli, rather than provide attributes for the subjects to rate, as in many of the other studies detailed in the preceding sections. Berg would, in his later experiments, provide the attributes previously elicited to new sets of subjects (Berg and Rumsey 2001, 2002).

Berg’s studies moved from work on the differences between stimuli using attributes elicited from different recording/reproduction methods of the same events, to a validation experiment rating the previously elicited spatial attributes of different simultaneous five-channel recordings of identical events. The work involved three separate experiments, summarised below.

#### 2.1.7.1 Experiment 1

Berg describes his 1998 experiments – reported in (Berg and Rumsey 1999) – as a pilot study to test the repertory grid technique as a method for eliciting spatial sound attributes from various recorded environments and reproduction formats.

Eighteen subjects described as Swedish-speaking students (ten studying sound recording, eight studying music or media) were used in an elicitation experiment to rate various reproduction techniques of six different programme items. The programme items that Berg chose included single source items, orchestral works, pop music and an outdoor scene. Three different recording techniques were used for each item, in most cases these were a mono recording, a stereo recording and a multichannel recording (sometimes three or four channels, but mostly five channels). One of the items included a phase-reversed signal. Stimuli were replayed over one or more of a five-loudspeaker system that appeared to conform to (ITU-R 1994-1997) (although this was not stated). The three stimuli for each programme item were loudness aligned to within 2 dB of one another.

Berg later reports that of the eighteen subjects, one of each of the subgroups did not complete the tests leaving sixteen subjects in total (Berg and Rumsey 2000b). He also states that half of the subjects were given an additional instruction to concentrate on the differences in “the three-dimensional nature of the sound sources and their environment” (Berg and Rumsey 2000b). This is perplexing, as it appears to be an intentional pre-biasing of the subjects to consider a paradigm that was later adopted (that of source and environment attributes). Unfortunately this author has found no explanation as to why this was done. It is possible that certain subjects required more information than others before being able to complete the task. It could also be possible that this was a scientific biasing to work out the effect that it would have,
however it was not stated which “half” of the subjects were pre-biased, and what effect this pre-biasing had.

Inspired by the repertory grid technique, the elicitation process consisted of the subjects listening to six “triads” (which were the three recording techniques for each of the six programme items). Subjects wrote down the ways in which two of the techniques were similar to each other and different from the third. They continued listening to the stimuli until no further constructs were elicited. It is worth noting that no direct comparisons were made between different programme item types – as happened in the Zacharov & Koivuniemi experiment (see Section 2.1.6.3). The elicitation was effectively being performed using differently downmixed version of the same events, and it is likely that the sound sources would change position (and hence direction relative to the listener) between downmixed items naturally. This author conducted experiments (Zielinski, Rumsey, Kassier and Bech 2005) involving rating the spatial quality of downmixed items and found it difficult to assign a unified set of attributes to the changes perceived by subjects of a pilot elicitation test. The solution found in (Zielinski, Rumsey, Kassier and Bech 2005) was to use a type of mean opinion score for spatial quality for certain segments of the replay field.

As a result of the elicitation exercise and a rating experiment (which consisted of two iterations of a reduced set of the total stimuli) conducted by each subject using their own individual descriptors to rate the various stimuli, Berg was able to suggest a number of commonly used descriptors. Elicited descriptors that involved preference had been discouraged during the elicitation phase, with subjects being rather asked to explain why certain stimuli were or were not preferable. Although elicited descriptors in (Berg and Rumsey 1999) ranged between nine and thirty for different subjects, Berg was able to group similarly rated descriptors together to form a reduced set of what he claimed to be “unrelated constructs” (Berg and Rumsey 1999). Berg summarised the most common of these as:

- **Authenticity/naturalness**
  This involved some form of naturalness, authenticity or the sensation of being present at the recorded venue.

- **Lateral positioning/source size**
  Which included descriptors such as “mono-stereo”, “one direction – many directions” and “narrow sound source – wide sound source”. As a word of caution, this author suggests that this concentration on directional and width attributes could simply be a direct consequence of the crude changes between mono-stereo-multichannel replay environments which were used in the elicitation process. This does not preclude the possibility that these descriptors relate to subjective perceptions of specific spatial sound fields, as similar attributes have been used in previous studies. It is worth noting that, like the Zacharow & Koivuniemi studies (see Section 2.1.6), elicited attributes are linked to the programme material used in the elicitation and the methods used to create differences between the stimuli. Care must be taken when applying attributes found during elicitation exercises more generally.

- **Envelopment**
  Described by Berg as “the feeling of being surrounded” (Berg and Rumsey 1999), with descriptors including “sound everywhere – sound from a part of
Background Theory – Spatial Attributes

"the room" and "in the center of the event – outside the event". Again, the feeling of envelopment is likely to have arisen as a direct consequence of switching from frontal mono or stereo reproduction to multichannel sound during the elicitation stage. Envelopment is a concept that appears in other studies, and in the field of concert hall acoustics (Beranek 1996), so it is likely that this descriptor is applicable when considering stimuli other than examples of downmixing.

- Depth

Berg stated that less than half of the subjects used depth descriptors such as mono – depth or sound source in the loudspeaker – sound source between the speaker and me, all of which relate to concepts of depth in previously cited studies, and can also be explained in terms of varying distances of the individual sound sources.

In summary, the main descriptors found in (Berg and Rumsey 1999) seem to relate very closely to the actual method of affecting change in the stimuli. This is unsurprising, but this must be born in mind if these attributes are to be used in a generalised scheme. Berg later used various different five channel recording techniques in a pilot elicitation and rating experiment (Berg and Rumsey 2002), which could potentially provide more insightful information into spatial sound attributes that that used in the study reported in (Berg and Rumsey 1999).

It is also worth clarifying the difference between the Berg studies and those reported by Zacharov & Koivuniemi (see Section 2.1.6). In Berg’s studies, he elicited attributes from subjects using different recording/reproduction formats of the same event and then uses these to rate all the stimuli across similar scales. Zacharov & Koivuniemi used various recording/reproduction formats of different events to elicit attributes. These methods are likely to produce attributes with differences in emphasis as the complicated task of applying an attribute between different auditory stimuli falls in the elicitation stage during the Zacharov & Koivuniemi experiments, and the attribute rating stage in the Berg studies.

In (Berg and Rumsey 2000b) data from (Berg and Rumsey 1999) was further examined using a technique called verbal protocol analysis to segregate the subjective responses and possibly find additional attributes. Crucially, Berg decided to classify the elicited constructs as either descriptive or attitudinal. This ties in with concepts of judgements and sentiments (Nunally and Bernstein 1994), one relating to a reference or external standard or measure, the other resulting from an emotional response with no “right” or “wrong” answer.

In all, 342 constructs were elicited, a mean of 21 per subject. Berg makes the point that different subjects may have had very different descriptive skills, as certain subjects used mostly descriptive constructs and others mostly attitudinal constructs.

Berg concentrated on those constructs considered to be descriptive and grouped similar terms into groups, which he summarised in (Berg 2002) as:

- Localisation, left-right and front-back
- Depth / Distance
- Envelopment
- Width
Background Theory – Spatial Attributes

- Room perception
- Externalisation
- Phase
- Source width
- Source depth
- Detection of background noise
- Frequency spectrum

Berg makes the point that some attributes seem to be inter-related, and uses the example of externalisation. A source that is externalised has a certain distance, whereas an internalised source has “zero” distance.

The room perception attribute is multidimensional and includes room size, reverberation and the feeling of a room. Technical and timbral attributes are also evident here, with phase, frequency spectrum and background noise attributes. It seems odd that the phase attribute features in the results of both of the sequences of data (there were two sequences of stimuli presented in the rating experiment, one of that contained phase-reversed signals, one that did not). It is possible that the sound recording students were pre-biased due to their experience and were aware of out-of-phase signals and elicited this as a concept that they understood. It is not clear if the music/media students were using this term in the same way as the sound recording students were.

Source depth is Berg’s label for a group of constructs that only appear in one of the sequences (the one involving the phase reversed item), and do not specifically include the term source depth, but rather constructs such as sound source is V-shaped/sound source sits closer to the listener and large sound source/small sound source. It is the suspicion of this author that Berg considered the concept of a source having depth fitting nicely alongside source width as another group label, as the supporting evidence that the concept of source depth is an important spatial attribute is simply not provided.

In his discussion, Berg notes that aspects of naturalness features strongly, with the subjects distinguishing between being in the same room as the recorded sound source, and the experience of listening to a reproduction, which is expressed as presence. He suggests that “the other attributes are supporting the natural feeling through localisation of sound sources that have width and depth and are at certain distances from the listener in a room that envelops the listener” (Berg and Rumsey 2000b). This author’s view is that the sensation of presence is more akin to a mean opinion score or preference rating than a judgement of a specific attribute, as a number of different attributes of the soundfield would need to contribute coherently to provide an overall sensation of a natural reproduction.

In (Berg and Rumsey 2000a), Berg further classified constructs as either descriptive, emotive or naturalness classes, and attempted to create mappings between them. Within the naturalness class were three sub-classes:

- natural / normal / real (or the opposite: unnatural / not common)
- technical device involved (loudspeaker, microphone, recording)
- feeling of presence (in the room or at the venue, or its opposite: absence)
Here we can see that the presence attribute which also featured in previous studies is categorised as a descriptive attribute. As it is expected that a general spatial audio training system would first focus on descriptive attributes, the attitudinal attributes naturalness and presence appear to be unsuitable for inclusion initially.

Within the emotive group lay a number of terms that were classed as attitudinal constructs in (Berg and Rumsey 2000b) but were found to be descriptors of timbral components such as sharp or dull in (Berg and Rumsey 2000a). Such ambiguity in classifying verbal data not only means that such timbral descriptors should be used with caution, but also points to the susceptibility of all methods to the interpretation of the experimenter(s) and the method chosen for analysis.

The descriptive group contained:

- **Localisation**, which is the ability to pinpoint directions, both lateral (left-right) and front-back.
- **Depth/distance**, which is a perceived distance to the sound source, or a depth localisation. Another feature of depth is a perception of the source’s shape, the source depth.
- **Envelopment**, which is when the listener feels surrounded by sound or feels like being within the sound source.
- **Width**, which has different aspects, both general remarks on the width of the overall sound stage or image and specific references to the source’s width.
- **Room perception**, which denotes the subjects’ experience of room size, reverberation, or just the ability to perceive the ‘feeling of a room’
- **Frequency spectrum**, which is description of bass, treble and other spectral components.

Additionally, the following two attributes were found to be necessary:

- **Lack of room perception**, which is a difficulty to perceive a room (that ‘should’ be there).
- **Lack of (normal) width**, which is when the width is ‘artificial’ or even ‘too large’.

Berg did not believe these to be explained by room perception and width respectively, because they did not directly measure those attributes, but rather a lack of a clear indication of them. They seem to have been included for completeness.

The results of the correlation analysis in (Berg and Rumsey 2000a) confirm that naturalness and presence can be considered overall attributes (either descriptive or emotive) that are built from individual constituent parts because they were found to be present in a number of individual groups containing various descriptive attributes.

### 2.1.7.2 Experiment 2

The second of Berg’s reported experiments (Berg and Rumsey 2001) aimed to investigate the applicability of a selection of the previously used attributes using partly new stimuli.
Background Theory – Spatial Attributes

### Table 5: Grouped attributes and descriptions from (Berg and Rumsey 2001)

<table>
<thead>
<tr>
<th>Group</th>
<th>Attribute</th>
<th>Description from (Berg and Rumsey 2001)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naturalness</td>
<td>How similar to a natural listening experience (i.e. not reproduced through e.g. loudspeakers) the sound as a whole sounds. Unnatural = low value. Natural = high value.</td>
<td></td>
</tr>
<tr>
<td>Presence</td>
<td>The experience of being in the same acoustical environment as the sound source, e.g. to be in the same room. Strong experience of presence = high value.</td>
<td></td>
</tr>
<tr>
<td>Preference</td>
<td>If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the content of the programme, i.e. do not assess genre of music or content of speech. Prefer the sound = high value.</td>
<td></td>
</tr>
<tr>
<td>Envelopment</td>
<td>The extent of how the sound as a whole envelops/surrounds/exists around you. The feeling of being in the centre of the sound. Feel enveloped = high value.</td>
<td></td>
</tr>
<tr>
<td><strong>Source Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Width</td>
<td>The perceived width of the source. The angle occupied by the source. Does not necessarily indicate the known size of the source, e.g. one knows the size of a piano in reality, but the task to assess is how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e.g. reverberation - only assess the width of the sound source. Narrow sound source = low value. Wide sound source = high value.</td>
<td></td>
</tr>
<tr>
<td>Localisation</td>
<td>How easy it is to perceive a distinct location of the source - how easy it is to pinpoint the direction of the sound source. Its opposite (a low value) is when the source's position is hard to determine - a blurred position. Easy to determine the direction = high value.</td>
<td></td>
</tr>
<tr>
<td>Source distance</td>
<td>The perceived distance from the listener to the sound source. Short distance/close = low value. Long distance = high value.</td>
<td></td>
</tr>
<tr>
<td><strong>Room Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Room width</td>
<td>The width/angle occupied by the sounds coming from the sound source's reflections in the room - not the sound source itself. Narrow room = low value. Wide room = high value.</td>
<td></td>
</tr>
<tr>
<td>Room size</td>
<td>In cases where you perceive a room/hall, this denotes the relative size of that room. Large room = high value. If no room/hall is perceived, this should be assessed as zero</td>
<td></td>
</tr>
<tr>
<td>Room spectral bandwidth</td>
<td>The perceived bandwidth of the room. Room with large bandwidth = high value.</td>
<td></td>
</tr>
<tr>
<td>Room sound level</td>
<td>The level of sounds generated in the room as a result of the sound source, e.g. reverberation - i.e. not extraneous disturbing sounds. Weak room sounds = low value. Loud room sounds = high value.</td>
<td></td>
</tr>
<tr>
<td><strong>Other Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background noise level</td>
<td>The level of sounds not generated by the sound source itself. Weak background noises = low value. Loud background noises = high value.</td>
<td></td>
</tr>
</tbody>
</table>

Berg omitted some of the attributes that were previously used (externalisation, phase, technical device) because they were said to have resulted from the phase reversed stimuli used in (Berg and Rumsey 1999), which would not be re-used. Externalisation was also excluded because it was considered to be a "dichotomous attribute hard to grade on the linear scales intended for the experiment" (Berg and Rumsey 2001). On this basis, this author would also see grounds to exclude a number of other attributes, for example presence, as it is unclear if this translates to a linear scale (you are either at the venue or you are not). In any case, the particular
attribute *externalisation* was previously (Berg and Rumsey 2000b) argued to be explicable through use of the *distance* attribute. In another significant move, the twelve attributes chosen for this experiment were separated into *general, source, room* and *other* attributes and were defined as shown in Table 5:

It is interesting to note that *source depth* previously described in (Berg and Rumsey 2000b) is not present here.

The stimuli used were taken from the previous stimuli. Four of the multichannel recordings (they were all five-channel) were downmixed to two-channel stereo and two channel phantom mono. One of the programme items (the speech recording) was also downmixed to single channel mono to create thirteen stimuli. The stimuli were replayed via one, two or five loudspeakers of a five-channel loudspeaker set-up conforming to (ITU-R 1994-1997).

Nineteen sound recording students were used as subjects. They had not taken part in the previous experiment (Berg and Rumsey 1999). A training session was provided using a quarter of the stimuli to avoid fatigue. Subjects were familiarised with the attribute definitions that were used and the use of the computer system that controlled the test. As for the test itself, it took the form of a multiple stimulus comparison, where all push buttons controlling individual playback of all thirteen stimuli were aligned with continuous rating sliders labelled zero and max at the extremes. The position of each of the stimuli was randomized on successive iterations, each iteration required a different attribute to be rated. The attributes were pseudo-randomised, with groups of attributes being selected randomly, all the attributes of the selected group would then be rated in a randomised order before moving on to the next attribute group, which was selected at random. Subjects were asked to grade at least one of the stimuli at “max” for each attribute, presumably to attempt to normalise the data provided by the subjects.

Berg’s criterion for attributes having a common meaning was whether statistically significant differences could be made between the stimuli through its use. This author would like to point out a fictitious scale as preposterous as “contains saxophone” could also be used to make significant differences between the stimuli containing the solo saxophone and the stimuli that do not. Berg’s method maybe enough to suggest a common meaning, but is not likely to be enough to prove or even to suggest that the attributes form a valid set for the overall description of these or generalised spatial sound stimuli.

The subjects’ responses were normalised to adjust for subjective sensitivity, and subjected to an ANOVA test. The consistency with which the various subjects used the attributes was examined by comparing residuals of the ANOVA. It was found that certain attributes were being used in a more consistent manner than others, shown in Table 6.

Whilst it is possible that listener training could improve the consistency of the attributes, it is worth noting that without specific training in spatial attribute detection, Berg’s subjects were fairly consistent in the use of following concepts (amongst others): *room width, room size, source distance* and *envelopment*. This indicates that these attributes were perceived in a unified manner across the subjects, pointing to their validity as reliable descriptors.
Table 6: Consistency of attributes shown from (Berg and Rumsey 2001)

<table>
<thead>
<tr>
<th>high consistency attributes:</th>
<th>mid-consistency attributes:</th>
<th>low consistency attributes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>room width</td>
<td>presence</td>
<td>localisation</td>
</tr>
<tr>
<td>room size</td>
<td>source width</td>
<td></td>
</tr>
<tr>
<td>source distance</td>
<td>naturalness</td>
<td></td>
</tr>
<tr>
<td>envelopment</td>
<td>room sound level</td>
<td></td>
</tr>
<tr>
<td>preference</td>
<td>room spectral bandwidth</td>
<td></td>
</tr>
<tr>
<td>background noise level</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Berg's subjects were not consistent in their use of (amongst others): *naturalness*, *localisation* and *source width*, indicating that these attributes may be less reliable as concepts to describe spatial audio. The fact that Neher (2004) encountered difficulties with the simulation of *source width* seem to confirm this to some extent.

Berg reported a negative correlation between the grades for *localisation* and *source width* scales, indicating that these terms could be interdependent. Examination of their definitions reveals the possibility that if a source is perceived to be *wide* it is likely to be difficult to pinpoint its direction.

Other correlated items include *naturalness* and *presence* (a highly natural-sounding environment gives a realistic experience that you are at the venue), and *room width* and *room size* which is not as highly correlated as could have been expected (because rooms can be *deep* and *tall* as well as *wide*).

In summary, whilst providing a set of attributes that have a common meaning, there is not necessarily proof in (Berg and Rumsey 2001) that this set of descriptors is enough to describe these or other generalised spatial sound stimuli.

The subdivision of attributes into *general*, *source* and *room* is significant here, as it is the first time that these concepts have been separated out in the spatial attribute studies previously cited. Considering the source and room separately when listening to spatial audio allows for a logical decomposition of spatial audio scenes into the elements that make it up, in a similar way to auditory system's method of segregation of streams (Bregman 1990), promising an analytical paradigm that is closer to the auditory system's innate methods.

### 2.1.7.3 Experiment 3

A third experiment was conducted by Berg and published with Rumsey in 2002. The main difference between this experiment and the previous one (Berg and Rumsey 2001) was the nature of the sound stimuli.

Two programme items were used: a solo viola and a duet (singer and piano), each recorded in a hall using five simultaneous five-channel microphone techniques. This created ten stimuli, which would be reproduced via a five-channel loudspeaker set-up.
conforming to (ITU-R 1994-1997). It was Berg’s intention that the differences between these stimuli were smaller and would therefore be a better test for the spatial audio attributes elicited in previous experiments.

Sixteen Swedish male sound recording students participated as subjects. Six had taken part in the experiments reported in (Berg and Rumsey 2001), but otherwise there was “no special training” given to the subjects (Berg and Rumsey 2002). The subjects were, however, considered to be “experienced” (Berg and Rumsey 2002).

**Preliminary elicitation experiment**

Four of the sixteen subjects took part in a preliminary elicitation test to see if subjects could detect differences in the stimuli, and elicit any additional attributes as needed – Berg does not say whether these four subjects also took part in (Berg and Rumsey 2001).

For the preliminary elicitation test, five of the stimuli were arranged into ten possible triads for each stimulus for elicitation, with each triad of stimuli being presented twice. Although this would mean that fifty comparisons would be needed to have experienced all combinations, Berg reports that only 24 comparisons were undertaken, resulting in 49 “bipolar constructs” (Berg and Rumsey 2002). These constructs were considered for inclusion with the other attributes used previously, resulting in two timbral attributes (not of immediate concern to this work) and three spatial attributes being added to other attributes used in previous experiments.

The spatial attributes are shown with their definitions in Table 7.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition from (Berg and Rumsey 2002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>room envelopment</td>
<td>The extent to which the sound coming from the sound source’s reflections in the room (the reverberation) envelops/surrounds/exists around you - i.e.: not the sound source itself. The feeling of being surrounded by the reflected sound. Low extent of envelopment = low value. High extent of envelopment = high value.</td>
</tr>
<tr>
<td>source envelopment</td>
<td>The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source. If several sound sources occur in the sound excerpt: assess the sound source perceived to be the most enveloping. Disregard sounds coming from the sound source’s environment, e.g. reverberation - only assess the sound source. Low extent of envelopment = low value. High extent of envelopment = high value.</td>
</tr>
<tr>
<td>ensemble width</td>
<td>The perceived width/broadness of the ensemble, from its left flank to its right flank. The angle occupied by the ensemble. The meaning of “the ensemble” is all of the individual sound sources considered together. Does not necessarily indicate the known size of the source [sic. “ensemble”]. Example: we know the size of a string quartet in reality, but the task to assess is how wide the sound from the string quartet is perceived. Disregard sound coming from the sound source’s environment, e.g. reverberation - only assess the width of the sound source [sic. “ensemble”].</td>
</tr>
</tbody>
</table>
The new attribute *room envelopment* – previously *envelopment* (Berg and Rumsey 2001) – appears to resemble the attribute *listener envelopment* (LEV) used in studies into concert hall acoustics (Beranek 1996).

The concept of *source envelopment*, however, is not entirely clear from the definition in (Berg and Rumsey 2002) (also shown in Table 7). An enveloping sound is defined by Berg as to be existing around the listener, so could be considered to have poor localisation, perhaps a large extent (therefore large source width) and since the direct rather than reflected sound is coming from all around, the centre point of the sound source is probably perceived to be coming from relatively close to the listener (i.e.: the listener feels that they are within the sound). One can also conceive of a scenario in which the listener could feasibly be surrounded by a single sound event that stretches around him at a certain distance (which could also be considered being *enveloped* by a source) - in this case the event could be described as having a large width, measured as an arc about him, with the source's centre existing at a certain distance close to the listener.

The concept of *ensemble width* is useful, as it allows the categorisation of an *ensemble* of sources.

**The main experiment**

Other attributes were brought forward to the main experiment that were taken from previous experiments, with inconsistently used ones removed. The attributes used in the main experiment are shown in Table 8:

<table>
<thead>
<tr>
<th>General Attributes</th>
<th>Source Attributes</th>
<th>Room Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>naturalness</td>
<td>localisation</td>
<td>room size</td>
</tr>
<tr>
<td>preference</td>
<td>ensemble width</td>
<td>room width</td>
</tr>
<tr>
<td>presence</td>
<td>source width</td>
<td>room sound level</td>
</tr>
<tr>
<td>low frequency content</td>
<td>source envelopment</td>
<td>room envelopment</td>
</tr>
<tr>
<td></td>
<td>source distance</td>
<td></td>
</tr>
</tbody>
</table>

These attributes were provided to the sixteen subjects, who graded all ten stimuli using all attributes via a multiple stimulus comparison technique similar to that used in (Berg and Rumsey 2001).

Because the two different programme materials have one and two sources each respectively, there was a potential problem regarding which source to rate. Berg reports that the source-related attributes *source width* and *localisation* but intriguingly not *source distance* or *source envelopment* were judged separately for each source for the stimuli that included two sources. *Source distance* and *source envelopment* were judged for the closest and the most enveloping respectively for the stimuli with two sources. No explanation is offered as to why these were not dealt with separately, and there appears to be no obvious reason for doing this, except for experimental expediency. This is not ideal as there could well be confusion as to which sound source is to be judged. Although this may be acceptable with the relatively simple stimuli used in this experiment, this rating of potentially different sources depending on subjective judgements could be expected to be problematic when complex stimuli involving multiple sources are used.
Background Theory – Spatial Attributes

As in (Berg and Rumsey 2001), there was a training phase – this time four attributes and all stimuli were used to familiarise the subjects with the experimental procedure.

Berg’s analysis of the results pertains mostly to whether significant differences can be found between the stimuli using the attributes. His analysis showed that all attributes enabled the group to produce judgements significantly indicating differences between the stimuli, which he claimed meant that the attributes had a common meaning to the subjects.

Other analysis showed that the least consistent items included naturalness and presence which seems to confirm the findings in (Berg and Rumsey 2001). The source width attribute was rated both relatively consistently (for the piano source) or relatively inconsistently (for the viola or voice sources). A correlation was also found between source width and ensemble width (probably a general broadening of the entire spatial audio scene would broaden the sources and therefore the increase the width of any ensemble that they are within) and a negative correlation was found between source width and localisation, in a similar way to (Berg and Rumsey 2001).

As an aside, an investigation by Lee (2006) found that source width and locatedness (the ability to localise the sound) were not particularly negatively correlated. However, this author has reservations about the potential biases involved in asking subjects to rate just two specific attributes in the study – it is feasible that subjects expected that they were required to rate the stimuli in different ways for each attribute which decreased the possible negative correlation of the attributes.

In a potential corroboration of the concept of the volume of tones (Stevens and Davis 1938), both source width and source envelopment appear to have been correlated with low frequency content. In addition, source width and source envelopment appear to be correlated with one another, which could potentially confirm at least some of the suspicions of this author about the source envelopment attribute expressed above. A final correlation was reported by Berg, that of source distance with room level, which Berg notes could be to do with the direct to reverberant sound level, with more distant sounds creating a larger relative level of reverberation at the listening position. From the evidence above, this author suggests further investigation of source width as a spatial attribute is needed if it is to be included in a generalised training scheme. Source envelopment and the previously used source depth attributes are not proven to be valid descriptors of spatial sound to this author based on the research presented by Berg.

Regarding perception of the recorded environment, Berg notes that the experiments reported in (Berg and Rumsey 2001) and (Berg and Rumsey 2002) indicated that room perception is twofold: one relating to sensation/impression of presence, and one relating to the judgement of certain room characteristics, such as room size and level of the reflected sound in the room.

Because judgements refer to some external reference or standard, they may be the most appropriate type of attribute for a generalised training scheme. It is likely that a future spatial audio training system would initially concentrate on training subjects to perceive and discriminate between judgements of source and/or ensemble dimensions and room characteristics such as room size and room level. Additionally it is possible that a closer study of subjective attribute judgements could lead to a greater understanding of the link between objective metrics and certain subjective perceptions.
Summary
The Berg study provides a number of very important points regarding spatial attributes. The most important being the separation of attributes that correspond to sources and their environment (which he terms room attributes) which was not previously delineated. The additional classification of ensembles is also important as this allows for analysis of more complex spatial scenes which would be too complicated to analyse using every individual source. These classifications also address a problem with previous studies, as it was frequently not clear which part or parts of the spatial scene were to be graded using the various attributes. Considering the source and room separately when listening to spatial audio allows for a logical decomposition of spatial audio scenes into the elements that make it up, in a similar way to auditory system's method of segregation of auditory streams (Bregman 1990), promising an analytical paradigm that is closer to the auditory system's innate methods.

Berg also distinguishes between descriptive attributes (relating to a reference or external standard or measure) and emotive or naturalness attributes. This is a useful distinction, as it provides a framework for future spatial attribute specifications. This author expects that a training system for the detection of and discrimination between spatial audio attributes could focus on descriptive attributes initially. Descriptive attributes should be based on some external reference and therefore should not be affected significantly by listener preference. In this way the training system would be orientated towards improving the analytical and descriptive skills of subjects, potentially avoiding accusation that it could influence or and “train out” subjective preference.

2.1.7.4 Summary of Berg’s Studies
In his overall summary of (Berg and Rumsey 2001) and (Berg and Rumsey 2002) presented in (Berg 2002), Berg provides a useful list of criteria that he used in the modification and exclusion of attributes, which is expected to help this author in the creation of an attribute scheme for the proposed training scheme. The conditions stated in (Berg 2002) are:
- inapplicability to the context of spatial audio
- inapplicability to linear scales
- low listener consistency in rating

2.1.8 Rumsey’s Scene-Based Paradigm
Francis Rumsey (2002) drew upon previous studies involving spatial audio attributes, and his work with Jan Berg (see Section 2.1.7) to propose a scene-based paradigm for the evaluation of spatial audio scenes. It was intended to be “a contribution to the debate” and “ongoing work”, but is the most complete description of spatial audio published to date. The work done by this author in studying the previous spatial audio studies has aided him to clarify many of the concepts used in Rumsey’s scene-based paradigm, and informed the modification of the paradigm to form the basis of a training system for a spatial attribute training system.
The scene-based paradigm is a way of classifying the various elements of a spatial audio scene in terms of dimensional, immersion and miscellaneous attributes, shown in Table 9:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Source Width</td>
<td>Width of individual source(s) within a scene</td>
</tr>
<tr>
<td>Ensemble Width</td>
<td>Overall width of a defined group of sources (may be all the sources in the scene if required)</td>
</tr>
<tr>
<td>Individual Source Depth</td>
<td>Depth of individual source within a scene</td>
</tr>
<tr>
<td>Ensemble Depth</td>
<td>Depth of a group of sources</td>
</tr>
<tr>
<td>Individual Source Distance</td>
<td>Distance from listener to perceived location of the source</td>
</tr>
<tr>
<td>Ensemble Distance</td>
<td>Distance from listener to perceived midpoint of an ensemble</td>
</tr>
<tr>
<td>Environment Width</td>
<td>Broadness of (reflective) environment within which individual sources are located</td>
</tr>
<tr>
<td>Environment Depth</td>
<td>Depth of (reflective) environment within which sources are located</td>
</tr>
<tr>
<td>Scene Width</td>
<td>Composite or global width of entire scene</td>
</tr>
<tr>
<td>Scene Depth</td>
<td>Composite or global depth of entire scene, including environment</td>
</tr>
<tr>
<td>Individual Source Envelopment</td>
<td>Sense of being enveloped by a single sound source</td>
</tr>
<tr>
<td>Ensemble Source Envelopment</td>
<td>Sense of being enveloped by a group of sound sources</td>
</tr>
<tr>
<td>Environmental Envelopment</td>
<td>Sense of being enveloped by reverberant or environmental (background stream) sound</td>
</tr>
<tr>
<td>Presence</td>
<td>Sense of being inside an (enclosed) space or scene</td>
</tr>
<tr>
<td>Scene L-R Skew</td>
<td>Degree to which a spatial audio scene is skewed to the left or right from a stated reference position.</td>
</tr>
<tr>
<td>Scene F-B Skew</td>
<td>Degree to which a spatial audio scene is skewed to the front or back from a stated reference position.</td>
</tr>
<tr>
<td>Source Stability</td>
<td>Degree to which individual sources remain stable in space with respect to time (assuming nominally stationary sources)</td>
</tr>
<tr>
<td>Scene Stability</td>
<td>Degree to which the entire scene remains stable in space with respect to time</td>
</tr>
<tr>
<td>Source Focus</td>
<td>Degree to which individual sources can be precisely located in space (this may be closely related to individual source width)</td>
</tr>
<tr>
<td>Scene Width Homogeneity</td>
<td>Evenness of distribution of scene elements compared with a reference scene</td>
</tr>
</tbody>
</table>

The dimensional and immersion attributes follow a similar pattern of decomposition of the spatial scene into (amongst others) source and room attributes (see Section 2.1.7.2) and expanded on with an ensemble attribute (see Section 2.1.7.3).

The scene-based paradigm provides a framework to describe a wide range of spatial audio scenes – from single sources within a room (which would entail source and environment attributes) to a group of sources within a room (which may entail just ensemble and environment attributes), to a combination of individual sources (e.g.
Background Theory – Spatial Attributes

soloists) and groups of sources within a room (which may entail source, ensemble and environment attributes).

Within the paradigm, it is possible to see a hierarchical structure, with sources and ensembles having similar attributes, but describing the group of sources, rather than the individual sources. An example would be source distance, where a sound source is perceived at a certain egocentric distance – to use the terminology proposed in (Loomis 1995) – and ensemble distance, where a group of sources is perceived to be at a certain egocentric distance.

Rumsey (2002) points out that “attributes should be chosen for an evaluation based on the task and context in question”, and that every attribute in the paradigm is not intended to be included for every experiment. Although a training system for every one of the spatial audio attributes presented in Rumsey’s scene-based paradigm may be possible, it is not feasible within the time-scale of this study, so some simplification and perhaps modification will be required for the specific purpose of creating a reduced but coherent spatial attribute training system. Whilst this will be the focus of Chapter 3, some initial thoughts and concerns about some of the concepts within the scene-based paradigm are discussed in this section.

An important point to bear in mind is that Nunally and Bernstein (1994) warned that “attributes should not be confounded with each another”. This author has examined Rumsey’s paradigm, and has found that it is possible to explain some of the attributes in terms of some of the other attributes, which calls into question whether they do not confound one another.

Whilst the dimensional and immersive attributes seem to conform more-or-less to what is found elsewhere in the literature, the miscellaneous attributes seem to have been included for completeness. There does not appear to be much supporting evidence for the existence of the attributes referring to skew in previous studies. Perhaps skew was logically deduced by Rumsey to supplement the paradigm, explaining potential spatial audio scenes that were otherwise impossible to describe using the dimensional and immersive attributes on their own. The stability attributes show the influence of Toole’s provided attribute definition of the sound images (defined in Table 2 in section 2.1.4.2), which also concerns positional stability. Rumsey admits that source focus attribute could be closely related to individual source width. The scene width homogeneity attribute appears to have been inspired by Toole’s provided attribute continuity of the sound stage (defined in Table 2 in section 2.1.4.2). Many of the miscellaneous attributes are defined in terms of being compared with a reference, which this author would imply means that they appear to describe a number of changes between two spatial scenes rather than the actual changes themselves. For example, a spatial scene that is skewed may in fact be describable in terms of positional changes of either sources or ensembles at different positions within the scene (it seems more likely to this author that any skewing perception would be more to do with the sources than the environment).

Regarding the immersive attributes, it is the view of this author that the individual and ensemble source envelopment attributes can be explained and conceived in terms of varying distances of the sources or ensembles. Individual source envelopment was first introduced as source envelopment by Berg & Rumsey (see Section 2.1.7.3). In a similar way ensemble source envelopment (which is Rumsey’s extension of the concept of envelopment to ensembles) can, in this author’s opinion be simplified by
the use of ensemble distance or ensemble width changes. If the ensemble appears to envelop the listener at a certain distance from the listener, for example, a line of choristers encircling the listeners, then the listener could use Rumsey’s terminology of being enveloped by the ensemble, or could use other terms defined by Rumsey to describe a wide ensemble, at a certain distance from the listener. Here, ensemble width is considered to be constant distance ensemble width as used by Neher (2004). In this, the width of an ensemble is considered to subtend an arc about the listener, rather than linear ensemble width which was considered perpendicular to the egocentric axis. If, however, the listener appears to be within the centre of a choir, with sources all about him, the listener could use Rumsey’s terminology of being enveloped by the ensemble, or could use other terms defined by Rumsey to describe that he feels close to, or perhaps at zero distance from the midpoint of the ensemble. Environmental envelopment, on the other hand, could relate to a new dimensional attribute: environment distance, where the feeling of being enveloped by the environmental sounds and the feeling of presence could be described by using the concept of an egocentric distance from the mid-point of the environment, potentially including the possibility of being outside the environment, looking in (which would cover the perspective attribute used by Toole (see Section 2.1.4 and Table 2). From the evidence from previous studies, it appears that the presence attribute is more akin to naturalness, in that it is an aggregated perception of a number of attributes. For example, Berg states that “the perception of different aspects of the room was most important for the feeling of presence” (see Section 2.1.7.1). Presence, whilst an important concept reported in many of the previous studies, also appears to be dichotomous in nature. In other words, you either feel that you are at the recorded venue, or not. Other forms of presence have been suggested (such as in the perspective attribute used by Toole (see Section 2.1.4 and Table 2) the concepts of you are there; they are here or outside looking in, but these seem to be abstract descriptions without a scale – linear or otherwise – between them.

Certain dimensional attributes that appear in the scene-based paradigm have already been questioned by this author. For example, source width, source depth and source envelopment. The concepts of source and ensemble could be combined using the term scene component first coined by Mason et al. (2001). Using this terminology, diagnostic tests (see Section 2.1.5) could use scene components defined by the experimenter, whereas the subjects in heuristic tests (if trained to use the concept) – see Section 2.1.5 – could be asked to decompose spatial audio scenes in terms of scene components of interest to them, or which they perceive as being together, which could be described by each subject and compared with one another.

It is prudent then, that the initial stage of the development of a training system for spatial audio should use a simplified set of attributes that can be widely applied by subjects across a wide variety of spatial scenes. If the attributes used are judgemental (rather than sentimental) (Nunally and Bernstein 1994) and can be classified as mainly dimensional attributes, then this author believes that the training scheme can be defended against the accusation that it will train out important subjective perceptions. If more general concepts such as scene component dimensions are used for the training scheme, it is expected that subjects will be able to apply their training to different spatial paradigms and concepts that they are required to use in other experiments.
2.1.9 Summary of Spatial Audio Attributes Used in Previous Studies

Experimenter bias appears to be ever-present. Studies that involve the elicitation of attributes have aimed to avoid biasing the listeners, yet the resulting data needs to be analysed and interpreted by someone, most likely the experimenter, who will bring their own opinions, knowledge and associated biases to bear on the results. In this way, none of the aforementioned studies are free from experimenter bias, giving a case for the use of provided (as opposed to elicited) attributes, which have been criticised in the studies involving elicitation of spatial attributes – see Sections 2.1.6 and 2.1.7. Another common factor running through the experiments is how the choice of programme material dictates the attributes that are used or elicited in any of the experiments. Despite this, there do appear to be a number of common and somewhat distinct spatial sound attributes described by subjects, or defined by experimenters.

By way of a summary, the attributes arising from each of the aforementioned studies are discussed in terms of their position inside or outside Rumsey’s scene-based paradigm (see Section 2.1.8) and placed in appropriate positions in Table 10, Table 11 and Table 12.

<table>
<thead>
<tr>
<th>Table 10: Summary of Dimensional Spatial Attributes from previous studies, placed within Rumsey’s Scene-Based Paradigm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eisler</strong></td>
</tr>
<tr>
<td>Clearness, fullness</td>
</tr>
<tr>
<td>fullness</td>
</tr>
<tr>
<td>depth of the image sources</td>
</tr>
<tr>
<td>depth of the image sources</td>
</tr>
<tr>
<td>depth of the image sources</td>
</tr>
<tr>
<td>environmental information</td>
</tr>
<tr>
<td>environmental information</td>
</tr>
</tbody>
</table>

- Scene Width
- Scene Depth
Table 11: Summary of Immersive Spatial Attributes from previous studies, placed within Rumsey's Scene-Based Paradigm

<table>
<thead>
<tr>
<th>Eisler</th>
<th>Nakayama et al.</th>
<th>Gabrielson et al.</th>
<th>Toole</th>
<th>Letowski</th>
<th>Zacharov &amp; Koivuniemi</th>
<th>Berg &amp; Rumsey</th>
<th>Rumsey</th>
</tr>
</thead>
<tbody>
<tr>
<td>fullness</td>
<td>fullness</td>
<td></td>
<td></td>
<td></td>
<td>Sense of direction, Broadness</td>
<td>Source Envelopment</td>
<td>Individual Source Envelopment</td>
</tr>
<tr>
<td>fullness</td>
<td>fullness</td>
<td>perspective</td>
<td>presence (nearness)</td>
<td>Sense of direction, Broadness</td>
<td>Ensemble Source Envelopment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>environmental information</td>
<td>fullness</td>
<td>feeling of space</td>
<td>Ambience / presence (nearness)</td>
<td>Sense of space</td>
<td>Room Envelopment</td>
<td>Environmental Envelopment</td>
<td></td>
</tr>
<tr>
<td>environmental information</td>
<td>feeling of presence / fidelity?</td>
<td>perspective</td>
<td>Ambience / presence (nearness)</td>
<td>Sense of space</td>
<td>Presence</td>
<td>Presence</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Summary of Miscellaneous Spatial Attributes from previous studies, placed within Rumsey's Scene-Based Paradigm

<table>
<thead>
<tr>
<th>Eisler</th>
<th>Nakayama et al.</th>
<th>Gabrielson et al.</th>
<th>Toole</th>
<th>Letowski</th>
<th>Zacharov &amp; Koivuniemi</th>
<th>Berg &amp; Rumsey</th>
<th>Rumsey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene L-R Skew</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scene L-R Skew</td>
<td></td>
</tr>
<tr>
<td>definition of the sound images</td>
<td>Source Stability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Source Stability</td>
<td></td>
</tr>
<tr>
<td>definition of the sound images</td>
<td>Scene Stability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scene Stability</td>
<td></td>
</tr>
<tr>
<td>fullness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scene Stability</td>
<td></td>
</tr>
<tr>
<td>definition of the sound images</td>
<td>directional selectivity</td>
<td>Sense of direction</td>
<td>Localisation</td>
<td></td>
<td></td>
<td>Source Focus</td>
<td></td>
</tr>
<tr>
<td>Clearness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Source Focus</td>
<td></td>
</tr>
<tr>
<td>continuity of the sound stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Source Focus</td>
<td></td>
</tr>
<tr>
<td>panorama</td>
<td>Scene Width Homogeneity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Scene Width Homogeneity</td>
<td></td>
</tr>
<tr>
<td>disturbing directional effects</td>
<td>abnormal effects</td>
<td>Penetration</td>
<td>Outside the paradigm</td>
<td></td>
<td></td>
<td>Outside the paradigm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sense of movement</td>
<td></td>
<td></td>
<td></td>
<td>Outside the paradigm</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Room Sound Level</td>
<td></td>
</tr>
</tbody>
</table>

Eisler’s experiment (see Section 2.1.1), whilst predating quadraphonic and surround sound (and possibly using predominantly monophonic reproduction methods), resulted in the labelling (by four experienced listeners) of two multidimensional spatial descriptors: environmental information – which seems to include source location and environmental attributes and possibly even presence as defined by Rumsey (see Section 2.1.8) – and disturbing directional effects – which does not seem to fit within Rumsey’s scene-based paradigm, seeming to describe similar anomalous images that Toole’s abnormal effects attribute aims to describe. This type of descriptor will either relate to a specific distortion, such as penetration as defined by Koivuniemi & Zacharov (see Section 2.1.6), or more generic, like Toole’s abnormal effects (see Section 2.1.4).

Nakayama et al. (see Section 2.1.2) varied the number of reproduction channels and defined three dimensions within the resulting (possibly naïve) preference data in terms of spatial descriptors: depth of image sources, fullness, and clearness. Depth
of the image sources refers to the distance of sources from the listener, but could also be considered to refer to ensemble depth. Fullness seems to refer to the concepts of envelopment and source or ensemble width, whereas clearness (which was not a particularly important dimension in Nakayama et al.’s study) could indicate source focus or (a lack of) individual source width.

In a series of experiments (see Section 2.1.3) Gabrielsson et al. studied spatial and other qualities of various forms of reproduction. Important descriptors to come from their work included perceived distance, feeling of space, fidelity (naturalness), feeling of presence, and fullness (see Section 2.1.3.2 - concerns a concept more akin to source or ensemble size). Importantly for this study, subjects seemed to find the set of scales which were developed using monophonic tests (see Section 2.1.3.3) inadequate to describe the spatial scenes and requested the inclusion of a stereo impression or image scale.

In the second of his mono/stereo loudspeaker tests (see Section 2.1.4), Toole used a subdivision of audio attributes along sound quality and spatial quality lines based on comments made by listeners during pilot tests. Present are the concepts of image definition (reminiscent of Gabrielsson’s clarity or Rumsey’s individual source width); continuity of the sound stage (Rumsey uses the miscellaneous attribute scene width homogeneity to describe this concept); width of the sound stage (which is more like ensemble width than scene width as it refers to sources) and distance also features in the term impression of distance or depth. Other attributes include abnormal effects which are not covered in Rumsey’s scene-based paradigm explicitly, but it is possible that the skew and possibly stability attributes could be thought of as abnormal effects (Eisler’s disturbing directional effects attribute seems to be close to Toole’s abnormal effects attribute). The perspective attribute seems to relate most closely with Rumsey’s presence attribute, but is actually defined in such a way as to include the concepts of distance and envelopment. Reproduction of Ambiance, Spaciousness and Reverberation refers to environmental cues such as the size of the environment or perhaps its envelopment of the listener.

Letowski (see Section 2.1.5) made a number of observations regarding the use of terms. Importantly, he drew the distinction between judgemental and emotional evaluation of sound, and introduced the concept that spatial qualities of sound can be separated from timbral qualities. Letowski also suggests that naturalness has both timbral and spatial (or his term spaciousness) aspects, and should be used as some form of overall descriptor. Naturalness can be seen to be a judgemental attribute, but one based on the overall judgement of a number of different perceptions. He also drew attention to the multidimensional nature of sound quality, and the need for parametric assessment. His MURAL suggests a decomposition of the various factors affecting sound quality with spaciousness being subdivided into three subgroups: directness (which contained directional selectivity similar to Toole’s definition of sound images); Stereophonic Impression (which included panorama which seems closest to Toole’s continuity of the sound stage quality scale, and ambience, which could be similar to Gabrielsson’s spaciousness (feeling of space) scale and Toole’s Reproduction of Ambiance, Spaciousness and Reverberation quality scale); the reverberance (liveness) group contains ambience, but also perspective (which could refer to a visual analogy of depth) and presence (nearness) (which must have something to do with distance due to the use of the nearness term, but could also refer to envelopment and involvement in the scene).
Zacharov & Koivuniemi (see Section 2.1.6) presented an involved study of sound quality including a lengthy attribute elicitation phase. Subjects were requested to listen to the various stimuli presented to them and “write down every spatial sentiment” that they perceive when listening to the stimuli. Koivuniemi & Zacharov are careful to describe the resulting attributes as a “first step” and that “new rounds of language development, and a new set of sound samples are necessary to build the attribute list”. The resulting attributes have a number of similarities with those elicited using different subjects, recordings and reproduction methods by Berg & Rumsey (see Section 2.1.7).

The spatial attributes elicited in the Zacharov & Koivuniemi studies (see Section 2.1.6) are shown below (with this author’s comments in parenthesis):

- **Sense of direction** (potentially involves width and focus concepts, as well as source/ensemble envelopment)
- Sense of depth (involves source-related distance and ensemble depth concepts)
- **Sense of space** (from the definition, it is equivalent to Rumsey’s presence attribute. Although not directly mentioned, the definition does not preclude environmental dimensions from being noticed, but due to the emphasis on sentiments rather than judgements, they are of secondary importance)
- **Sense of movement** (many of the stimuli used involved moving sources, whereas Rumsey defined his scene-based paradigm with stationary sources in mind. This attribute exists outside Rumsey’s scene-based paradigm).
- **Penetration** (the definition contains descriptors about source distance and naturalness but is most closely associated with Eisler’s disturbing directional effects or Toole’s abnormal effects)
- **Distance to events** (is the familiar concept of egocentric distance, describable between the listener and either individual sources or ensembles of sources).
- **Broadness** (includes the concepts of source or ensemble width and source or ensemble envelopment)
- **Naturalness** (does not fit within the paradigm and is not considered a spatial attribute, but rather a more an amalgamated grade that depends on a number of other attributes)

Berg & Rumsey (see Section 2.1.7) conducted three experiments into spatial attributes of sound. The first was an elicitation experiment inspired by the repertory grid technique (see Section 2.1.7.1), which provided over three hundred verbal descriptors. One of Berg’s crucial contributions was the classification of attributes as either descriptive, emotive or naturalness classes (see Section 2.1.7.1). This not only ties in with concepts of judgements and sentiments (Nunally and Bernstein 1994), but also segregates the naturalness and the presence attribute (which is similar to naturalness in that it is a mean opinion score-type grade based on a number of other attributes) from descriptive attributes. Concentrating on descriptive attributes, Berg & Rumsey began to formulate a list of attributes that was refined over the next two experiments (see Sections 2.1.7.2 and 2.1.7.3) to include many of the attributes that would form the basis of the scene-based paradigm. Of particular note is the second
Background Theory – Spatial Attributes

experiment (see Section 2.1.7.2), where attributes were grouped for the first time in terms of source and room attributes, allowing the decomposition and description of stimuli with multiple-sources in a similar way to how Bregman (1990) described the auditory streaming process in the psychoacoustic mechanism. The third experiment (see Section 2.1.7.3) introduces the important concept of ensembles of sources, which allows decomposition and analysis of potentially very competitive scenes. Between the second and third experiment, Berg dropped source depth as an attribute; split width into source width, ensemble width and room width; dropped externalisation and phase because they referred to a stimulus that is no longer used. The remaining attributes were used in the final experiment and are listed below and placed at their appropriate level with respect to the scene-based paradigm in Table 10, Table 11 and Table 12.

As can be seen by the above summary tables, the clarity and rigour in definition of spatial sound attributes that Rumsey introduced with his scene-based paradigm has mostly been absent in past studies.

The lack of a rigorous training method for the detection of, and discrimination between the various spatial aspects of these studies has inspired the author to investigate what is achievable with such a system.

This summary will be used in the creation of a new descriptive paradigm for use in the proposed spatial attribute training system, which is covered in Chapter 3.
2.2 Previous Work in Listener Training

In order to build a training system for application in the field of spatial audio, it is necessary to examine previously published studies into learning and training relating to audio. Section 2.2.1 examines general studies that have either focussed on the training of hearing skills or have incorporated audio reproduction into other training systems. Section 2.2.2 details previously published *Timbral* ear training systems, whilst Section 2.2.3 covers previous work in *spatial* ear training.

2.2.1 General Training Studies Involving Audio

A number of studies have looked into issues of learning, training or conditioning of subjects with respect to audio and hearing.

2.2.1.1 ITU-R Standards in Subjective Assessment of Audio: ‘Familiarisation and Training’

Before examining particular studies it is useful to see what standards pertaining to audio evaluation have to say on the matter of training.

Two ITU-R (International Telecommunications Union – Radiocommunications) standards (1994-1997, 2001) state the importance of training listeners involved in subjective assessment of audio. They both recommend a similar regime of what they call “Familiarisation or Training” (emphasis added). It is not entirely clear what is meant by the use of the conjunction ‘or’ in the standards, as their descriptions of the ‘Familiarisation or Training’ experimental phase demonstrate that subjects are both familiarised with the stimuli and are given the chance to train (by repetitively practising the tasks using the test equipment) before beginning with the main experiment.

The ‘MUSHRA’ standard (ITU-R 2001) explains that “[in] order to get reliable results, it is mandatory to train the subjects in special training sessions in advance of the test... The training should at least expose the subject to the full range and nature of impairments and all test signals that will be experienced during the test” (ITU-R 2001). In Appendix 1 of the MUSHRA standard, the ‘familiarization or training phase’ is described thus: “The first step in the listening tests is to become familiar with the testing process. This phase is called a training phase and it precedes the formal evaluation phase. The purpose of the training phase is to allow you, as an evaluator, to achieve two objectives” (ITU-R 2001) and goes on to explain that the two objectives are “to become familiar with all the sound excerpts under test and their quality level ranges; and to learn how to use the test equipment and the grading scale.” (ITU-R 2001). This author would describe the first objective as *familiarisation* and the second objective as *training* (albeit involving repetitive practice of the task – at least part of which would involve *procedural learning* of the experimental equipment and test paradigms – see Section 2.2.1.6).

In ITU-R BS 1116-1 (1994-1997) the ‘Familiarization or training phase’ is described as a chance for subjects “to become thoroughly familiar with the test facilities, the
test environment, the grading process, the grading scales and the methods of their use. Subjects should also become thoroughly familiar with the artefacts under study. For the most sensitive tests they should be exposed to all the material they will be grading later in the formal grading sessions... By the end of the familiarization process, subjects should have arrived at a stable sense of the scale that will be used in the formal grading phase which will follow familiarization or training” (ITU-R 1994-1997). Again the experimental phase is labelled ‘familiarization or training’ – the two terms are seemingly used interchangeably.

This author would like to draw the distinction between familiarisation which implies that elements of the task need to be presented to the subject so that they are allowed to calibrate themselves to the range of stimuli that they will encounter (so there are ‘no surprises’ during the test), and training which implies a regime of tuition and practice.

2.2.1.2 Kirk’s Preference Training Experiment

Kirk found that previous experience and conditioning influenced the listening preferences of subjects for various methods of stereo audio reproduction (Kirk 1956). Five programme items were played back through four different playback bandpass filter configurations (180Hz-3kHz, 120Hz-5kHz, 90Hz-9kHz and 30Hz-15kHz). The experiment involved an initial pre-test where preferences were gathered for the various stimuli from 210 music students of varying listening experience. The subjects were then separated into four groups – with two pairs of preference-matched groups forming two experimental/control group sets. The two experimental groups were then given a different treatment for six and a half weeks, one group only listened to the most narrow bandwidth reproduction, the other only to the widest. Thereafter all four groups (the two experimental groups and their control groups) repeated the initial preference test as a post-test.

The results of the pre-test were rather surprising – rather than preferring the widest bandwidth, the group tended to prefer the bandwidth 90Hz-9kHz. Kirk argued that this result was due to the listeners adopting a set (see also Section 2.3.1.9) where they were expecting to hear reproduced music rather than live music. However this author believes that technical limitations of the playback system employed at the time could have contributed to this result – perhaps objectionable signal distortions or other noises were present below 90Hz and above 9kHz in certain recordings.

The post-test results were also surprising – both experimental groups demonstrated an increased general preference for the type of playback that they had encountered over their respective control groups. Kirk also explained this result of set (see Section 2.3.1.9) - where the listeners in the experimental groups had grown accustomed to the particular method of playback to which they were previously exposed.

Kirk also presented the results of a further re-analysis of the preference data – although it is actually unclear which data set he analysed. Kirk divided his listeners once again, this time (seemingly arbitrarily) into three groups of listeners. The three groups consisted of those with less than one year of musical study (95 individuals), those with between one and seven years of musical study (102 individuals), and those with between eight and twelve years (22 individuals). The reason for the choice of division points at one and eight years is not explained in the paper. Neither
Background Theory – Listener Training

is the source of the data – this author presumes that Kirk re-analysed the pre-test preference data as it would be less representative to take the preference data from all users after various experimental manipulations. Kirk presents evidence that the group with less than one year of musical study (95 individuals) is statistically significantly more likely to prefer fuller bandwidth signals than the group with between eight and twelve years of experience (22 individuals), and concludes that this is further evidence that those who have been listening to reproduced music for longer adopt a set and expect to hear reduced bandwidth reproduction. Unfortunately Kirk does not provide information on the age of the students – it could be possible that the more experienced group was also older and had potentially different audiometric hearing curves. The choice of the age divisions is baffling and open to the criticism that the upper division could have been chosen in order to demonstrate an experimental effect. This is especially difficult to justify when one examines the widely different number of individuals in each group.

What can be drawn from Kirk’s study for purposes of the current work though is that by adopting a pre-post methodology with both experimental and control groups, he was able to accurately examine learning effects that occurred as a result of the specific conditioning between the pre- and post-tests.

2.2.1.3 Bech’s Selection & Training Experiment

Bech published the results of an investigation into the effects of repetitive practice and other factors (hearing threshold levels and experience) upon performance in a subjective audio evaluation experiment (Bech 1992). The experiment involved six iterations of the paired comparison of a total of four loudspeakers using four programme items by twelve subjects.

Bech (1992) analysed the consistency and sensitivity with which each subject rated the loudspeakers and compared this with their hearing threshold levels, previous experience and across their experimental iterations. All subjects that took part in the test had no greater than 15dB of hearing loss at any tested frequency, but Bech (1992) found no correlation between hearing loss and performance in the task for the subjects used in his tests. He also published evidence from a separate study that showed that far fewer experienced listeners than naïve listeners were required to produce statistically significant results in listening tests, by a factor of between two and seven (Bech 1992). Olive (2003) also showed that trained listeners were far more sensitive than experienced listeners and naïve listeners. For a further discussion of levels of experience and training see (Bech and Zacharov 2006).

Bech (1992) uses the term ‘training’ to mean repetitive practice of a task, but also to mean learning in general and to cover the rapid gaining of procedural learning that takes place at the beginning of experiments (see also Section 2.2.1.6). Bech states that the main purpose of ‘training’ is “to ensure that different aspects of the subjects’ performance have reached an asymptotic level” (Bech 1992). He examined the performance of his test subjects over six experimental iterations and concluded that the majority of his subjects had reached an asymptotic level (and therefore had acquired at least most of the procedural knowledge – see Section 2.2.1.6) after four experimental iterations – the other subjects achieving asymptotic level within six or potentially eight sessions (although he only conducted six trials).
### 2.2.1.4 Retholtz’s Training System for Sonarmen

An early example of an auditory training system was reported in (Retholtz 1961). Retholtz argued for the need to establish or improve selection and training standards and the creation of adequate benchmarks for US Navy sonar operators.

He described a method to train sonarmen to identify individual sound elements with the intention of training with combinations of the sounds at a later stage.

In the training system actual recordings of “passive” sonar signals (ambient sound of the sea rather than reflections from emitted sonic pulses) were replayed to trainees via a tape. It can therefore be seen to fulfil the constructivist idea that learning should be situated (Alessi and Trollip 2001) because the learning task closely resembles the real-world application. Retholtz also stated that the use of recorded stimuli would provide consistency in training that would not otherwise be possible with individual instructors.

Unfortunately, this author was unable to find any follow-up reports on the system including measures of how successful it was in training Sonarmen.

### 2.2.1.5 Zwislocki et al. Study into Practice and Feedback

Zwislocki et al. (1958) investigated whether practice and feedback would affect hearing thresholds in a series of experiments using separate groups of listeners.

Tests were conducted to measure the threshold of audibility of a 100 Hz tone. Without incentives for good performance and a standard payment rate, it was found that (over six weeks and one session per week) whilst subjects would improve their detection thresholds with practice, they routinely got worse within the course of a single session. Zwislocki et al. attributed this to a lack of motivation.

Additional sessions involving negative reinforcement (fining the subjects for giving incorrect answers), positive reinforcement (paying more money for correct answers) and giving the subject feedback allowed thresholds to be further reduced, but did not stop the intra-session deterioration experienced before.

By using fresh listeners, Zwislocki et al. managed to show that feedback and incentives were able to remove intra-session deterioration, but only if the subjects had not already been “pre-habituated” to non-incentive conditions.

Using yet another set of listeners, Zwislocki et al. were also able to show that pretraining with a 1000 Hz tone did not transfer to increased sensitivity with 100 Hz. They were able to motivate the subjects after they had been “pre-habituated” without incentives and feedback by including incentives and feedback after just four sessions.

In order to check that these effects were not due to listeners selecting their own audibility threshold, a new paradigm was used. A further six listeners were tested using a more objective method of detecting thresholds where they would state whether one or two stimuli had been replayed. All subjects were motivated and improved throughout the test.

In a final test, a further ten subjects were trained with no monetary rewards. The threshold measuring paradigm was changed to one where the threshold level was tracked. On the whole, listeners improved their thresholds by about 10 dB.
These experiments are useful because they show that prolonged practice can be beneficial for hearing tasks. The study also showed evidence that motivational techniques can counteract boredom sometimes felt during listening tests.

2.2.1.6 Hawkey et al. Initial Learning Experiment

In recent experiments by Hawkey et al. (2004) "procedural" and "perceptual" learning were defined. "Learning that is specific to the perceptual judgment as 'perceptual' and contributions to improvement resulting from aspects of task performance not involved in making a perceptual judgment as 'procedural'".

Hawkey et al. questioned what they described as established notions that learning which takes place early in a task (so-called "initial learning" is accounted for by procedural learning and that perceptual learning happens once procedural learning has taken place).

They assembled four groups of 20 people and tested for frequency difference thresholds at 1 kHz using a "2I-2AFC" (two interval, two alternative forced choice) experimental paradigm (where subjects needed to correctly identify the tone that had higher frequency). However each group was trained in a different way. The first group was trained using the target paradigm. The second group was trained in listening for frequency differences using an AXB paradigm (3 successive tones were played. The subject was then required to identify which tone was different to the other two). The third group was trained to recognize level rather than frequency using the target paradigm. The fourth group was trained on visual images using the target paradigm (they had to say which one had more contrast).

Each subject was tested individually in two-hour sessions where they went through a round of training followed by two tests using the target paradigm and task.

By testing each of the groups using the target paradigm after their training, it was possible for Hawkey et al. to show that the target group had reached an asymptotic level during the training phase, and that the training received by the second group (using the AXB paradigm) had transferred to the target task. The two groups trained in the procedure but not the perceptual task showed signs of improvement during the tests. The difference between this behaviour and that of the subject trained using the target task and using the same perceptual task but different paradigm showed that "perceptual" learning as well as "procedural" learning must be taking place during initial practice sessions. Indeed, they were able to quantify this, stating that 76%-98% of initial learning improvement was due to perceptual learning, whereas only 2%-24% could be attributable to procedural learning.

2.2.1.7 Drennan & Watson's Extended Training Experiment

Drennan & Watson (2001) reported findings of experiments that investigated whether subjects who had poor sensitivity to 'spectral shape' (the frequency content of stimuli) "could eventually acquire finer sensitivity". The effect of extended training on "good" and "poor" subjects was also investigated. Initially 41 subjects were judged on their ability to discriminate various sounds using Watson's TBAC (Test of Basic Auditory Capabilities) – described in (Surprenant and Watson 2001). Five subjects were added to the group and the 46 subjects then tested for their
threshold of discriminating differences in frequency content of a complex signal consisting of eleven tones was judged. After an average of 2000 trials they were able to select seven “good” judges and four “poor” judges of “spectral shape” in order to test the effect of extended training on their performance.

In the extended training experiment, the subjects were given a further 2000 trials of the same complex signal used previously, followed by 6000 trials using three other complex signals, followed by a further 1000 trials of the original signal. Due to “confounding effects of the order of training” (not explained in their report), Drennan & Watson (2001) only presented the data for the original complex signal.

The “good” listeners did not continue to improve after the first 2000 trials of training, reaching an asymptotic performance. Most of the “poor” listeners (three of the four) continued to improve during the 9000 trials, and only one of the four “poor” listeners improved to the level of the “good” listeners.

The experiments performed by Drennan & Watson like those of Retholtz (see Section 2.2.1.4) did not involve active learning (Alessi and Trollip 2001). The stimuli were played to the listeners as 400ms signals with 500ms gaps between them. Subjects were therefore neither able to pace the sessions for themselves nor interactively switch between stimuli in order to gauge the differences between them. It is likely that if subjects had active control over the stimulus presentation, their motivation and concentration levels would have been higher (Alessi and Trollip 2001). This can be seen as a common factor for all subjects in the experiments, but it would be unwise to draw significant conclusions from this data as their findings may not pertain to an interactive training system.

Drennan & Watson showed that extended training of passive learning in timbral detection threshold tasks can help certain (“poor”) listeners achieve a better sensitivity, but does not necessarily benefit other (“good”) listeners. Time savings can thus be achieved by monitoring the point at which asymptotic performance is achieved.

2.2.1.8 Jones et al. Close-Combat Trainer Involving Spatial Audio

Jones et al. (2005a, b) investigated the use of spatial audio in a training system for military close-combat operations. Subjects were asked to navigate through a building displayed visually using a virtual reality headset. Audio, when present, consisted of alerts that would warn the subjects if they were in dangerous areas (such as windows or doorways) and various sounds to help them identify friends and foes in the building. 36 subjects were randomly divided into four groups, one for each of four audio conditions. The conditions were: no audio, non-spatial audio, generic HRTF (head-related transfer function) and best-fit HRTF. The performance measures used related to the amount of time subjects would spend in dangerous areas, and how quickly they would identify and “clear” hostile units. Subjects were tested once before training and once again after training. The results show that spatial audio (via either generic or best-fit HRTFs) allowed subjects to perform quicker than non-spatial audio or no audio at all. Regarding training effects – all subjects across all conditions seemed to show similar improvements in post-training tests. The only exception was the time spent by subjects in “door” entry areas, where the non-spatial audio seemed to confuse subjects as to where they were supposed to
move and hence caused them to not improve with training, whereas the spatial audio and non-audio conditions improved with training.

Overall the training effects seem to have been dominated by an overall improvement by all subjects in all conditions. Jones (2005) attributes this to the fact that “the trainees were not trained enough on how to interact with the system before the initial performance measure was taken”. Thus, because the subjects were so new to the system, performance increases due to familiarity with the control system post-training were potentially obscuring training effects due to the type of audio reproduction employed. This has an important ramification for training systems, namely that potentially important between-subject training effects (particularly with respect to timing) should be isolated by allowing the subjects to get familiarised with the control systems before the initial pre-training performance measure is taken. Using the definitions stated by Hawkey et al. (see Section 2.2.1.6), it can be hypothesised that procedural learning dominated the initial learning results of Jones et al.’s study.

Care must be taken to avoid the dominance of procedural learning in experiments connected to this research. If initial pretraining familiarisation sessions are employed, procedural learning effects should be minimised within the main experiments.

2.2.2 Previous Timbral Ear Training Systems

This author has found five published “timbral” ear training studies, and has attended the presentation of Moulton’s ‘Golden Ears’ training programme (see Section 2.2.2.5). Only Quesnel (Section 2.2.2.4) publishes any experimental verification of his system, the others simply cite anecdotal evidence of the effectiveness of their programmes. Everest’s (Section 2.2.2.2) and Brixen’s (Section 2.2.2.3) make use of passive learning cassettes with limited active learning. Moulton’s course is run purely from CD (Section 2.2.2.5). Quesnel’s (Section 2.2.2.4) and Olive’s (Section 2.2.2.6) are multimedia learning tools (Jonassen 1988; Alessi and Trollip 2001; Mayer 2005), with administration, data collection and tracking handled via computer.

2.2.2.1 Warsaw Academy of Music’s Timbre Solfege

The Timbre Solfege system of timbral training at the Academy of Music in Warsaw, Poland has been reported in three papers (Rakowski and Trybula 1975; Letowski 1985; Miskiewicz 1992).

Timbre Solfege is a three year course designed to improve timbre sensitivity and memory. These two abilities are useful for sound engineers because they need to try to recreate the sound in the recorded venue, and need to recognise and remember how the live sound differs from the sound in their control room. Although the course appears to have evolved over 18 years spanned by the three papers, the course has two main task groups: active tasks and passive tasks.

Active tasks involve active learning (Alessi and Trollip 2001). The instructor (unseen by the students) alters a signal by using any one or any combination of pieces of studio sound processing equipment (most commonly a 1/3rd octave band parametric EQ). Students are then required to mirror the changes to a second copy of
the original signal using identical pieces of equipment. These tasks are "usually well received" (Letowski 1985) by the students.

Passive tasks require the student detect and describe timbral differences between sounds and eventually to listen and describe absolute timbral properties of a signal without a reference. Initially at least the students learn from how the instructor describes the signals.

Training is based within the framework of international octave bands from 63 Hz to 16 kHz, further subdivided into intervening 1/3rd octave bands later in the course. This has the advantage of referring to direct physical parameters that sound engineers can alter using equalisers. It is therefore practically useful to them.

The programme is described in some detail in (Miskiewicz 1992), and even includes "spatial hearing" (which appears to be mostly related to detection in timbral changes in reverberation). Tasks in the course get gradually more challenging throughout each year, with difficulty being increased through increasing the number and type of changes occurring to the sounds.

Unfortunately there is no published formal evaluation of the Timbre Solfege system. All evidence of its success is anecdotal. Letowski claims that "almost every student improves greatly in both [timbre sensitivity and timbre memory] after several laboratory classes", and that timbre perception skills are "easy to gain" but also "easy to lose" and need to be regularly practiced. Miskiewicz (1992) claims that "experience with the Timbre Solfege course makes it evident that technical listening skills may be improved by systematic training" and that "reports from graduates working in the field attest to the success of the course".

It may be unreasonable to expect formal testing with a control group that was not trained as this would involve potentially damaging the learning of the control group. Formal evaluation of the Timbre Solfege course appears not to have been a priority for the course organisers.

Timbre Solfege includes familiarisation of the theory behind the perceptual tasks involved, passive learning tasks, some active learning tasks and has anecdotal evidence of success.

### 2.2.2.2 Everest’s ‘Critical Listening’ Ear Training Programme

Everest (1982) published details of his "Critical Listening" course. It is a programme of ten lessons covering various technical and timbral distortions presented over audio cassette with an accompanying booklet (which contains a written-out version of the cassette narrator’s script along with explanatory diagrams). Everest explains about the necessity of comparative judgements with a reference, but describes only one test (in the final lesson) that is presumably only prepared in one configuration – this would be of little use as a repetitive practice drill. There is no published experimental verification of the effectiveness of the programme. As Olive (2001) points out, Everest’s course “serves more as an introductory course for naïve listeners”.

- 60 -
2.2.2.3 Brixen’s Spectral Ear Training Programme

An often-quoted study of timbral training is that published by Brixen (1993). However, remarkably little detail is included in the paper making it difficult to decipher how courses were put together. Unlike the Timbre Solfège system (see Section 2.2.2.1), it seems to be mostly based on passive learning (relying upon 3-4 weeks of self-study using supplied materials on cassette tape), but does have 2 days of training with a computer system that can process signals in various ways using a MIDI (Musical Instrument Digital Interface)-controlled parametric equaliser. It is unclear whether the computer system was exclusively operated by the instructor to create degradations for groups of subjects or individual subjects to listen and respond to, or whether individual subjects would use the computer system and listen to the results of their manipulations. There appear to be no active tasks as were found in the Timbre Solfège system (Section 2.2.2.1). The training was based around 10 standard octaves (from 32 Hz to 16 kHz). Tasks for participants included identifying altered frequency bands (in terms of the actual frequency bands or the perceived timbre) and using a “sweep mode” to identify changes to the spectrum (Brixen 1993). Whilst there is a description of how the changes were created, there is no description of how they were tested. There is evidence that the programme started with easier changes to spot that became increasingly subtle, and that theory was also taught to the subjects in-between practical sessions. The reliance upon pre-prepared cassette tapes was probably an unavoidable limitation at the time, but with the recent proliferation of computers, it would be conceivable that a modern version of such a system could be delivered using an interactive computer presentation that could harness the benefits of active learning whilst not being as predictable as a pre-prepared cassette.

Brixen’s Spectral Ear Training Course was designed specifically to train engineers quickly and using the minimum number of contact hours, whilst using the equipment available at the time. Brixen says that his system was “widely used” and that “It has been experienced that to some extent it is possible to establish an ability of being a ‘human spectrum analyzer’, within the frame of short-time courses”. Unfortunately there is also no formal evaluation of the effectiveness of the system presented by Brixen, so it is not possible to say by how much the system helped people, or to suggest any improvements.

2.2.2.4 Quesnel’s Timbral Ear Trainer

The timbral ear training system at McGill University (Quesnel and Woszczyk 1994; Quesnel 1996; Quesnel 2001) combines the use of computer-controlled equaliser found in Brixen’s system (see Section 2.2.2.3) with active and passive tasks inspired by the Timbre Solfège system (see Section 2.2.2.1).

The first paper (Quesnel and Woszczyk 1994) describes the system and reports on a formal evaluation of training effects. The second paper (Quesnel 1996) describes an advanced performance tracking and training management system that was later implemented in the system, but contains no further formal evaluation of the system.

The system is based around a computer that administers the tasks and controls two parametric equalisers. Active tasks involve “comparative listening” and “bring-to-flat” exercises. In comparative listening tasks students are required to match EQ modifications to one signal performed in secret by the computer by adjusting one or
Background Theory – Listener Training

more EQ settings on a similar equaliser via the computer interface. In bring-to-flat exercises students were required to remove modifications made by the computer to one channel by applying inverse filter(s) to the second channel. Passive tasks involved specifying frequency bands that were modified or categorising the changes in terms of their similarity to spoken vowel sounds. Stimuli included broadband noise and music. The difficulty of the tasks was gradually increased over the 8-month training period.

The system was evaluated using seven students but importantly did not include a untrained control group (to see what effect no-training would have upon the results). Criteria for the evaluation were speed and accuracy. The tests were administered at the beginning and end of the eight-month training period and included ten comparative listening tasks (white noise in the pre test, and a snare signal in the post test), five bring-to-flat tests involving two 12dB cuts and/or boosts in a cello recording and ten absolute identification tasks involving one cut or boost in a snare track.

The results show that during comparative listening tasks many perfect scores were achieved in the pre and post tests, therefore showing not much in the way of improvement. However because different signals were used, important interactions may have been missed. An improvement in the design of the experiment would have been to include both noise and snare signals in the pre and post tests. If the snare was found to be too difficult during pre-training, then this would have shown up as a useful training effect. The bring-to-flat and absolute identification tests showed that improvements were made by all subjects. One listener who had taken part in previous timbral ear training achieved perfect scores for the bring-to-flat tests pre- and post-training and had the highest initial scores for absolute identification (but also improved the least). This shows that this subject was able to retain some of their previous training.

Regarding speed as a measure of effectiveness, it was clear that students who practice most were not necessarily quicker in the tests. Also the slowest subjects in the post tests were not necessarily the slowest working in the practice sessions, they were probably being more careful in the tests. Quesnel (1996) makes the point that speed was not systematically enforced as a criterion during training and included it in the second paper. An important outcome is therefore that factors for assessment should be enforced during training. Notwithstanding, response time did improve overall between pre- and post-training tests.

Whilst no control group was used, Quesnel claims to have controlled for course-based factors affecting students because students from courses other than their Tonmeister course were included in the training group.

In a later verification experiment Quesnel (2001) compared the performance of five university students that had been trained using his timbral trainer (for an average of 2.4 years) with five "audio professionals ... with an average of 19 years of experience working in the audio engineering field and 2.2 years of training in audio". Quesnel explains that the experienced listeners did not have formal aural training, yet also states that they training in audio without further explanation of the nature of this training, or whether it was comparable to that which the students had received. Quesnel labelled the five experienced listeners a "control group", although this was not a control group as would have been understood by Kirk (Section 2.2.1.2), or
Hawkey et al. (Section 2.2.1.6). The task involved what appears to be a 'bring-to-flat' task although one of Quesnel’s stated research questions regarded testing the validity of his training system by using a task that was similar to what would be experienced in the industry. This author contests Quesnel’s claim that this task was more than a variant on the training tasks.

Quesnel was however, able to demonstrate that the student subject group performed better than the experienced subject group. The small group size and questionable task and group composition mean that his findings need to be treated with caution.

Overall then, factors included in the evaluation need to be enforced during training. Control groups should be included and a spread of difficulty options should be included in the pre- and post-tests in order to avoid missing important training effects.

### 2.2.2.5 Moulton’s ‘Golden Ears’ training programme

Moulton (2007) developed a compact disc (CD) -based ear training system called “Golden Ears”. He identified that technical listening tuition and audio equipment review media contains jargon, with assumed meanings, private audio terms and exclusive knowledge. The “Golden Ears” system aims to provide a unified and unambiguous approach to describing sound. Golden Ears is orientated towards sound engineers and so mostly involves commonly used audio engineering tools (like EQ, delay and reverb). Moulton recommends that it be exclusively run using loudspeakers, and not using headphones (the reason he gave was that headphone listening would be too easy).

The system is supplied on CDs containing a number of training drills involving recordings in three sections: A-B-A, where “B” features an alteration (which can eventually be any number of different alterations) and “A” is the original unprocessed item.

A varied selection of degradations is potentially applied to the training signals. Timbral alterations include octave boosts and cuts, 1/3rd octave boosts and cuts and eventually 2 bands altered at the same time. For “reverb” training, a snare drum recording is used with a varying reverb time and pre-delay. Left-right delays can also be applied as can dynamic range compression.

Because of its reliance on pre-prepared CD media, the “Golden Ears” course is inflexible, relying on pre-prepared training stimuli. However the number of different potential degradations maintains interest in the programme.

Moulton (2007) offers no experimental verification as to whether his training programme helps individuals perform better as balance engineers, but claims that repeated practice of the course drills aids performance on the trials themselves.

### 2.2.2.6 Olive’s ‘So you want to be a Harman Listener?’

Olive (2001) identified the need for a computer program that could be used by subjective evaluation panel trainees in order to allow them to classify and rate audible artefacts commonly found in loudspeakers, and to train them to “report these problems in precise and unambiguous terms”. Olive (2001) contended that only a
computer-based training programme could overcome the limitations of previous compact-disc-based systems. He explained that computer-based training could provide tasks that can adapt to the trainee’s skill level, would not be limited to a finite number of test signals, could handle randomisation and blind presentation of trials and could collect and analyse training data.

The resulting software program “So You Want to be a Harman Listener?” (SYWTBAHL) provides a variety of different tasks based around the detection of frequency-based distortions of stereo audio signals. Trainees need to compare the unprocessed original signal (known as the “flat” signal) and a processed signal (the “EQ” signal), then specify which type of distortion is present. In order to start a training task, trainees select the type of degradation to be studied and the skill level and stimuli to be studied.

SYWTBAHL can be seen (at least in part) as an automated version of the Timbre Solfege programme (see Section 2.2.2.1), in that it requires comparison of timbrally altered signals. Because the nature of the degradations is known - subjects know that they will be hearing either bandpass-filtered items, or items with peaks or troughs (or both) of say 6dB) – the possible degradations are not as complex as Moulton’s “Golden Ears” programme (see Section 2.2.2.5).

Most of the tasks use a visual representation of a logarithmic frequency graph which shows the number of possible degradations. See Figure 1 for a screenshot of the user interface.

Feedback is presented immediately in the form of a dialog box with a smiling cartoon face that informs the listener if the answer they gave was correct. If they were “incorrect” they are notified (visually) of the actual correct answer. There is unfortunately no way (in the current version of SYWTBAHL) for the trainee to listen back to the EQ and FLAT audio files once an incorrect answer has been given. For
most of the tasks the trainees need to do at least four trials and get at least 80% correct. Up to two incorrect answers (strikes) are also allowed, but if a trainee answers three times incorrectly on the same skill level the training session is brought to a close. Once at least four trials have been answered and an average score of 80% has been achieved the trainee progresses to the next level, whereupon the number of possible degradations increases and the number “strikes” is cleared, allowing for a maximum of another two incorrect attempts. A number of elements were included to make it “more entertaining and fun to use, encouraging its continued use” (Olive 2001), these included a fictitious stock price indicator that charts trainees’ progress and three “training aides” that could be called upon (once each) to provide different levels of guidance on the correct answer.

In addition to the main trials there is also a preference test mode, where subjects are asked to listen to four sound files: one is the “FLAT” reference file along with three degraded versions. The software keeps track of the scores given by the subjects to each of the stimuli and a measure of their consistency – the “t-statistic” (Olive 2001) – is calculated and displayed in the stock-price indicator.

Olive (2001) does not make any mention of whether or not the programme has any positive benefit on listeners’ performance in the rating of loudspeakers, or of any verification tests run to quantify such improvements.

2.2.3 Previous Spatial Audio Ear Training Systems

The author can find evidence of four previously published training studies involving listening to spatial audio, or the spatial aspects of audio reproduction.

2.2.3.1 Zacharov & Koivuniemi’s Spatial Ear Training

As part of their elicitation study into spatial audio attributes (see Section 2.1.6, and specifically Section 2.1.6.4) Zacharov & Koivuniemi claim to have trained their listeners in listening to specific spatial audio attributes through the use of stimulus sets that “indicate” the attribute and polarity of the scale, supplemented with a verbal description (Koivuniemi and Zacharov 2001). However, details of the training system were never published.

2.2.3.2 Corey’s Reverberation Matching Training System

Corey (2004) details what could be considered an extension to the Timbre Solfège programme (Section 2.2.2.1) and Quesnel’s timbral trainer (Section 2.2.2.4) for matching artificial reverberation parameters (rather than perceptual spatial attributes). The computerised training programme allows subjects to adjust reverberation parameters such that one reverberant signal matches another. Because of the reliance on adjusting the artificial reverberation parameters – rather than perceptually unidimensional spatial attributes (Neher 2004) – this method is more suitable for the direct training of sound engineers who will be utilising artificial
reverberation equipment than for a generalised system involving a universal language for perceived spatial audio attributes.

Unfortunately Corey (2004) does not provide any experimental verification regarding the effectiveness of his training system.

### 2.2.3.3 Merimaa & Hess: ‘Training of Listeners for Evaluation of Spatial Attributes of Sound’

Merimaa & Hess (2004) reported a study where they claimed to have trained naïve subjects in spatial audio attribute evaluation. It is, however, more a study about the terms ASW (auditory/apparent source width) and LEV (listener envelopment) than a study about training. ASW and LEV are attributes used mainly in concert hall acoustics (Beranek 1996) as a way of classifying two subjective attributes thought to describe room acoustics. However, as Rumsey (2002) pointed out, spatial audio reproduction can evoke spatial perceptions that are not covered by ASW and LEV alone.

Sixteen naïve subjects took part in the experiment. They were paid and therefore extrinsically motivated – see Section 2.3.2.1. Each subject listened to a number of sound stimuli before being asked to discuss the terms ASW and LEV with each other (with the authors moderating the discussions). Binaurally processed sound was reproduced using headphones – albeit without “head-tracking” (Merimaa & Hess 2004), within an isolating booth. All subjects took part in the ‘training’ therefore there was no control group (which could have been utilised to show any differences between pre- and post-training). Each subject also took part in an evaluation phase in two different experiments. In ‘Experiment 1’ each subject separately rated the ASW and LEV of 12 stimuli a total of five times each on a scale that ranged from 0 to 6 (no evidence is published of whether there were any decimal points allowed in the values). In ‘Experiment 2’ 15 of the previous 16 subjects were asked to repeat the evaluation this time using a graphical method to simultaneously evaluate ASW and LEV. There is evidence that at least two trials were attempted (as the first one was used as a practice run). Merimaa & Hess look for evidence of learning in changes in the way in which subjects evaluated the spatial audio over the various sessions, and whether the subjects agreed with one another. They were not able to conclude a great deal from their experiments, as the subjects did not show much in the way of variation throughout the trials, and had tended to rate ASW and LEV in a highly correlated fashion.

There are a number of problems with this particular study. This author doubts whether the reproduced sound would have conjured a sufficient sense of ASW or LEV due to the visual cues associated with being seated inside an isolating booth, and the use of binaural playback where the head is free to move, but the soundfield is not updated due to those movements. Because no control group was used, Merimaa & Hess were not able to show data from untrained listeners in order to gauge the effect of their training – it could have had no effect, or could even have been detrimental to the subjects. Because no differences could be found across the five repetitions during Experiment 1 the (naïve) subjects could be said to have been relatively consistent. This could imply that the tests were too straightforward for the naïve listeners to require any specific training – even though Merimaa & Hess (2004) stated that “the experiment was not easy”. This author would suggest that as the
Background Theory – Listener Training

subjects were relatively consistent in the face of adverse sound reproduction conditions, it is likely that they could tell the different processes apart (by reverberation characteristics, for example), but were not able to get a sense of envelopment or perhaps even the apparent width of sources within the room, which is why they rated the two attributes in a similar way. It is also possible that the source material was inappropriate for demonstrating differences between the two attributes.

The main problem as far as the research presented in this thesis is concerned is that the training programme employed did not constitute a programme of training. Instead it allowed naïve subjects to come to their own understanding of what unfamiliar and provided terms mean. Provided terms should be exemplified using clear standards, or subjects should be allowed to use their own terms to describe their experiences.

In summary though, little can be learned from this study because it used problematic reproductions of spatial audio and no formal training programme.

2.2.3.4 Neher’s Spatial Trainer

In Appendix E of his thesis, Neher (2004) described an investigation into training using two of the unidimensionally varying spatial audio attribute stimulus sets that he had simulated for such purposes.

Time constraints allowed him to employ just five naïve subjects who were paid to take part in the experiment (paying listeners is considered extrinsic motivation and is not considered to be as effective as intrinsic motivation – see Section 2.3.2.1). All five took part in a ranking task where they had to place five Source Width and nine Source Distance stimuli in order of their specific spatial attribute.

The subjects were then separated into two groups, one that would undergo additional training, and one that would act as a control group. Unfortunately Neher (2004) does not leave information about how he separated the subjects. The training group had three subjects, the other two formed the control group.

The training group subjects were then given a short tutorial and familiarised with the stimuli before being allowed to progress through three types of practice drill at their own pace. All drills at all levels needed to be completed for the subjects to be considered ‘trained’. The five Source Width and nine Source Distance stimuli were used to create various difficulty levels of stimuli, with the easiest difficulty level utilising the two extremes of each attribute simulation, and the most difficult utilising every stimulus within the attribute simulation. Drills consisted of discrimination (whether two stimuli the same or different), pairwise ranking (which source in one of two stimuli was either wider or further away), and multi-stimulus ranking (put a number of stimuli in a rank order of source width or distance). Each drill had 10 trials and had a pass mark of 80% (similar to Olive’s SYWTBAHL – see Section 2.2.2.6), however once the subject had answered 4 incorrectly the drill would be aborted. Drills were administered by Neher (he would load each training difficulty level for each subject), but the trials within each drill were presented by the computer which also displayed the current score visually. Visual and auditory feedback was presented through well-known cartoon characters.
The effects of the training program were demonstrated by testing each of the subjects with the original task – that of ranking all levels of the source width and source distance stimuli. No further tests were attempted to investigate possible transfer of any trained skills. Because there was an assumed correct order for the stimuli (they had been simulated and verified in a certain rank order) it was possible to check how 'correct' each of the responded rank orderings were. Neher measured this using the Sum of Squared Euclidean Distances (SSED, although Neher used the term “SED” (Neher 2004) for this measurement). The lower the SSED of a particular rank ordering is, the closer it is to the 'correct' rank order. The time taken for the tests was also recorded in minutes.

Neher did not specify how he separated the subjects into groups, but by utilising an odd number of participants, Neher’s subject groups were uneven in numbers and with so few subjects it was not possible for him to draw any firm conclusions about the programme.

In general the results show that the trainees tended to achieve better scores – sometimes perfect and never worse – after training. The control subjects did not improve greatly, actually got worse in one case and never achieved perfect scores.

The most interesting results come from one of the ‘training’ group subjects who achieved a perfect score in the source distance pre- and post-tests, and also did not improve upon his SSED of 2 between the source width tests. Whilst not improving upon their SSED scores, this subject was slowest in both pre-tests but got quicker in both post-tests. This shows that whilst this subject did not improve, they actually became more confident in the answers that they gave. However, that one naïve subject was able to get a perfect score on the source distance test shows that the test should have been more difficult in order to warrant training in the first place.

In summary several things can be learned from Neher’s training experiment. By using a control group as well as an experimental group in a pre-post test methodology the experimenter has the ability to compare the performance both before and after training as well as control for any learning effects as a result of repetition of the pre- and post-test tasks. Like Olive (see Section 2.2.2.6), Neher also used an 80% pass mark for his training drills. Pre- and Post-tests need to be difficult enough to warrant training, as learning effects will be missed if the pre-test is too easy. It is important to note that Neher had confined his training and testing to the same stimulus sets – however a training system would need to show that learned skills could transfer beyond the training environment (see Section 2.3.1).

### 2.2.4 Summary of Work in Listener Training

The timbral and spatial training systems outlined in Sections 2.2.2 and 2.2.3 are anecdotally successful, but only Quesnel (Section 2.2.2.4) and Neher (Section 2.2.3.4) provide evidence that their training systems helped the subjects improve.

The standards (Section 2.2.1.1) recommend that tasks be practiced to ensure that procedural learning occurs. Repetitive practice of a task has been shown to be beneficial to listening skills by Bech (Section 2.2.1.3), Zwislocki et al. (Section 2.2.1.5) and Drennan & Watson (Section 2.2.1.7). Drennan & Watson showed that performance benefits through practice tended to reach an asymptotic level, and Bech
stated that the goal of repetitive practice is to allow subjects to reach this asymptotic level.

Many of the training systems (for example Timbre Solfege – Section 2.2.2.1, Quesnel – Section 2.2.2.4 and Corey – Section 2.2.3.2) utilise changes in physical parameters that equate to the manipulation of controls on sound engineering equipment (because the goal of these systems was the training of sound engineers). The focus of Neher (Section 2.2.3.4) was for the simulation of perceptual spatial audio attributes for the purpose of training subjects to describe how they perceived spatial audio reproduction.

The standards (Section 2.2.1.1) call for a familiarisation phase in which subjects are exposed to the range of stimuli that they will encounter in a test (a form of internal calibration). Jones et al. (Section 2.2.1.8) reiterates the need for familiarisation of subjects with the procedural elements of the task. The need for familiarisation must be fitted within given time constraints and balanced against the potential pre-biasing of control groups.

The need for matched control groups was demonstrated by Kirk (Section 2.2.1.2), whose method allowed experimental effects to be observed and measured. Most of the training systems in Sections 2.2.2 and 2.2.3 did not include comparison with a control group.

The value of pre-testing and post-testing subject groups either-side of experimental treatments can be seen in Kirk (Section 2.2.1.2, where pre-testing allowed for separation into matched experimental groups), and Hawkey et al. (Section 2.2.1.6), who was able to gauge experimental effects in his post-tests.

Motivational techniques (the subject of Section 2.3.2) where shown by Zwislocki (Section 2.2.1.5) to help counteract boredom.
2.3 Transfer & Motivation

How well learned knowledge and skills transfer from the training environment to other situations will define how useful the training system is seen to be. In order to evaluate a training system for spatial audio attributes, the degree to which the spatial audio listening skills can be transferred to other situations needs to be investigated. Relevant literature involving the transfer of learning is therefore detailed in Section 2.3.1.

As indicated by Zwislocki (Section 2.2.1.5), the motivation of subjects is likely to play an important role in the effectiveness of the training programme. Relevant literature involving motivation in learning is therefore summarised in Section 2.3.2.

Some of the content of this section has been published previously in (Kassier, Brookes and Rumsey 2006b).

2.3.1 Transfer of Training

If spatial audio listening skills learned through training can be used in situations outside the training context, then a case can be made for their wider applicability.

2.3.1.1 Definition of Transfer

Various authors have attempted to describe or define transfer. Ellis (1965) and Wittig (1981) use very similar, generalised descriptions that could imply transfer between very similar environments. Detterman (1993) is clearly concerned with transfer that occurs between two quite different conditions:

"Transfer of learning means that experience or performance on one task influences performance on some subsequent task" (Ellis 1965).

"Transfer of training ... describes situations where the learning of one task influences the later acquisition of some other task" (Wittig 1981).

"If two situations where the same behaviour occurs are obviously different in important ways, interest is in transfer" (Detterman 1993).

Transfer is often subdivided into two categories: near transfer and far transfer. Clark and Voogel (1985) explain that near transfer refers to target contexts that are similar to the training setting, whereas far transfer is achieved when skills are applied in "very different" contexts to the trained one. Detterman (1993) distinguishes near transfer as involving identical situations apart from specific differences, from far transfer which describes a "continuum of situations progressively more different from the original learning experience" (Detterman 1993). Near/far transfer in the pilot referred to performance in the training and ecologically valid tasks.

Whilst Ellis (1965) does not make the near/far transfer distinction, he draws the distinction between positive transfer (which aids the target task), negative transfer (which hinders the target task) and zero transfer (which either indicates that no effect has occurred, or that positive transfer has been cancelled by any negative transfer present). It can be said that, for the most part, the training system used in the pilot
study resulted in zero transfer (because there was no significant performance improvement over the control group).

Transfer can also be considered as specific or general (Hulse, Egeth and Deese 1980; Detterman 1993). Specific transfer involves skills from the initial task aiding learning in the target task. General transfer is that which occurs not as a result of specific elements in the original task. General transfer involves warm-up and learning to learn (Hulse, Egeth and Deese 1980).

In addition, Detterman (1993) also describes the difference between Surface structure (similar controls or overall views) and Deep structure (similar internal workings) in training studies.

2.3.1.2 Transfer of learning: Experiment designs

Wittig (1981) and Hulse et al. (1980), describe a basic transfer experiment design identical to that presented in (Ellis 1965).

In the simple transfer experiment an experimental group learns a certain task. Thereafter a control group and experimental group perform a target task. The difference in performance between the two groups can therefore be attributed to transfer between the two tasks that the experimental group achieved. It is important for this experimental design that the groups be “equivalent with respect to factors important in learning the tasks” (Ellis 1965) – Hulse et al. (1980) suggest randomly selecting the control and experimental groups.

Ellis (1965) goes on to describe four additional transfer test paradigms, all of which are shown in Table 13.

<table>
<thead>
<tr>
<th>Test</th>
<th>Group</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Experimental</td>
<td>Learn A</td>
<td>Learn B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>(rests)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Experimental</td>
<td>Pretest B’</td>
<td>Learn A</td>
<td>Learn B</td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td>(rests)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Experimental</td>
<td>Learn A</td>
<td>Learn B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Experimental</td>
<td>Learn A</td>
<td>Learn B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Experimental</td>
<td>Learn A</td>
<td>Learn B</td>
<td></td>
</tr>
</tbody>
</table>

Task "A" is normally the initial task, task "B" is normally the transfer task.

Experiment 1 is not particularly useful for working out exactly what “A” does for “B”, as general factors are not controlled. Experiment 2 uses part of the target task to pre-test both groups (allowing for similar ability groups to be assembled) and control for certain specific factors (see ‘warm-up’ in Section 2.3.1.4). The experimental group will still undertake more practice than the control group on the whole. Wittig (1981) suggests using a ‘filler task’ that does not have the specific features of the original “A” task as a means of controlling practice. Experiment 3 is supposed to be useful for inter-sensory transfer experiments (Ellis 1965), but this assumes that transfer will be symmetrical from “A” to “B” and from “B” to “A”.

- 71 -
Experiment 4 compares transfer to two similar but not identical transfer tasks. This controls 'warm-up' and practice. Experiment 5 uses time intervals between task “A” and “B” in order to investigate temporal issues. Control groups can be created by not providing them with task “A”.

2.3.1.3 Specific factors in transfer

Specific factors are dependent on the nature of the original task and how that affects the transfer task.

Task Similarity

Transfer is aided if the training and transfer tasks are similar. According to a study by Osgoode – cited in (Ellis 1965) – identical conditions maximise transfer. In addition, if responses are the same, positive transfer will result from similar stimuli, negative transfer will result from “antagonistic responses to identical stimuli” (Ellis 1965), but new responses to previous stimuli will not necessarily result in negative transfer. Near transfer is therefore easier to achieve than far transfer.

Variety of Previous Tasks

A study by Duncan – cited in (Ellis 1965) – showed that a small increase in the variety in original tasks resulted in an increased positive transfer. The increase was largest from 1 to 2 tasks, progressively less from 2-5-10 (Ellis 1965).

Time Interval between Tasks

If memorisation is not required, the time-elapsed between training and transfer tasks does not seem to be an issue (Ellis 1965). As memory is not the focus of this study, it is not expected that time differences between subjects’ experimental sessions need to be strictly controlled in the training system.

Degree of Original Learning, Intelligence and Motivation

Ellis (1965) recommends extensive practice of the original task, as this reduces the chance of negative transfer. He also advises that intelligence and motivation are factors in transfer. Testing for and controlling intelligence and motivation is a challenge with any type of test. By selecting subjects from existing student groupings (such as first year sound recording students at the University of Surrey) should give a certain level of control over such factors.

Stimulus Predifferentiation

“The greater the relevancy of the initial ... responses to the later ... task, the greater the positive transfer expected” (Wittig 1981). Stimulus predifferentiation takes place when subjects are conditioned to provide a response to a given stimulus. The more relevant the training tasks appear to the transfer tasks, the greater the transfer can be expected to be. Familiarisation tasks and previous exposure will all increase stimulus predifferentiation.
**Task Difficulty**

Practising an easier task may sometimes facilitate better performance in a subsequent task than training on the task itself (Wittig 1981). This may seem counter-intuitive (given that transfer increases with the similarity of the two tasks).

### 2.3.1.4 General factors in transfer

General factors are independent of the nature of the initial task.

**Warm Up**

Warm-up results from practice and aids learning by allowing the subject to prepare themselves to attend to the stimuli or adjust to the rhythm of the task. It disappears within hours of the trials (Hulse, Egeth and Deese 1980). If the transfer task is temporally close to the training task, warm-up should be controlled using a “put in time” task (which warms the subject up using non-related stimuli) (Wittig 1981).

A related concept is *fatigue*, which is the “opposite of warm-up” (Wittig 1981). Too much practice is likely to make subjects unresponsive to learning opportunities. There is therefore a compromise to be made between sufficient practice (Section 2.3.1.3) and *fatigue*.

**Learning to Learn**

Learning to learn refers to the process where tasks become easier with practice (Ellis 1965). It is also the process when subjects learn general principles that can be applied to other situations (Wittig 1981). Learning acquired through practice is more permanent than temporary warm-up exercises (Hulse, Egeth and Deese 1980). Hulse et al. (1980) cited a study by Ward who showed that practice in learning (memorising) different lists of words allowed subjects to more easily learn (memorise) a specific task. From the presented data, subjects seemed to reach an asymptotic level of performance after around five previous repetitions (Hulse, Egeth and Deese 1980). Bech had found the subjects needed between four and eight repetitions to reach an asymptotic performance level during loudspeaker quality evaluations (see Section 2.2.1.3).

**Measuring Transfer**

In order to measure the amount of transfer that has occurred in an operation, Wittig suggests looking at either absolute or percentage transfer (Wittig 1981). In absolute transfer a performance measure (for example the number of errors in the task) are directly compared between the groups. In order to standardise these absolute figures, Ellis (1965) provides a number of *transfer formulae* that allow the calculation of a percentage of transfer, enabling positive and negative transfer to be quantified in a standard way (numerators in the equations are switched over if the performance measure is desirable as low as possible – e.g. errors):

Equation 1 Ellis (1965): compares absolute performance of experimental (E) and control (C) groups.

\[
\frac{E - C}{C} \times 100 = \text{PercentageTransfer}
\]

\[
\frac{C - E}{C} \times 100 = \text{PercentageTransfer}
\]
Equation 2 Ellis (1965): compares absolute performance between E, C and the total possible (T). It is potentially useful if you know the total possible grade achievable.

\[
\frac{E - C}{T - C} \times 100 = \text{PercentageTransfer} \quad \frac{C - E}{T - C} \times 100 = \text{PercentageTransfer}
\]

Equation 3 Ellis (1965): compares absolute performance of experimental (E) and control (C) groups. This equation always has a range of -100% to +100%.

\[
\frac{E - C}{E + C} \times 100 = \text{PercentageTransfer} \quad \frac{C - E}{E + C} \times 100 = \text{PercentageTransfer}
\]

Wittig (1981) also recommends testing over extended periods of time in order to “catch” any transfer effects that did not show up in the initial transfer tests.

### 2.3.1.5 Teaching for transfer

It is possible to teach for transfer by following certain guidelines – adapted from (Ellis 1965):

- Train and test for specific outcomes - devise the training and transfer tasks so the skills are practised in a realistic environment that is as similar as possible to the original setting.
- Analyse the important outcomes of the task and teach and test for those.
- Provide practice in a “real-world” environment, or final task environment.
- Allow extensive practice of the original task.
- Provide examples of concepts and non-concepts in order to demonstrate the applicability of the training.
- Draw attention to the most important features of the task.
- Explain general principles in order to facilitate for transfer.

### 2.3.1.6 Procedural and Declarative Objectives and Transfer

Clark and Voogel (1985) attempted to explain the many transfer failures that have occurred throughout the literature in terms of a confusion that exists between behaviourist and cognitive procedures.

From their perspective, near transfer is limited to specific skills that are not generalisable, and far transfer involves decontextualisation of skills so that they are widely applicable. They hypothesise that near transfer seems to be at the expense of far transfer and vice versa. They do not expect procedurally trained subjects to be able to easily generalise their skills, and they do not expect those that have generalisable skills to be able to easily use these practically.

They argue that by catering for near transfer one is potentially reducing the possibility for the subject to generalise their knowledge, and suggest a number of ways to aid this.
They distinguish between "procedural objectives" (which are useful for near transfer and specify objectives and procedures that need to be mastered), and "declarative objectives" which are more suitable for far transfer (objectives are written in a less rigid manner, allowing more room to experiment.).

Clark & Voogel also suggest using a variety of different contexts for the practice sessions, and the use of analogies as this will help decontextualise the specific skills from the specific stimuli.

They believe that, on the whole, behaviourist studies (which tend to foster near transfer in their opinion):

- direct and monitor progress
- provide feedback and reinforcement
- test after practice

Whereas cognitive model studies (which tend to foster far transfer in their opinion):

- encourage decontextualisation
- encourage discovery
- paraphrase
- use advance organisers
- use analogy
- test the generalisability of learning

They go on to argue that the use of advance organisers (explanatory tutorials) in more cognitive-based studies can increases far transfer further than in more behaviourist studies.

### 2.3.1.7 High and Low Road Transfer

Salomon & Perkins (1989) attempted to explain transfer in terms of two different phenomena, each being capable of producing flexible skills. They called this low road transfer and high road transfer.

Low road transfer describes the process by which subjects can learn practices in various situations such that their response becomes automatic. The mechanism for achieving low road transfer is to practice until responses become automatic, and to vary the practice so that new situations are encountered and assimilated by the learner.

The main issue with high road transfer is that it involves "mindful abstraction" (Salomon and Perkins 1989), the decontextualising of the task to allow prior knowledge to help to find a solution.

Unlike Clark & Voogel (Section 2.3.1.6), Salomon & Perkins (1989) believe that if one reflects upon and practices the behaviour, it is possible that both high and low roads of transfer can be utilised.
2.3.1.8 Ahissar's Reverse Hierarchy and Transfer

Ahissar (2001) commented on a paper presented by Wright & Fitzgerald (2001), and explained that their experiment showed that transfer for certain learned skills did not occur due to the hierarchical level at which the skills were learned.

The experiment involved the use of inter-aural time differences (ITD) and inter-aural level differences (ILD) to localise sounds on headphones (Ahissar and Hochstein 2002).

Ahissar argued that the learning for each condition must have occurred before the concept of auditory localisation was formed. This has important implications, because the current study is based around the use of higher level perceptual or cognitive concepts and not looking at low-level physical parameters. Learning taking place at these higher (fused) levels is expected to be more transferable.

Ahissar (2001) also agrees that variety of stimuli is useful in learning, explaining that "initial learning" begins to get the "gist" of the task and that this begins at "generalizing high-level sites". Initial learning appears to be very useful in the quest for optimal transfer of concepts.

2.3.1.9 Sternberg & Frensch's Four Mechanisms of Transfer


Encoding specificity refers to how learning needs to be encoded in the brain in such a way that it is possible to use it in other situations. This can be achieved by explicitly showing students how to apply information and require that they find their own applications for their skills.

Organisation refers to the observation that experts organise learning in a deeper structured level than novices. Organisation can be aided by ensuring that information to be learned is connected logically, either by the trainer or by the trainee.

Discrimination means that information is deemed either relevant or non-relevant for particular situations. If relevant areas are selected for the subject, this will allow them to aid students in choosing relevant objects.

Set relates to having the appropriate mind set required for transfer. Testing for application rather than recall will create a mind set ready to understand concepts rather than facts.

2.3.1.10 Transfer on Trial

Detterman (1993) conducted a literature survey of transfer studies and came to the opinion that transfer rarely happens, and when it does it is normally because the investigators have specifically explained to the subjects what is needed in order to facilitate transfer. It is worth noting though, that Detterman is almost certainly only interested in far transfer.

He gives the following advice:
- Use double-blind procedures, especially with investigations into general transfer.
- Provide a “filler” task for the control group.
- If subjects are told that something will be useful it should not be a surprise when they use it during a subsequent test. The peril is that they might use a trained method in an inappropriate manner.

This last point conflicts with Sternberg & Frensch’s advice to assist students as much as possible (see Section 2.3.1.9).

### 2.3.1.11 Transfer Summary

Transfer of learning has been classified in terms of whether it relates to tasks and/or situations which are identical (near transfer) and tasks and/or situations which are different (far transfer). Generalised learning with a variety of tasks promotes far transfer whereas task-specific learning promotes near transfer. Clark & Voogel (see Section 2.3.1.6) contend that near and far transfer cannot be catered for simultaneously, whereas Salomon & Perkins (see Section 2.3.1.7) believe that if tasks are practiced and reflected upon, both near and far transfer is possible. Ahissar (see Section 2.3.1.8) recommends that initial learning is best for promoting far transfer as further practice on specific tasks results in more specific learning that is less generalisable.

In order for the training system to show applicability, as wide a transfer as possible is sought.

Firstly, the training task needs to be similar enough to the transfer task to encourage transfer of training from one environment to the other. The tasks also need to have a wide enough variety to allow the subjects to decontextualise the stimuli, facilitating further transfer. Difficulty should be set very low to begin with.

Warm-up and practice effects should be controlled for the non-experimental group, in this case possibly by creating a “timbral” ear training package for them to use instead.

An ecologically-valid task needs to be provided for the transfer task, and transfer needs to be assessed by setting a specific and meaningful goal. This needs to be measured and assessed using Equation 3 to gauge transfer. The test needs to be geared for application, not recall.

Analogies should be used and reflection fostered in order to encourage generalisation of the skills.

Aid should be provided to the students as far as possible in order to assist encoding specificity. Familiarisation and regular exposure to the stimuli should provide stimulus predifferentiation.

### 2.3.2 Motivation

Motivation is an important factor in learning (Alessi and Trollip 2001). To better understand how to optimise learner motivation in any spatial audio training system, two theories of motivation were studied: Malone & Lepper’s Taxonomy of Intrinsic
Motivations for Learning is covered in Section 2.3.2.1, and Keller & Suzuki’s ARCS Motivation Model is detailed in Section 2.3.2.2.

### 2.3.2.1 Malone & Lepper’s Taxonomy of Intrinsic Motivations for Learning

After studying the motivating effects of computer games, Malone & Lepper (1987) attempted to summarise the factors affecting motivation levels for learners in their Taxonomy of Intrinsic Motivations for Learning.

They argue that **intrinsic** motivation – “learning that occurs in a situation in which the most narrowly defined activity from which the learning occurs would be done without any external reward or punishment” (Malone and Lepper 1987) – is more beneficial than **extrinsic** motivation (which relates to doing an activity to achieve an external goal).

They also make the distinction between **endogenous** and **exogenous** forms of motivation. **Endogenous** motivation having “motivating embellishments” (Malone and Lepper 1987) as part of the instructional content. **Exogenous** reward utilises motivating elements that are separated from the instructional content.

In other words a task could be **intrinsically** motivating if useful skills are gained by doing it, whereas a task could be **extrinsically** motivating if you are paid to do it. A task can also be **endogenously** rewarding by including an internal rewards structure, whereas **exogenously** rewarding tasks could reward participants for simply taking part a given number of times.

They hypothesise that **intrinsic** and **endogenous** motivations will provide higher levels of motivation and greater learning when harnessed within a learning environment, but postulate that any learning environment can benefit from the use of motivational techniques described in their taxonomy.

The taxonomy has four elements: Challenge, Curiosity, Control and Fantasy (CCCF).

**Challenge**

Malone & Lepper (1987) claim that “people prefer an optimal level of challenge”. They argue that there is little intrinsic value in activities that are either too difficult or too easy for the learner.

In order to optimise **challenge** in a learning environment, Malone & Lepper suggest the setting of **explicit goals**, the use of **uncertain outcomes**, provision of **performance feedback**, and enhancement of the user’s **self-esteem**.

Goal setting can either be done by the instructor (through explicit provision), or in open-ended environments could be set by the user themselves. If learners set their own goals care must be taken to not allow them to set unrealistic goals as this can demotivate them.

In order for the learner to feel challenged by the environment, they should not feel assured of success. Malone & Lepper suggest that motivation is highest when the probability of success in a task is exactly 50%. They suggest varying difficulty levels (either determined by the learner’s performance or chosen by the learner themselves), and including multiple levels of goals (either where goals are made
more difficult, or by using different goals such as speed and accuracy). Including random elements into the instruction and omitting information are also ways to enhance the challenge of the environment.

Feedback allows learners to appraise how they are doing in relation to the goals of the environment. Performance feedback should be frequent, clear and constructive (providing corrective information) and encouraging.

Malone & Lepper (1987) also recommend that the user’s self-esteem should be promoted because “success makes people feel good about themselves, failure can make people feel worse about themselves”. They suggest a number of ways to promote self-esteem. Feedback should promote the feeling that subjects are achieving success. There should be a series of increasingly difficult levels within an activity, allowing all learners to progress. Finally, performance goals should be made meaningful to the individual learners by demonstrating relevance.

Curiosity
For curiosity, as with challenge, an optimal level should be sought.

They delineate two forms of curiosity: sensory curiosity and cognitive curiosity. Sensory curiosity refers to the use of attention-drawing affects using light and sound. Cognitive curiosity is utilised by stimulating the need to modify one's current understanding. The use of incomplete or inconsistent information is expected to stimulate the learner to seek the full information and resolve the inconsistency.

Control
Providing the learner with control (or the perception of control) over their own destiny within a learning environment is a powerful motivational tool. Control can be provided through contingency and choice. As the learner develops, so the learning environment can reduce the assistance provided by feedback and vary the tasks posed. The learner can also be provided with the ability to personalise their environment, being allowed to change the type of learning task, format of instruction, fantasies evoked by the environment and use and choice of audio/visual effects.

Fantasy
Malone & Lepper (1987) define a fantasy environment as “one that evokes mental images or physical or social situations not actually present”. They go on to specify two forms, exogenous fantasy (where the fantasy is not directly linked with the skill to be learned) and endogenous fantasy (where the fantasy and skill depend on one another). Because different people are motivated by different fantasies, the provision of a number of fantasy choices can allow the widest audience to be motivated. Malone & Lepper hypothesise that the use of particularly endogenous fantasy can provide useful metaphors and evokes vivid images that aid learning and recall.

Interpersonal Motivations
In addition to the taxonomy, Malone & Lepper also discuss factors relating to the interaction of different learners with one another. Competition can be a strong motivational factor, but can also be demotivating. Opportunities can be provided depending on whether learners are stimulated by competition or affiliation. Another factor affecting motivation is recognition. Learners will be motivated to perform
well in a task if their achievements are visible to others. This could be a list of prize
winners. Such a list should not include all learners, as this would undoubtedly
demotivate the learners who were not doing as well within the group.

**Summary of Malone & Lepper’s Taxonomy**

Malone & Lepper’s taxonomy describes a number of motivational factors that can be
considered when designing learning environments. Optimal levels of challenge and
curiosity should be harnessed to attract and maintain attention. Control should be
offered to the learner in order to allow them to personalise the learning experience
and to concentrate on the elements that are relevant to them.

**2.3.2.2 Keller & Suzuki’s ARCS Motivation Model**

Keller and Suzuki (1988) describe the use of their ARCS (Attention, Relevance,
Confidence, Satisfaction) motivational model in computer-based instruction,
summarised below.

**Attention**

The first factor in the ARCS model is attention. Keller & Suzuki (1988) explain that
in order for the user to learn, their attention should be attracted. They outline three
strategies for gaining attention: perceptual arousal, inquiry arousal and variability.
Perceptual arousal is the use of surprising or unexpected events. Inquiry arousal
refers to the fostering of curiosity within the learner, for example using mystery to
make the learner create questions to solve for themselves. Variability is the
infrequent change of parts of the programme. In contrast to perceptual arousal
(which is deliberately “catchy” and designed to gain attention), variability is used to
maintain interest by reducing monotony. Keller & Suzuki warn against the use of
too much variability and recommend the use of a degree of continuity which will
increase the familiarity with which the learner views the instruction (see relevance,
below).

**Relevance**

Keller & Suzuki (1988) argue that motivation is engaged by demonstrating personal
relevance to the learner once their attention has been aroused.

They continue to explain that relevance is both an “ends-related aspect” (if the
content can be shown to be useful in the future, or perhaps could have been useful in
the past or present, then perceived relevance will increase), and a “process aspect”
related to satisfaction – students may be motivated by collaborative or competitive
aspects of the instruction). Keller & Suzuki recommend that the course content
should be written in an enthusiastic style which could engender a similar attitude in
the learner.

Keller & Suzuki believe that familiarity is a useful strategy for increasing relevance.
Relating the instruction to the learner’s experience (by using anecdotes for example)
fosters an affinity with the programme. They recommend the retrieval and use of the
learner’s first name in interactions as this strategy mostly enhances familiarity.
Elements that add continuity to the instruction also increase the learner’s affinity
with the course. Another strategy is goal orientation. A clear statement of the
purpose or importance of the lesson with a statement of the learning objective helps
to generate relevance in the perception of the learner, especially if their own motives have been taken into consideration. The learner could also be required to generate their own goals, or to complete certain stages of the instruction before being allowed to participate in more advanced (or rewarding) stages.

Confidence

Keller & Suzuki (1988) hypothesise that there must be an “acceptable probability of success” in order for learners to feel motivated. They argue that challenge must be within reasonable limits.

They divide confidence into three dimensions: perceived competence, perceived control and expectancy for success. Learners need to feel that they have the competence to take part in the task. If they feel that they have control over the consequences of their actions, and feel that success in the task is obtainable, learners will feel more confident and even try harder.

Keller & Suzuki provide three distinct strategies for improving confidence: the statement of learning requirements, provision of success opportunities and provision of personal control.

If the learner is fully aware of how they will be evaluated, they will have the confidence that there will be no mystery objectives and be allowed to estimate the likelihood of their success.

Distinct from Malone & Lepper’s theories on optimal challenge levels (see Section 2.3.2.1), Keller & Suzuki believe that subjects learning new skills need a fairly large initial opportunity for success. They argue for the provision of multiple entry points into the course, allowing learners with better initial competence to begin the course at a higher level. They also suggest that learning tasks should start with a high chance of success and get gradually more and more difficult. Keller & Suzuki recommend that challenge should be used to increase the motivation of competitively motivated students. Difficulty levels in the course should be altered by changing the evaluation parameters to achieve a “personally meaningful level of challenge and develop both confidence and self-esteem” (Keller and Suzuki 1988).

Control also features in the ARCS motivation model – Keller & Suzuki (1988) recommend that the learner be given an optimum level of control in order to build “internal attributions for success”. The learner should be given control over potentially irritating issues such as pacing (allowed to skip or dwell on certain elements of the instruction), have the ability to end any lesson at any time and have immediate access to the instruction without having to watch long introduction sequences. In addition, access to different parts of the course and the ability to change the difficulty level should be given to the learner. Keller & Suzuki also recommend using language that personalises effort and subsequent success to help learners attribute the success to their own achievements. In order to minimise frustration, it is also important to terminate any task that is doomed to failure and encourage the student to begin again (or presumably begin another easier task to build confidence).
Satisfaction
Satisfaction mainly affects motivation to continue with the learning programme. If the learner expects more or less from the programme than it delivers, motivation to remain on the programme will wane. Keller & Suzuki explain that predictable rewards encourage a feeling of security, and that these rewards should be *intrinsic*, rather than external. They also describe the process whereby an initially negative view of a learner’s own performance can be altered upon reflection with appropriately chosen positive feedback.

Satisfaction can be enhanced by three specific strategies: *natural consequences*, *positive consequences* and *equity*. Natural consequences include allowing the learner to use the learned skill in a specific setting – real, fantastical or simulated. Positive consequences involve providing positive motivational (rather than strictly corrective) feedback being used to help sustain the desired behaviour. Because the novelty can wear off, however, Keller & Suzuki recommend that such reward systems be a user-selectable option that can be turned off (or perhaps changed) in the future. Equity refers to the maintenance of “standards and consequences” (Keller and Suzuki 1988). Good communication and a fair enforcement of standards will prevent the learner becoming demotivated through a perception of injustice in the way in which they are being evaluated.

Methods of testing for motivation
Keller & Suzuki recommend that motivation levels should be measured using either *affective reaction* (obtaining feedback from the learner through questionnaires) or *achievement* (by examining performance metrics). Rather than enquiring about motivation in general, Keller & Suzuki suggest first determining the motivational effect that is required, and creating an objective for that. An example could be that if increased satisfaction is sought, a questionnaire could be devised that aims to measure the perceived fairness of the evaluation, or metrics could be acquired that relate to the continued motivation of the learner.

Summary of Keller & Suzuki’s ARCS Motivation Model
Keller & Suzuki recommend that knowledge of the intended audience is important for designing learning environments, especially their need for challenge or confidence. Another useful general point that they mention is that it is advisable to reduce irritating factors (such as long introductions that cannot be skipped, or delays in loading of elements from disk) which will not *improve* motivation if done correctly, but will *reduce* motivation if done badly. Like Malone & Lepper (see Section 2.3.2.1), they also recommend that their model be used to inform the creation of learning environments and advise against using the model in its entirety.

2.3.2.3 Summary of Motivation
Motivation is seen as an important factor in successful learning. Training is more effective if people are motivated to do it. *Intrinsic* motivation is better than *extrinsic* motivation. Two theories of motivation in learning have been studied, and a number of common threads have been found, but also some important differences. Table 14 shows a classification of the similarities and differences between Malone & Lepper’s Taxonomy and Keller & Suzuki’s ARCS model.
**Table 14: Similarities and differences between the Taxonomy and the ARCS Model.**

<table>
<thead>
<tr>
<th></th>
<th>Malone &amp; Lepper</th>
<th>Keller &amp; Suzuki</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Challenge</strong></td>
<td>Explicit goals</td>
<td>Relevance *</td>
</tr>
<tr>
<td>(optimal level)</td>
<td></td>
<td>Goal orientation *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Learning requirements</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equity</td>
</tr>
<tr>
<td><strong>Uncertain outcomes</strong></td>
<td></td>
<td><strong>Success opportunities</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relevance *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goal orientation *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceived competence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cognitive evaluation</td>
</tr>
<tr>
<td><strong>Performance feedback</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reinforcement/Feedback</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intrinsic rewards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Positive consequences</td>
</tr>
<tr>
<td><strong>Self-esteem</strong></td>
<td></td>
<td>Relevance *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goal orientation *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Perceived competence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cognitive evaluation</td>
</tr>
<tr>
<td><strong>Curiosity</strong></td>
<td>Sensory curiosity</td>
<td>Attention</td>
</tr>
<tr>
<td>(optimal level)</td>
<td></td>
<td>Perceptual arousal</td>
</tr>
<tr>
<td></td>
<td>Cognitive curiosity</td>
<td>Attention</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inquiry arousal</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>Making level more difficult adaptively</td>
<td>Relevance *</td>
</tr>
<tr>
<td>(perception of control)</td>
<td></td>
<td>Goal orientation *</td>
</tr>
<tr>
<td></td>
<td>Personalise the environment (task, fantasy, feedback)</td>
<td>Confidence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Personal control</td>
</tr>
<tr>
<td><strong>Fantasy</strong></td>
<td>Metaphors / vivid images</td>
<td>Relevance *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goal orientation *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Satisfaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Natural consequences</td>
</tr>
<tr>
<td><strong>Attention</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variability</td>
</tr>
<tr>
<td><strong>Relevance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Familiarity</td>
</tr>
</tbody>
</table>

* Denotes that the item appears in more than one group.

Many issues feature in both paradigms, for example Keller & Suzuki’s perceptual and inquiry arousal (part of attention) are equivalent to Malone & Lepper’s sensory and cognitive curiosity. Regarding the differences between the two paradigms, Keller & Suzuki note that their idea of fantasy differs from that of Malone & Lepper in that they classify fantasy as being over-and-above what would happen in reality and therefore, according to them potentially very much more relevant. Other differences visible in Table 14 are the apparent conflict between Malone & Lepper’s call for uncertain outcomes and Keller & Suzuki’s recommendation for success opportunities. Keller & Suzuki explain that the need for success opportunities is more important during the learning of new skills, and that uncertain outcomes and challenge is more important when practising those skills. Apparently missing from the CCCF Taxonomy is an equivalent notion for familiarity (part of establishing relevance) and the use of variability to encourage attention. It is possible that variability could have been considered tacitly within control and that familiarity could be fostered through the use of consistent fantasies.
Overall there are some strong recommendations:

**Goals:** Ensuring that the goals of the training environment are well understood by the learner, or having the learner set and modify their own goals is an important motivational factor. Having specific goals satisfies the need for goal orientation (*relevance*), learning requirement (*confidence*) and helps to promote the perception of equity (*satisfaction*).

**Feedback:** Continuous and constructive feedback is deemed necessary to inspire motivation, not only to help with goal orientation (*relevance*) and enhance perceived competence (*confidence*), but also to supply the main strategies for engendering *satisfaction* (reinforcement/feedback, intrinsic rewards and positive consequences). Feedback should always be encouraging and should be corrective rather than negative. Having a user-controllable feedback system would allow this to be tailored to the individual’s requirements and reduce the possibility of irritation.

**Attention:** The need to obtain and maintain attention appears to be very important. Without initially getting the attention of the learner, they will not have a chance to benefit from the instruction. Without being able to maintain that attention, motivation will wane and the programme will lose its audience. Judicious use of attention-grabbing effects and occasional variation in the instruction could be coupled with user-control over variability and audio/visual effects to maximise the motivational potential of the programme.

**Personal Control:** Because the perception of control is such a powerful motivating factor, the user should be allowed to personalise elements such as type and amount of feedback, instructional content, presentation style, goals and difficulty. Allowing personalisation of elements of the instruction will also aid the designer in achieving an optimal balance of factors such as *challenge, goals* and the type and amount of *feedback*, due to the ability for the learner to adjust these factors to be personally appropriate. See also (Laurillard 1987).

In order to gauge motivation levels, objectives that measure important elements of the motivational outcomes should be considered and measured, either using pre- and post-programme questionnaires or by recording appropriate performance measures (or both).

**Practice Opportunities:** Provision of the ability for learners to practice what they have learned in real, simulated or fantastical settings allows for a number of motivational factors to be addressed. Practice can provide *challenge* as well as *success opportunities* (to inspire *confidence*). An optimal level for each of these can be provided by increasing the difficulty of the practice gradually, or providing specific practice sessions to inspire confidence or create challenging situations. It can also help to achieve *satisfaction* in the learner through the demonstration of natural consequences.

**Testing for motivation:** In addition, measures of motivation levels can be obtained through the use of performance metrics or questions that pertain to one or more of the specific motivational factors specified in either of the paradigms.
2.4 Overall Summary

This chapter has covered four main areas of the literature: spatial audio attributes, listener training, transfer of learning and learner motivation.

Section 2.1 and its subsections covered previous studies that involved spatial audio attributes. In Section 2.1.9 a case was made for the use of a provided attributes. Previous studies were summarised and Rumsey's Scene-Based Paradigm was found to be the most rigorous method for describing reproduced spatial audio scenes. See Section 2.1.9 for a further summary of Section 2.1.

Previous work in listener training was discussed in Section 2.2 and its subsections. Much of the research described mostly-annecdotally successful listener training systems. Other research investigated the effects of repetitive practice. Important outcomes of the study of the literature outlined in Section 2.2 included the existence of repetitive practice as a method for training listeners before participation in listening tests. The need for a familiarisation phase, pre/post test methods and the usefulness of matched control groups were all demonstrated. See Subsection 2.2.4 for a full summary of Section 2.2.

Section 2.3.1 and its subsections dealt with the important issue of transfer. Transfer has been classified as near and far. Near transfer relates to situations and tasks identical or close to the trained task and/or situation. Far transfer related to tasks and situations that are increasingly difficult from the trained ones. An important outcome from the survey was that generalised training promotes far transfer, whereas task-specific training promotes near transfer. See Subsection 2.3.1.11 for further summary of Section 2.3.1.

Motivation in learning was highlighted by Zwislocki et al. (see Section 2.2.1.5), and previous research in this field was investigated in Section 2.3.2 and its subsections. The two main theories of motivation in learning (CCCF and ARCS) were compared and discussed. Important outcomes of the survey were the need for user control, goals, feedback and attention. It is important also to note that extrinsic forms of motivation (such as the payment of participants in previous studies – see Sections 2.2.3.3 and 2.2.3.4) was likely to be less effective than employing intrinsically motivated subjects (those who understood the benefit of the training itself). See subsection 2.3.2.3 for a full summary of Section 2.3.2.

The following chapter details the specification and creation of a paradigm for the description of reproduced spatial audio scenes that is optimised for training purposes.
3 THE SIMPLIFIED SCENE-BASED PARADIGM (SSBP)

Previous studies involving spatial audio attributes have been investigated and summarised in section 2.1. However the only rigorous set of attributes flexible enough to describe various spatial scenes is Rumsey's Scene-Based Paradigm (discussed in Section 2.1.8).

In order to train listeners in the identification and specification of spatial audio attributes, a framework within which to base the training is required. Neher (2004) specified that the eventual goal of a spatial training system would be to simulate every attribute within Rumsey's Scene-Based Paradigm. However, as will be discussed in Section 3.1, this is not the ideal paradigm to use.

This chapter identifies criteria for the inclusion of spatial attributes into a paradigm that is suitable for training. Attributes from previously published studies are examined and included or rejected as part of an overall paradigm for use in spatial audio training. Sections 3.2.3 and 3.2.4 detail a new, Simplified Scene-Based Paradigm (SSBP) for use in spatial audio training.

Some of the content of this chapter has been published previously in (Kassier, Brookes and Rumsey 2004).

3.1 Criteria for the inclusion of spatial audio attributes in a training paradigm

Berg states that three conditions for the modification and discarding of attributes were applied in his studies (see Section 2.1.7.4) and lists them as:

• inapplicability to the context of spatial audio
• inapplicability to linear scales
• low listener consistency in rating

Additionally, work carried out in the field of food sciences, brought to the attention of the audio research community in (Bech 1999), features important discussions about the selection of attributes for descriptive analysis. (Lawless and Heymann 1999) list the desirable characteristics for terms to be used in descriptive analysis experiments (in their approximate order of importance) as:

• Discriminate
• Nonredundant
• Relate to consumer acceptance/rejection
• Relate to instrumental or physical measurements
• Singular
• Precise and reliable
• Consensus on meaning
• Unambiguous
The Simplified Scene-Based Paradigm

- Reference easy to obtain
- Communicate
- Relate to reality

Whilst the senses of taste and hearing may not be directly comparable, studies involving these senses share similar problems relating to the fact that tangible references are difficult to isolate and agree upon. The requirements for descriptive analysis in sensory analysis of food are therefore similar to those for the evaluation of sound.

The above criteria will be discussed individually in light of their applicability to the task of the creation of a paradigm of spatial audio attributes for a training system in the following subsections.

3.1.1 Discriminate

Lawless and Heymann (1999) suggest that subjects should be able to use scales of each of the terms to distinguish between the samples. Simply put, if there are no changes within the specific attribute scale (in the stimulus set under test) it should not be included.

Regarding the proposed training system itself, perceptually unidimensionally varying auditory description stimulus sets are expected to be created for each attribute in the training system. Each attribute should therefore be usable to discriminate between the stimuli in its auditory description training set (which by definition will exemplify a unidimensional change in the perceptual attribute that it is describing). In this way, the ability for subjects to discriminate between the stimuli using the attribute in question is more a criterion for the auditory description stimulus sets than the attributes themselves. It is also perhaps useful to point out here that perceptual attributes arising from sensations that are not possible to simulate using the experimental system (such as the perception of “height” using 3/2 Stereo reproduction) will not be usable to discriminate between potential stimuli and should hence be excluded from the current training system. Should the training system be extended to cover different replay formats, for example, then attributes that allow for the description and discrimination of the perceptions arising from such changes should be incorporated.

Regarding the wider applicability of the proposed training system, the attributes chosen should usefully discriminate between as wide a variety of potential experimental stimuli as possible. This will maximise its benefit as trained subjects should be capable of applying their training to as large a variety of tasks as possible. Whilst the ability of the selected attributes to discriminate between stimuli will be dependent on the programme items and experimental techniques used, verification of this ability could form a part of the test for the training system.

3.1.2 Nonredundant

Lawless and Heymann (1999) suggest that attributes should be put together which “have little or no overlap with other terms used”. They also describe terms that are not correlated with each other as “orthogonal”. The justification given for this criterion is that “it is very confusing, demotivating, and mentally frustrating to the panellists when they are asked to score redundant terms”.

- 87 -
Regarding a paradigm of spatial attributes for use in training, this means that the terms used should be carefully chosen not to overlap conceptually. If individual spatial audio perceptions are found to be describable using more than one attribute, the most appropriate one should be included. Lawless and Heymann’s argument also provides further evidence that the training system should not rely solely on evidence from experimentation (which will have unique, possibly freak, set-up conditions), but also that logic needs to be applied when combining the terms within the training paradigm.

Lawless and Heymann (1999) also go on to say that “panelists often have preconceived notions about which terms are correlated and which are not. During training it is often necessary to help panelists ‘decorrelate’ terms.... Exposing panellists to [exemplary reference] products would help decorrelate [perceptually covarying] terms, allowing the panel to understand that the two terms do not always have to vary together”.

This argument lies at the heart of the need for perceptually unidimensional reference stimuli. Through the use of such stimuli it will be possible to exemplify how each of the chosen attributes can vary independently of the other attributes in the scheme.

### 3.1.3 Relate to consumer acceptance / rejection

Lawless and Heymann (1999) deem this criterion to be necessary because “data from descriptive analyses are often used to analyse consumer hedonic responses to the same samples”. They go on to explain that it would be helpful if the selected terms could be “related to concepts that lead consumers to accept or reject the product”.

Whilst this may be useful in listener training, specifically for the testing of products, the focus of the proposed training scheme is to train subjects to be aware of spatial attributes (as opposed to timbral attributes) of audio in general and not necessarily to train them to describe or rate consumer preference-related perceptions. Whether current understanding of how spatial perceptions map to consumer preference is sufficient to be able to screen attributes for their applicability is debatable, and something the proposed training system may eventually aid.

### 3.1.4 Relate to instrumental or physical measurements

Lawless & Heymann (1999) explain that “ideal descriptors can be related to the underlying natural structure (if this is known) for the product”. The examples given by Lawless & Heymann regard the use of descriptors for one food which are based around the chemical compounds of another - such as using “bell pepper” to describe wine, or “butyric acid” to describe “a specific odor in aged cheese”.

Regarding measurement, Nunally and Bernstein (1994) explain that “one cannot measure objects, one measures their attributes”. Subjects do not tend to cognitively perceive attributes such as source distance in terms of the individual physical changes in the signals arriving at the ears – physical variables” as defined in (Bech 1999) – but rather the perception of the source being at a certain distance or changing its perceived distance from the listener – an “individual impression” as defined in
(Bech 1999). It is therefore right to relate the terms used to “individual impressions” rather than to go all the way down to “physical variables”, using the terminology found in (Bech 1999).

3.1.5 Singular

Lawless & Heymann (1999) note that “descriptors should be singular rather than combinations of several terms. Combination or holistic terms ... are very confusing to panellists. These terms should be broken down into their elemental, analytical, and primary parts”.

In terms of the spatial audio attribute paradigm, the singular criterion is open to interpretation, but can be thought of as equating to a single perceptual dimension. Hence, if a spatial attribute can also be described by using a number of other constituent spatial attributes, it is better to use these than the combined term. The constituent attributes will also need to satisfy the criteria for inclusion themselves (for example, they must not also overlap conceptually with one another).

A balance must be found between such constraints as singularity and what is actually perceivable as an attribute by subjects. Previously studied spatial audio attributes are an obvious starting point, as these are likely to have been perceived as spatial audio attributes. Each included attribute will also need to fulfil the other criteria.

3.1.6 Precise and reliable

Lawless & Heymann (1999) suggest that “suitable descriptors are ones that can be used with precision and reliability by the panellists”.

This criterion will not only affect which attributes are selected for inclusion in the paradigm, but can also provide a means to evaluate the paradigm, as the selected attributes could be checked for the reliability and precision in which they are used by the subjects once they have been trained to do so.

3.1.7 Consensus on meaning / unambiguous

Lawless & Heymann (1999) explain that “panelists should fairly easily agree on the meaning of a specified term, the term should thus be unambiguous.... They should be able to agree on the prototypical examples related to the descriptor, and they should agree on the boundaries of the descriptor. Using reference standards to signify these boundaries is encouraged”.

There are two ramifications of this criterion in terms of the proposed training system.

Firstly, attributes should be avoided that have been shown to be divisive, or have been used by different experimenters in different ways (Shaw and Gaines 1989). A holistic or amalgamated (i.e. non-singular) term could also be described as ambiguous, as there could be many conflicting subjective views as to what constitutes the attribute.

Secondly, it could be argued that if a term is unambiguous, there is no need to include it in a training system. However, a number of factors mean that we cannot assume that terms that are considered unambiguous to current experts in the field of spatial audio are unambiguous to subjects that have little or no previous formal
spatial hearing experience. Such factors are language translation issues (Teunissen 1996; Martens and Giragama 2002), issues of terminology usage (Shaw and Gaines 1989) and the potential problems of concept misalignment (O'Mahony 1991). Lawless & Heymann (1999) recommend the use of reference standards to help define the terms, and the proposed training system will aim to provide a consensus of meaning for the terms used through the use of auditory description stimulus sets.

The combined effect is that the attributes in the paradigm will need to be selected to be as unambiguous as possible from the outset, and through training will be exemplified in an unambiguous manner to the subjects to facilitate consensus (Shaw and Gaines 1989) and concept alignment (O'Mahony 1991).

### 3.1.8 Reference easy to obtain

Lawless & Heymann (1999) argue that “it simplifies the life of the panel leader if the physical reference standards for the descriptor are easy to obtain. However, difficulties in obtaining physical reference standards should not prevent the panel leader or the panelists from using terms that are ideal in every other way”.

A goal of the proposed training system is to exemplify spatial audio attributes using auditory description stimulus sets, and these will be sought for all attributes concerned. If unidimensional auditory description stimulus sets are not feasible, then either new attributes should be defined using achievable references, or the burden of perceptual unidimensionality (Neher 2004) may need to be eased so that the attribute could be defined by indicative references (as in the Zacharov & Koivuniemi studies – see Section 2.1.6). The focus of this chapter is the creation of a theoretical paradigm of spatial attributes based on previous studies, constructed in the most appropriate way for training purposes.

### 3.1.9 Communicate

Lawless & Heymann (1999) recommend that “terms should be understandable to the users of the information obtained in the study, not only to the descriptive panel and their leader”.

This relates to the ability for others not directly involved in the test to understand the attributes. The creation of auditory description stimulus sets that are commonly available should render the attributes widely communicable. This could also be part of the validation process for the training system.

### 3.1.10 Relate to reality

Lawless & Heymann (1999) explain that “it helps if the term has been used traditionally with the product or if it can be related to the existing literature”.

The terms in the proposed paradigm are intended to be drawn from previous studies. They should therefore be related to existing literature and previously (traditionally) used terms.
3.1.11 Summary of criteria

Regarding the selection of terms for the paradigm, Lawless & Heymann provide many useful pointers. Terms should be considered that were used in previous studies (i.e. relating to reality) in a reliable and consistent fashion (which Berg concurs with—see Section 2.1.7.4), but a systematic process of analysis and definition should be used in the selection of these terms: divisive or ambiguous attributes should be avoided. Attributes selected should also be singular (i.e. not holistic or amalgamated), yet non-redundant (i.e. have no conceptual overlap with one another, being notionally “orthogonal”). Thought should also be given as to how potentially useful the attributes will be in discriminating between as wide a range of eventual experimental stimuli as possible.

The following criteria must therefore be taken into consideration during the selection of attributes for inclusion in the training paradigm:

- Attributes should be drawn from previous studies in order to be seen to be “relating to reality” (Lawless and Heymann 1999). They should therefore also be “applicable to the context of spatial audio” (Berg 2002).
- Attributes should be continuously variable along a single perceptual dimension, and thus applicable to linear scaling (Berg 2002), rather than being holistic or amalgamated (Lawless and Heymann 1999).
- Attributes should also have no conceptual overlap with one another, being notionally “orthogonal” and non-redundant with respect to other terms (Lawless and Heymann 1999).
- Divisive or ambiguous attributes should be avoided, with attributes that have been shown to have been used reliably and consistently by subjects preferred (Berg 2002).

Once attributes has been selected that meet these criteria, the creation of perceptually unidimensionally varying auditory description stimulus sets for each attribute selected will allow other criteria taken from (Lawless and Heymann 1999) to be met. The use of auditory description stimulus sets containing stimuli that are discernible by changes in a single perceptual attribute will facilitate the generation of a consensus of meaning of the attributes amongst the subjects. They may also help to decorrelate non-redundant terms that are perceived by subjects to be correlated. Widespread access to the reference auditory description stimuli should render the attributes widely communicable.

Whilst the proposed training system is intended to help with the understanding of how various spatial audio attributes affect consumer preference in the future, not enough is known currently to make consumer preference the basis of any judgements as to the suitability of attributes to be included in the paradigm. The difficulty of obtaining reference stimuli is not going to be considered in the design of the training paradigm, but difficulties in the creation of reference stimulus sets may eventually require modifications in the paradigm to be made or a relaxation in the requirement for unidimensional variation of the auditory description stimuli.
The Simplified Scene-Based Paradigm

3.2 Constructing a paradigm for spatial ear training

Criteria for the construction of a paradigm for the purpose of spatial audio training have been identified in Section 3.1. The following sections detail the inclusion and rejection of spatial audio attributes found in previous studies in keeping with these criteria and an overall paradigm for use in spatial audio training is described.

3.2.1 Previous studies into spatial audio

Previous work involving spatial perception can be found in research literature on the subject of concert hall acoustics – for example see Beranek (1996) and Morimoto (2001), and sound localisation – see Blauert (1997) and Moore (1997). Localisation studies have tended to concentrate on the methods used by humans to determine the direction of lateral plane sources to the orientation of the head. Concert hall research has tended to concentrate on perceived spatial impression generated within live performance venues, which is subdivided into apparent source width (ASW) (Beranek 1996) – sometimes referred to as auditory source width (Morimoto 2001) – and listener envelopment (LEV) (Beranek 1996). As pointed out by Berg (2002) and Rumsey (2002), spatial audio reproduction is capable of the presentation of soundfields not naturally found in concert halls, so although elements of concert hall acoustics research can be seen to apply, other attributes are necessary in order to describe sound fields encountered in reproduced sound.

A number of previous studies into the quality of reproduced sound have involved various spatial aspects of sound. Examples include an early experiment by Eisler into the applicability of factor analysis in subjective audio tests (see Section 2.1.1); a ‘quad-era’ study by Nakayama et al. that involved multichannel sound recording and reproduction (see Section 2.1.2); various studies by Gabrielsson and his co-workers (see Section 2.1.3) into transducer sound quality; and a two year study of mono and stereo loudspeaker reproduction conducted by Toole (see Section 2.1.4). Studies specifically aimed at the identification of spatial audio attributes were carried out recently by Zacharov & Koivuniemi in the Finnish language (see Section 2.1.6), and by Berg & Rumsey in the Swedish language (see Section 2.1.7). Two theoretical methods of describing spatial audio attributes have also been published by Letowski (see Section 2.1.5) and Rumsey (see Section 2.1.8).

Neher (2004) proposed that perceptually unidimensional simulations of every attribute contained within Rumsey’s Scene-Based Paradigm (see Section 2.1.8) should be created for use in listener training, because it “is the only systematic approach to profiling spatial quality” (Neher 2004). However, according to the criteria previously delineated, it is possible to show that inclusion of all of the attributes from the Scene-Based Paradigm within a training system is not the most appropriate way to construct the training system. Rumsey also points out that “attributes should be chosen for an evaluation based on the task and context in question” (Rumsey 2002), and that not every attribute in the paradigm is intended to be included in every experiment.

In the following sections, the Scene-Based Paradigm is discussed in light of the previously delineated criteria. Where attributes are found to conflict with the
suggested criteria, modifications are proposed and a new paradigm that satisfies the criteria is presented.

3.2.2 The Scene-Based Paradigm revisited

In his Scene-Based Paradigm (see Section 2.1.8), Rumsey proposed a new method to decompose and describe spatial audio scenes in terms of a hierarchical system of attributes classified into three groups: dimensional attributes, immersion attributes and miscellaneous attributes. The Scene-Based Paradigm is a reasoned combination of spatial audio attributes that were previously either provided by experimenters in subjective tests or elicited from subjects, into a structured method that allows a multitude of different scenes to be described using rigorously defined terms. See Section 2.1.8 for further details, but the main attributes of the Scene-Based Paradigm will be outlined here, along with a discussion of their suitability for inclusion into a paradigm suitable for training.

At the heart of the Scene-Based Paradigm is the concept of auditory scene analysis (Bregman 1990) (the perceptual segregation of concurrent sound events into auditory streams perceived to be from the same sound-producing object) and the distinction between source-related and environment-related perceptions evident in the definitions of ASW and LEV by (Beranek 1996), and seen later in the separation into general, source and room attributes by Berg & Rumsey (see Section 2.1.7.2).

Implicit within the Scene-Based Paradigm (although not discussed in any detail due to the amount of work that exists on the subject), is the concept of lateral location of ensembles or the positioning of sources within the scene. This could be considered a source-related measure of source or ensemble direction.

Using the paradigm, a given spatial audio reproduction can be described using dimensional attributes in terms of a scene (of certain width and depth), containing one or more environments (each of a certain width and depth), each containing one or more ensembles (each of a certain width and depth and at a certain distance from the listener), each containing more than one source (each of which have a certain width and depth and are at a certain distance from the listener). Whereas source and ensemble attributes are both source-related, the dimensional attributes of environment width and environment depth are environment-related. The scene, however, contains all environments, ensembles and sources, and is therefore both source- and environment-related.

Using immersion attributes, the listener is able to describe the extent to which they are enveloped by individual sources, ensembles and the environment. Individual source envelopment and ensemble source envelopment is environment-related. There is also an additional immersion attribute termed presence which Rumsey defines as “the sense of being inside an (enclosed) space or scene” (Rumsey 2002). It is unclear if presence is classified as source- or environment-related – from the definition it appears to be environment-related, but source-related issues could affect it (presence will be discussed in more detail below).

There are also a number of miscellaneous attributes suggested in Rumsey’s Scene-Based Paradigm. These mainly refer to differences between the scene in question and a given reference. There are two source-related attributes: source stability and
source focus, and four additional scene attributes: scene left-right skew, scene front-back skew, scene stability and scene width homogeneity.

It is perhaps the need for non-overlapping, notionally “orthogonal” terms that conflicts most noticeably with the attributes within the Scene-Based Paradigm, when taken together. The hierarchical nature of the Scene-Based Paradigm allows for detailed description of scenes in terms of individual sources, ensembles and environments within the scene, but means that changes to attributes at various levels in the hierarchy will affect attributes on other levels of the paradigm and can therefore be said to overlap. Examples of this could be that changing the depth of an ensemble would necessitate the changing of the distances of various sources. In practical terms, a change in the depth of an ensemble could be confusing to subjects who are also being trained to listen for the distance of individual sources. Examples can also be found for overlapping concepts in ensemble width (which overlaps with individual source width and the position of individual sources) and ensemble distance (which would overlap with the individual source distances of the sources within the ensemble. In contrast, environment-related dimensional attributes (environment width and environment depth) do not overlap conceptually with source-related dimensional attributes (individual source width, depth and distance, or ensemble width, depth and distance). This is due to the distinction between sound that is fused perceptually with the direct sound, and that which is not (Morimoto 2001). The dimensional scene attributes (scene width and scene depth) can, however, be seen to overlap with both environment-related and source-related dimensional attributes. This is because changes in scene width or depth would not only accompany changes in the width or depth of the environment or environments contained within the scene, but also potential changes in sources that exist outside the environment(s), as expected in the paradigm. The concept of the scene as a whole is therefore a holistic term (in relation to the source/environment distinction), and therefore not suitable for a training system that also trains for source- and environment-related attributes.

Matters become more complicated when one considers the immersion attributes detailed in the Scene-Based Paradigm. Rumsey subdivides the concept of envelopment into three attributes, along the hierarchical levels of the dimensional aspects of the paradigm (individual source envelopment, ensemble source envelopment and environmental envelopment). Whilst this is a useful way to fully describe spatial audio scenes, the way in which these concepts overlap with the dimensional attributes contained in the Scene-Based Paradigm make it difficult to include them all in a new paradigm that needs to contain only non-redundant attributes.

For example, individual source envelopment is defined as the “sense of being enveloped by a single sound source” (Rumsey 2002). The examples given indicate that individual source envelopment seems to consist of the feeling of being within a sound source, or that a dry source has been panned very wide so that it wraps around the listener. This appears to be an amalgamated concept then, consisting of the phenomena of very wide sources (perhaps at a fixed radial distance from the listener), and the subjective sensation of being placed within the boundaries of a sound source. Any simulation of individual source width would therefore overlap with part of what Rumsey terms individual source envelopment. Attribute simulations of individual source distance would also be seen to overlap with the
concept of individual source envelopment as the source becomes close enough so that the listener perceives that they are within its boundaries.

Ensemble source envelopment is defined by Rumsey as the “sense of being enveloped by a group of sound sources” (Rumsey 2002). Neher (for the specific set of stimuli that he used) found that subjects perceived ensemble width to vary continuously when sources within an ensemble were spread along a circular arc about the listener, a term that he classified as constant distance ensemble width (Neher 2004). Therefore attribute simulations involving Neher’s definition of “constant distance” ensemble width, as well as ensemble distance and potentially ensemble depth can be seen to overlap with what is classified as ensemble source envelopment by Rumsey.

Regarding environmental envelopment, this is a term that Rumsey likens to LEV in concert hall acoustics research, defining it as the “sense of being enveloped by reverberant or environmental (background stream) sound”. According to Morimoto (2001), LEV is related to, amongst other things, the relative level of sound arriving from the rear hemisphere surrounding the listener to that arriving from the front hemisphere, and increases with relatively increasing rear incident energy. Attribute simulations involving changing levels of environmental envelopment can therefore be expected to overlap conceptually with what may be perceived to be environment depth. The notion of the environment extending around the side and to the rear of the listener could be described as the perception of increasing environment depth or increasing environmental envelopment. However environment depth has not (yet) been elicited as an attribute in previous studies, so may not be directly applicable to the training paradigm (it does not “relate to reality”).

The final immersion attribute in the Scene-Based Paradigm is presence, defined by Rumsey as “the sense of being inside an (enclosed) space or scene”. He mentions that presence and environmental envelopment may be closely related (may overlap conceptually), and states that one must feel presence in order to be able to feel enveloped by the environment. The presence attribute is seen to conflict with the previously delineated criteria in a number of ways. One definition of presence is “the state or fact of existing, or being present in a place or thing” (Pearsall (Ed.) 1998). It seems the latter part of this definition is closest to presence as defined by Rumsey. Berg defined presence as “the experience of being in the same acoustical environment as the sound source, e.g. to be in the same room” (see Section 2.1.7.3), but goes on to equate it to Toole’s perspective attribute (see Section 2.1.4.2) reproduced below:

“Refers to your general impressions of the experience. A good reproduction of a good recording with natural room or hall acoustics should suggest that ‘you are there’ at the performance, complete with a sense of the enveloping ambient sound. A less perfect reproduction could separate you from the performance, giving the impression that you are ‘close, but still looking on.’ In a still worse reproduction it may seem that you are listening through an opening between the loudspeakers. It is as though you were ‘outside looking in’ - there is no impression of being within the ambient sound. Other recordings may appear to transport the musicians to the listening room, ‘they are here.’ The ambiance is
The Simplified Scene-Based Paradigm

that of the listening room, and the instruments sound close. Still other recordings are created as abstract special effects, with no attempt to simulate a realistic experience" (Toole 1985).

The various levels of Rumsey's *presence* attribute can be seen as part of this definition. Examples of levels of presence given by Rumsey include "present", "absent", "outside the event", "in a corridor outside", in the "centre of the sound". Depending on the definition, presence can therefore be seen to be at worst dichotomous (subjects either feel present or absent from the space), or at best not applicable to a linear scale. In this way presence cannot be said to vary continuously along a single perceptual scale. Toole's *perspective* (and by extension, Rumsey's *presence*) can also be seen as an overall rating, based upon, amongst other things, the proximity of sources ("instruments sound close"), *environmental envelopment* ("sense of the enveloping ambient sound"), the *width* of the reproduction ("listening through an opening between the loudspeakers"). Toole also implies that *perspective* (and therefore *presence*), involves a measure of the realism of the reproduction ("simulate a realistic experience"). *Presence* is also found in the definition of the *sense of space* attribute used in (Koivuniemi and Zacharov 2001): "This attribute scales how well the space where the recording was made is perceived. A positive value could mean a strong sensation of being in a certain kind of environment, e.g. in a room". Put another way, this definition implies that *presence* (as used by Rumsey) is dependent upon how realistic ("well") the environment is reproduced. Further evidence of the amalgamated nature of *presence* is found in Berg's study, where he states that "the perception of different aspects of the room was most important for the feeling of presence" (Berg and Rumsey 2000a). The term *position presence* can also be found in the press (Salmon 1950), but it is used in a similar way to *ensemble depth*. Salmon (1950) also uses the terms *intimacy presence* (in a similar way to *source distance*) and *detail presence* (which can be thought of in timbral terms like *clarity* or perhaps to do with the *widths*, *focus* and *position* of the sources. All of this is evidence that *presence* is a holistic or amalgamated term, overlapping with *source distance*, *environmental envelopment* and *width* perceptions, and is possibly equivalent to, or a strong part of the feeling of *realism* generated by the reproduction.

In addition, *presence* is also seen to be potentially *ambiguous*, having been used in previously reported studies to indicate proximity of sound sources (*source* or *ensemble distance* using the Scene-Based Paradigm terminology). In audio mixing, *presence* is used to mean "bringing 'forward' an instrument (or voice) by selectively amplifying a range of frequencies which contains much of its character" (Nisbett 1993). Allen *et al.* (1969) specified the range of frequencies forming the "presence band" as 2-4 kHz. Toole also used the term *presence* in his *nearness/presence* scale (see Section 2.1.4.2), apparently using *presence* interchangeably with the perception of the proximity of sources. These instances of conflicting definitions can be seen as semantic problems that could be solved through rigorous definition and the use of *auditory description* stimulus sets. However, the differing use of the term presence, in addition to its amalgamated and holistic nature and overlap with other terms in the paradigm make it unsuitable for inclusion in the paradigm.

Regarding the *miscellaneous* attributes from the Scene-Based paradigm, changes in spatial audio scenes can be described in terms of *skewing* of the *scene* (either front/back or left/right), and changes in *scene width homogeneity*, *scene* and *source stability*, and *source focus*. 
Skewing was intended to be able to explain the panning of the entire scene from left to right or potentially from rear to front (and vice versa), but like other scene attributes, it overlaps with source-related and environment-related attributes, as skewing of the scene could involve, amongst other things, positional and width or depth changes of the sources or ensembles and changes in environment width, depth or envelopment.

Scene width homogeneity is defined by Rumsey as “evenness of distribution of scene elements compared with a reference scene”, and is reminiscent of Toole’s continuity of the sound stage attribute defined thus: “is the display of sound images continuous, left to right, or are there illogical groupings of images, with large gaps in between? Is the reverberation uniformly displayed or is it concentrated in strange places?” (Toole 1985). These attributes are used to describe changes in the position of sources, and can therefore be seen to overlap with attributes that describe the position and perhaps dimensions or sources. It does not necessarily overlap with environment width, because the definition leaves the possibility for two environments of equal width but a different distribution of reverberation across their widths. A final problem with scene width homogeneity is the inapplicability to linear scales. There appears to be no specific direction in which scene width homogeneity varies, making the creation of a linear scale of scene width homogeneity impossible. When considered in light of the overlapping with source-related attributes, the case has been made to exclude scene width homogeneity from the training system. Unresolved issues relating to the distribution of reverberation may have to be considered using a separate attribute that contains a direction. For example “concentration of reverberant energy towards loudspeakers” may be viable, but would necessitate that the loudspeakers were viewable by the subjects, which may not be ideal for experimental verification tests.

Two miscellaneous attributes deal with stability. Scene stability deals with the “degree to which the entire scene remains stable in space with respect to time” (Rumsey 2002), whereas source stability is defined as the “degree to which individual sources remain stable in space with respect to time (assuming nominally stationary sources)” (Rumsey 2002). Stability, whilst assuming notionally stable sources (and presumably environments), introduces an element of motion into the paradigm. In a similar way to scene width homogeneity, stability is a concept that is difficult to scale in a linear way. If a source is said to not be stable, it presumably moves spatially with respect to time, yet in order to describe this, and hence create a unidimensional simulation, one would need to break up this motion in terms of direction, range of motion, rate of motion and so on into continuously varying scales. Sense of movement has been elicited from subjects in the work of Zacharov & Koiivuniemi (see Section 2.1.6), due to the inclusion of scenes that included moving sources in their stimulus set (however, if movement attributes are to be included in the training set, then less amalgamated terms than sense of movement would be needed). Movement can also be thought of in terms of a simple change of position with respect to time, and in this sense a simulation of source motion would overlap with the concept of source position. The majority of the attributes in the Scene-Based Paradigm are considered to be time-invariant, so it may be possible for subjects to use other descriptive attributes (such as source position) along with an indication of time elapsed to describe such phenomena as source motion or the stability of sources or scenes.
The final miscellaneous attribute is source focus, defined by Rumsey as the “degree to which individual sources can be precisely located in space (this might be closely related to ISW)” (Rumsey 2002). Here the potential overlap with individual source width is found in the definition of the term. Source focus is not necessarily negatively correlated with source width, as Rumsey notes that it is possible to have a large source that is focussed, however it is more difficult to imagine a very narrow, yet defocused source. Other than in those of Berg & Rumsey outlined in Section 2.1.7.3, and in (Lee 2006) which was touched on in Section 2.1.7.3 (however this author believes that Lee’s test subjects were pre-biased to rate source width and locatedness differently because they were only asked to rate two attributes during the test), no distinction has been made in previous studies between source focus (or localisation/locatedness) and source width, and it is doubtful that the two concepts are really distinguishable by subjects. Even in the Berg & Rumsey studies localisation and source width were found to be negatively correlated and in addition source width and localisation were also found to be inconsistently used by subjects (see Section 2.1.7.2).

In summary, although the Scene-Based Paradigm allows for detailed and rigorous description of spatial audio scenes, many of the attributes within the paradigm can be seen to overlap with one another when used together. Inclusion of all such attributes into a training system for spatial audio attributes would prove problematic as “it is very confusing, demotivating, and mentally frustrating to the panellists when they are asked to score redundant terms” (Lawless and Heymann 1999). Certain terms, for example overall scene dimensional and miscellaneous attributes as well as presence were seen to be amalgamated terms. Another term, environment depth was deemed not to “relate to reality”, as it had not been elicited or used in previous experiments. Source movement issues were argued to be describable in terms of changes in the other attributes (most notably direction) providing that the subjects are given a time against which to make their judgements. The next section will outline a simplified scene-based paradigm that aims to negate the redundancy found in Rumsey’s Scene-Based Paradigm, whilst maintaining the ability to describe a wide range of spatial audio scenes with additional flexibility to specify descriptive elements according to the task or the stimuli under test.

### 3.2.3 The Scene Component (SC) Concept

As discussed in the previous section, the hierarchical nature of Rumsey’s Scene-Based Paradigm means that there is considerable overlap in terms, especially between the various source and ensemble terms. Scene terms were seen to be amalgamations of source-related and environment-related attributes, and the source-related envelopment terms were seen to overlap with source-related or dimensional attributes as well as source direction.

To overcome this problem of conceptual redundancy, the concept of the scene component (SC) is introduced.

Mason et al. (2001) used the concept of the scene component to allow separately perceivable elements to be identified that form part of the same sound object. This effectively allows what Rumsey (2002) defines a source to be classified as a scene component, or even a group of scene components themselves. Mason et al. (2001) indicate that this was necessary due to the fact that the source of the sound in
The Simplified Scene-Based Paradigm

reproduced sound is not the actual sound sources themselves, but rather loudspeakers or headphones. This author proposes an extension to the scene component concept to allow not only for a single source or individually perceivable elements of a source, but also the potential inclusion of a group of sources (for example, a choir or group of violins) into a single scene component – equivalent to Rumsey’s ensemble concept (see Section 2.1.8). In this way, a scene component can therefore be either a constituent part of what Rumsey termed a source, a source itself, or what Rumsey would term an ensemble (a group of sources). The definition of each scene component is flexible and can either be elicited from subjects in heuristically-orientated experiments (for example, experiments designed to examine subjective perceptions of stimuli), or specified by the experimenter in diagnostic experiments (for example, loudspeaker quality tests). The specific advantage of classifying all source-related elements of the spatial audio scene as being, or being part of scene components removes most of the hierarchical redundancy from the Scene-Based Paradigm, as each scene component can be specified so that it does not overlap with other separate scene components. Additionally, the use of specifiable scene components will allow for specific judgements to be made on individual areas of the reproduced spatial audio scene in a way which is consistent between subjects and hence possible to compare between subjects.

Figure 2 shows how three different scene components (SC) could be specified. SC (a) in Figure 2 constitutes part of what Rumsey would term a source. An example of this type of SC could be the different sound generating elements of a truck – the engine, exhaust pipe and a rattling plate could all be thought of as separate elements (SCs) within what could also be cognitively perceived as a single source (the truck) using Rumsey’s terminology. SC (b) in Figure 2 on the other hand could be specified as containing a single source using Rumsey’s terminology. An example could be a solo clarinet. SC (c) in Figure 2 however, is defined as a group of sources, for example four singers, or even a string quartet – this would be an ensemble using Rumsey’s terminology. Again, depending on the orientation of the experiment, subjects could either be provided with the defined SCs (for example: “consider the clarinet to be one SC, consider the string quartet to be another SC”), or they could be subjectively elicited (for example: “analyse the auditory scene and indicate how many SCs you perceive, and what is contained within each”). Flexibility and non-redundancy are the key benefits of the use of scene components. Because scene components can be specified in a rigorous manner, experimenters will also have confidence that they know which elements of the scene are being rated by subjects.

The scene attributes in the Scene-Based Paradigm (scene width, scene depth, and the scene skew, stability and width homogeneity attributes) as Rumsey defines them allow for multiple environments, and potentially for sources to exist outside these environments. Due to the previously delineated redundancy inherent in the overall scene attributes, they must be dispensed with for purposes of the training paradigm. The removal of the scene attributes has the ramification that multiple environments may not be specifiable as the environmental attributes will need to be referred back to a single environment. Scene components could, however, be described as being positioned outside the environment if they were perceived to be there. The inability to specify multiple environments is deemed to be an acceptable loss, as the majority of spatial audio scenes are at least intended to be reproduced in a single perceived
environment. A certain amount of simplification can also be justified within the context of a training system.

Figure 2: Three Examples of a Scene Component (SC)

The subdivision of source and room (or environment) was first performed in spatial attribute studies by Berg & Rumsey (see Section 2.1.7.2). It follows along the same lines as the distinction made in concert hall acoustics between ASW and LEV (Beranek 1996; Morimoto 2001), and this segregation into source-related and environment-related aspects is in keeping with the auditory system's method of segregation of streams (Bregman 1990), allowing the analytical paradigm to be close to the auditory system's innate methods.

The main advantage of the use of scene components for the proposed training system is that it removes the confusion that is created when changes in source attributes in the scene-based paradigm can conceptually give rise to changes in ensemble attributes and vice versa. Using scene components to define source-related elements of the scene allow changes therein to be described in a conceptually unidimensional manner.

To summarise then, spatial audio scenes can be analysed in terms of scene components (which consist of source-related sound) and the environment in which they exist. The next section will discuss how each of Rumsey's Scene-Based Paradigm attributes will need to be modified or excluded to take into account the new scene component concept and the previously noted concerns regarding adherence to the criteria for terms to be used in descriptive analysis.
3.2.4 Training System Spatial Audio Attributes

Implicit in Rumsey’s Scene-Based Paradigm is the concept of lateral position, the perceived direction to which the sound is located, measured from a given reference position (the centre-front reference, for example) to the sound’s midpoint. Although it was not discussed in detail in the Scene-Based Paradigm (mainly due to the amount of research already conducted on human perception of lateral angle), the lateral position of sources is considered to be a vital means of describing the sound scene and is therefore included within the training system paradigm as scene component direction. A definition of scene component direction could be “the perceived signed angle between the centre front reference position and a line connecting the listener with the perceived mid-point of the scene component”. Regarding the signed angle, this could be in the same form as those used in (ITU-R 1992-1994), where angles measured from the centre front reference line in a clockwise direction are positive, and those in an anticlockwise direction are negative.

Once scene component direction is determined, the other source-related dimensional attributes from the Scene-Based Paradigm can be seen to apply to the scene component.

Each scene component will therefore be at a certain scene component distance from the listener, defined as ‘the perceived egocentric distance between the mid-point of the scene component and the listener’. The use of the term “egocentric” is in line with two terms introduced by Loomis (1995). Egocentric distance (measured from the listening position) and exocentric distance (measured between two points that are not the listening position). Scene component distance can be seen to be equivalent to either individual source distance or ensemble distance, depending on the context in which it is used.

There will also be a specifiable scene component depth, defined as ‘the perceived distance between the nearest and furthest extent of the scene component, measured radially along the egocentric line connecting the perceived centre of the scene component’. Using terminology proposed by Loomis (1995), scene component depth is an exocentric distance. Scene component depth can be seen to be equivalent to either individual source depth or ensemble depth, depending on the context in which it is used. Individual source depth is an attribute that may not “relate to reality”, in that it was Berg’s label for a group of constructs that do not actually include the term source depth, but rather terms such as “sound source is V-shaped/sound source sits closer to the listener” and “large sound source/small sound source” (Berg and Rumsey 2000b). Additionally, this only appears where phase reversed items were included as stimuli. Supporting evidence that the concept of individual source depth is an important spatial attribute has not been found elsewhere in the literature. Ensemble depth on the other hand, was reported to have been successfully simulated in (Neher 2004).

In a similar way, scene component width can also be included in the training paradigm. Scene component width could be defined either as “the perceived width of the scene component, measured from the extremes of the perceived extent of the scene component on the lateral plane perpendicular to the line connecting the mid point of the scene component and the egocentre”, or as “the perceived width of the
scene component, measured from the extremes of the perceived extent of the scene component on the lateral plane in a circular arc about the egocentre”. The former definition follows Rumsey’s definitions of individual source width and ensemble width. The latter definition is in keeping with Neher’s finding that subjects perceived ensemble width for his stimuli as continuously varying when sources within the ensemble are repositioned at a constant distance about the listening position, a concept that he called constant distance ensemble width. Scene component width can be seen to be equivalent to either individual source width or ensemble width, depending on the context in which it is used. Whether individual source width is a viable attribute is debatable. Neher had difficulty simulating it in a unidimensional manner (Neher 2004), and Berg & Rumsey found that source width was used inconsistently by subjects (see Section 2.1.7.2). Ensemble width was, however, reported to have been successfully simulated in (Neher 2004). Whether scene component width should be rated perpendicular to the egocentric line or in a circular arc about the listener (or both/either) is a question that may need to be resolved with the help of a future experiment.

Because overall scene attributes are to be excluded, environment width can be included as the last width attribute, defined as ‘the perceived width of the (reflective) environment within which individual scene components are located’.

Because there is no evidence to say that environment depth is an important spatial attribute (or even that it is perceived with any reliability by subjects), the decision has been made to exclude it in favour of an immersion attribute: environment envelopment, with which, it was argued, it would have overlapped. Environment envelopment is a term that is much more established, being likened to LEV in concert halls in (Rumsey 2002). Environment envelopment could be defined as ‘the extent to which the listener perceives that he is immersed in a soundfield that extends to the sides and then to the rear of the listening position’. Whilst not ideal, this definition puts a direction to a term that is very loosely defined elsewhere in the literature.

Two other immersion attributes: individual source envelopment and ensemble source envelopment, can be described in terms of the above attributes, and have therefore been excluded. Individual source envelopment perceptions can be described in terms of scene component distance and depth – an enveloping source could be very near, or could be close and very deep (so that the boundary of the scene component envelops the listener). Alternatively it could also be very wide (with scene component width being considered along a circular arc about the listener), so as to encircle the listener at a specific distance. Ensemble source envelopment perceptions could now simply be described in terms of scene component directions and distances (extending around the sides and towards the rear) and scene component width and depth (extending around the sides and towards the rear).

The final immersion attribute, presence, has been excluded due to its dichotomous and/or non-linear nature as well it being ambiguous and amalgamated. Understanding of the included spatial attributes could help in future studies into what constitutes a reproduction exemplifying good and bad levels of presence.

Regarding the miscellaneous attributes, the case has been made for the exclusion of all miscellaneous attributes from the training paradigm, especially for this initial version of the training system.
This leaves the following attributes for inclusion in the paradigm for use in spatial audio training applications:

**Source-related attributes:**
- Scene component direction
- Scene component distance
- Scene component width
- Scene component depth

**Environment-related attributes:**
- Environment width
- Environment envelopment

Figure 3 shows how elements in the new paradigm will fit together to allow a description of spatial audio scenes.
3.3 Discussion

One potential accusation that can be levelled at the concept of *scene components* is that they are themselves amalgamated terms, referring to *sources*, elements of *sources* and *ensembles*. This is, however, not the case, as each *scene component* refers to a specific element of the scene either specified by the experimenter (for analytical experiments) or defined by the subjects according to how they perceive the scene to be divided up. In this way, the attributes that pertain to each *scene component* are orthogonal with respect to other attributes pertaining to the same scene component, and any attributes pertaining to other scene components or the environment. It is expected that a number of attribute simulations could be devised to exemplify changes of the attributes of various types of *scene components* in order to train the subjects fully for the terms and situations that they may encounter in listening tests. In devising this paradigm for use in a spatial audio training system, a certain degree of accuracy has had to be sacrificed for simplification and removal of redundancy. The potentially debatable level of amalgamation introduced in the *scene component* concept is considered to be acceptable when one considers the potentially intractable problems in the overlapping found in the hierarchical structure of the Scene-Based Paradigm. Training in the decomposition of various scenes into scene components is also intended to aid subjects in any future experiments involving the analysis of complex scenes, even when asked to use combined descriptors, like those in the ITU-R standard (ITU-R 1994-1997).

3.4 Summary

The issues facing the use of a paradigm for training in spatial audio attributes were discussed with reference to recommendations made for descriptive analysis terms by researchers in the food sciences. A set of criteria were studied and found to provide significant support for the use of reference stimulus sets, and a number of pointers to the selection of terms appropriate for inclusion in a descriptive analysis paradigm, such as the proposed training system. Rumsey's Scene-Based Paradigm was discussed in light of the criteria and found to contain, amongst other things, a number of overlapping terms. A new paradigm was designed that is based upon a modified version of the Scene-Based Paradigm and that keeps *source-related* and *environment-related* terms separate, allowing detailed description of spatial audio scenes through the concept of *scene components*. This avoids the confusion and frustration associated with grading redundant terms. Although *scene component* attributes can be criticised as being potentially holistic or amalgamated terms, they can be seen to be orthogonal within each scene component, and do not overlap between scene components. The paradigm developed in this chapter meets all of the criteria specified in Section 3.1 and is able to describe a wide variety of scenes. This enables the decomposition of any *source-related* sound into a number of *scene components* and allows for two orthogonal descriptors of the *environment* in which the *scene components* are located.
4 PILOT INVESTIGATION

An investigation was undertaken to determine whether formal training using unidimensionally varying stimuli such as those developed by Neher (2004) would benefit the performance of 'naïve' subjects in the evaluation of 'real-world' (i.e. not artificially contrived) audio stimuli.

The investigation incorporated five experimental sub-sections:

- The collection of ‘real-world’ stimuli for subsequent evaluation
- Subjective attribute rating by an experienced listener panel
- Subjective attribute rating by a group of ‘naïve’ listeners
- Training of a sub-group of the ‘naïve’ listeners
- Subjective attribute rating by the untrained and trained ‘naïve’ listeners

A series of objectives needed to be achieved in order to investigate the effect of spatial audio training on a real-world task. Firstly, a task was specified that involved the evaluation of one or more spatial audio attributes. The attribute or attributes involved needed to be identified and rated by experienced listeners in order to have a benchmark against which to measure any trainee’s performance, and in order to select appropriate stimuli for use in the training tests.

Next, naïve listeners were recruited to take part in the ‘real-world’ spatial attribute grading task. This allowed a benchmark of their pre-training performance to be recorded for subsequent analysis and to allow them to be separated into two evenly matched groups of listeners, one that was trained, and one that was not trained.

The training of listeners was achieved through a modified implementation of Neher’s pilot spatial audio attribute training system. This system had proved to be beneficial in (Neher 2004) and was available to this author. Performance was measured by examining the correctness of rank-ordering of the contrived stimuli within the training task – measures which had been utilised by Neher (2004) – and consistency and sensitivity measures of the grading data. The untrained subjects provided a control group to show the effect of repeating the task at a later date without the intervening training.

Finally, the trained and untrained listeners would need to take part in another grading exercise. This would allow their performance in the test to be compared between the two sets of sessions, allowing the effect of the training programme to be measured.

The pilot experiment was designed to examine the effect of participation in a training programme on listeners’ performance in subjective evaluation tasks using stimuli that are different to those in the training programme. This is in order to demonstrate far transfer (see Section 2.3.1.1) of the learned skills from one set of stimuli to another.
The overall methodology employed for the experiment is summarised below:

- The selection of an appropriate perceptual task to use as a verification of the training system.
- Recruitment, assessment and streaming of a group of 16 untrained (naïve) listeners, so that two sub-groups have similar performance skills.
- The training of one sub-group of the listeners using a spatial audio attribute training system.
- Verification of a training effect by testing both sub-groups of listeners using the perceptual task.

Any changes in performance of each group between the initial and final assessment would therefore be attributable to their participation or lack of participation in the training programme. Each of the above stages will be summarised in the following sections.

Some of the content of this chapter has been published previously in (Kassier, Brookes and Rumsey 2005, 2006b) and (Kassier, Lee, Brookes and Rumsey 2005).

4.1 The Collection of ‘Real-World’ Stimuli

For the planned investigation into the effect of spatial audio attribute training on subjective performance during listening tests, it was necessary to find a task that would be suitable for the subjects to perform pre- and post-training.

Such a task would need to involve the detection, identification and evaluation of changes in one or more spatial audio attributes. It would be possible to create a set of artificially generated stimuli that simulate changes in spatial audio attributes. Unlike Neher's study, the pre- and post-training task should utilise stimuli that were not found in the training task. This task would then be closer to the type encountered in real-world applications such as loudspeaker or microphone system evaluation.

It was therefore deduced that a set of acoustic recordings would be required for use in the pre- and post-training task. In order for subjects to detect, identify and evaluate spatial audio changes in a way that could be subsequently evaluated, the stimuli selected must demonstrate a range of ‘values’ for one or more spatial audio attributes.

There is, however, a lack of recorded audio test material in five-channel surround format. Particularly lacking are recordings made simultaneously using different microphone arrays that allow listeners to switch instantaneously between different recorded versions of the same acoustical event in order to compare them.

Multiple simultaneous recordings of the same acoustic event using different microphone configurations would create a set of stimuli with potentially varying spatial audio attributes that could be instantaneously switched between upon reproduction in order to evaluate them. It would also create stimuli that would allow the transfer of the learned skills to be tested between contrived stimuli used in training with non-contrived, complex signals.
An experiment was conducted by this author with a colleague (Dr. Hyun-Kook Lee). One of its goals was to create a set of simultaneous multichannel recordings for potential use in the pre- and post-training evaluation task. Full details of the background theory and experimental techniques are published in (Kassier, Lee, Brookes and Rumsey 2005), and the relevant sections have been reproduced for convenience in Appendix 7.1. A summary follows in Section 4.1.1.

4.1.1 Method

In order to produce multiple, simultaneously recorded versions of the same acoustical event, a method was devised to record 24 different microphone channels in a way which would allow them to be combined into 16 different simultaneous 5-channel recordings of various programme items.

This involved the placement of four ‘front’ microphone techniques (used to record primarily direct sound) and four ‘rear’ microphone techniques (used to record primarily diffuse reverberation) in Studio 1 at the Institute of Sound Recording (IoSR) at the University of Surrey. The studio is acoustically a typical concert hall, with a reverberation time of approximately 1.5 seconds. The signals from one of the four ‘front’ techniques could be mixed with the signals from one of the four ‘rear’ techniques to produce one 5-channel reproduction. Thus, 16 different combinations of the microphone signals could be assembled from the recorded signals. Figure 4 shows an overall view of the recording set-up in Studio 1.

*Figure 4: Photo of the recording set-up in Studio 1.*

The photo is taken from behind the centre of the ‘rear’ arrays, with the ‘front’ arrays visible in front of the piano.
Pilot investigation

The ‘front’ techniques (see Section 7.1.1.1 in Appendix 7.1) used were:

- ‘Fukada Tree’ technique consisting of three widely spaced cardioid microphones.
- A technique inspired by a ‘near-coincident’ technique suggested by Klepko, consisting of three cardioid microphones placed very close together.
- A technique inspired by the OCT technique proposed by Theile consisting of three cardioid microphones.
- INA-3 technique consisting of three cardioid microphones.

The ‘rear’ techniques (see Sections 7.1.1.2 and 7.1.1.3 in Appendix 7.1) used were:

- A four-channel ‘Hamasaki Square’ technique.
- A technique inspired by the four-channel ‘IRT-Cross’ technique.
- A two-channel ‘Dummy Head’ technique, suggested by Klepko.
- A two-channel ‘Spaced Cardioid’ technique.

The techniques were chosen mainly due to practical issues (the quantity, quality and type of microphones available), but also to give the author experience in using a variety of techniques to inform future experimentation.

Over thirty different programme items were recorded during the experiment, mostly involving solo instruments, or single instruments accompanied by piano.

The signals from the microphone arrays (24 microphone channels) were recorded onto digital audio tape then transferred to a Digital Audio Workstation (DAW) over analogue connections and sampled at 16-bit depth and 44.1 kHz rate. Once transferred to the DAW, the various microphone signals could be combined to create the sixteen simultaneous 5-channel reproductions.

The recordings and DAW files were then transferred to a portable computer system which was subsequently set-up in the listening room at the IoSR in order for informal evaluation to be conducted.

4.1.2 Informal Evaluation

Initial informal evaluation of the programme items took place in the listening room at the IoSR. The acoustical parameters of this room conform to the requirements of the ITU-R Recommendation BS. 1116 (ITU-R 1994-1997). Five active loudspeakers (Genelec 1032A) were arranged in 3/2 stereo configuration according to the ITU-R BS. 775 Recommendation (ITU-R 1992-1994), spaced a distance of 2m from the listening position. A portable DAW was placed just in front of the listening position and connected to the five loudspeakers. A computer monitor and mouse were provided in order to control the software.

During informal evaluation, it was found that the spatial reproduction of the two four-channel ‘rear’ microphone techniques was very similar, as was the spatial reproduction of both two-channel techniques. It was therefore decided that only one of each of the two-channel and four-channel techniques would be subsequently used. The Hamasaki Square (4-channel) and the Spaced Cardioid (two-channel) were...
chosen, partially because of certain unavoidable technical problems with the other two techniques.

All four 'front' techniques produced very different spatial reproductions, and informal evaluation suggested that the most salient spatial audio attribute change that occurred between the various items was Scene Component (SC) Width (referring to either the ensembles or the soloists themselves).

It was therefore decided to make SC Width the focus of the investigation and subsequent training system.

Twelve programme items were selected for further evaluation. There were:

- Piano (more continuous music from a romantic era Sonata)
- Piano (staccato music from a 20th century Toccata)
- Harpsichord (a minimalist piece for keyboard played on the harpsichord)
- Piano-Accordion (a mixture of continuous and transient music)
- Solo Soprano (romantic era aria)
- Accompanied Soprano (renaissance era aria)
- Solo Violin (a baroque era solo sonata)
- Accompanied Violin (modern Christian 'worship' music)
- Solo Trumpet (traditional melody, relatively continuous)
- Accompanied Trumpet (jazz)
- Solo Clarinet (classical clarinet concerto played solo, relatively continuous)
- Solo Trombone (bombastic music, relatively staccato)

Short loops were created at musically appropriate places within the programme items and eight different versions (called 'processes' in later stages of this report) of the twelve different programme items were created and saved as multichannel '.wav' files for subsequent use in the evaluation phases.

4.1.3 Summary

In order to provide a set of multiple simultaneously recorded programme items for use in comparative evaluation, a number of programme items were recorded using a multiple-microphone array. Using a combination of different microphone signals, up to sixteen simultaneous versions of each programme item could be created, although this was reduced to eight after informal evaluation. Informal evaluation also suggested that the experiment was successful in creating a set of stimuli that changed in terms of perceived spatial audio attributes. The most salient change across all versions of all programme items was SC Width (referring to either the ensembles or the soloists themselves). Twelve programme items were selected for subsequent evaluation. Eight simultaneous versions of the twelve programme items were created as multichannel '.wav' files that could be looped for subsequent evaluation.
4.2 Subjective Attribute Rating: Experienced Listener Panel

Before the multichannel stimuli that had been recorded were used in subjective evaluation tasks, they were evaluated by a panel of experienced listeners. This was done in order to obtain subjective values for the SC Width of the various versions of the various programme items as well as other performance-related measures. These values could later be compared to those obtained by the trained and untrained subjects in order to gauge training effects. SC Width values could also potentially be used to compare the various versions of the various programme items in order to reduce the number of stimuli based upon the subjective performance.

4.2.1 Method

The listening tests were conducted in the Listening Room at the IoSR. A computer (Silicon Graphics SGI O2) running proprietary listening test software (ALEX) was used to run the listening tests. It was connected to the loudspeakers in the listening room via an eight-channel digital audio interconnect and a digital mixing console (Yamaha O2R). A computer monitor and mouse were situated in front and below the listening position to allow the subjects to control the test computer.

Eight experienced listeners (referred to later in the report as subjects 1-8) took part in the informal comparisons (seven members of the IoSR and one final-year undergraduate student on the University of Surrey's Tonmeister course who was experienced in the recording of classical music in surround format). They were not paid for taking part in the investigation. During recruitment and during the introduction the benefits of taking part in the experiments were explained to the subjects in terms of improvements in their critical listening skills and experience in the listening room and with listening tests in general. This was done to foster beneficial intrinsic motivation – see Section 2.3.2.1.

The test consisted of six phases: familiarisation, practice and four experimental sessions. All subjects were also required to sign a standard departmental listening test consent form after having read the instructions.

Each of the experienced subjects was given the chance to familiarise themselves with all 96 stimuli to be used in the test (eight versions each of twelve programme items) during a familiarisation phase. Subjects were presented with eight unmarked (individually randomised) buttons for each of the programme items as shown in Figure 5.

During the familiarisation stage, subjects were asked to think about the ‘width’ of the soloists or ensembles, taken as the angle subtended between the left and right edges of the sound source or sources (SC Width of either a soloist or ensemble). They were asked to formulate their own scale of width on a 100-point scale that they would use in the upcoming experiments. Subjects were given as much time as they needed on the familiarisation task (which varied from seven to fifteen minutes).
After they were happy with the familiarisation page, the subjects were asked to move onto a ‘practice page’ which allowed them to begin using the scale that they had developed during familiarisation on a selection of the stimuli. The practice page contained three different versions (Fukada Tree & Hamasaki Square, Near Coincident & Hamasaki Square and INA-3 & Spaced Cardioid) of three different programme materials (Harpsichord, Accompanied Trumpet and Solo Clarinet). These were chosen because they were judged by the author to demonstrate a wide range of width values. They were labelled A, B, C; K, L, M; and X, Y, Z respectively by programme material (the order of these triplets of sound files was randomised between subjects). Figure 6 shows the practice page.

The subjects’ task was to evaluate the SC Width (referred to as ‘Width of soloist or ensemble’ in the instructions) of each of the items, then position the appropriate
slider on the 0-100 point scale according to their evaluation of the width of the soloist or ensemble. Subjects were informed that their results would not be used in the analysis, and that they should take as much time as they needed (which ranged from two minutes to eight minutes).

Once the familiarisation and practice phases were completed, the subjects took part in four nominally 30-minute experimental sessions. These were completed within two to five days of the familiarisation and practice stages. Each experimental grading session consisted of six pages, each containing eight versions of one of the programme items, available as eight on-screen buttons. As in the practice page, the subjects' task was to evaluate the SC Width of each of the items, then position the appropriate slider on the 0-100 point scale according to their evaluation of the width of the soloist or ensemble. The presentation of the programme items and position of the 'process' type behind each of the buttons was randomised for each subject across the two sessions that formed a complete iteration. See Figure 7 for the layout of each experimental page.

As six programme items were evaluated per session, two sessions were required to evaluate all twelve programme items under investigation. Over the course of the four sessions, the eight versions of each programme item were evaluated twice (once during the first two sessions and once during the third and fourth sessions).

4.2.2 Results

The results of the width rating experiment for the experienced listeners has been summarised in Figure 8, Figure 9, Figure 10, Figure 11 and Figure 12.

Figure 8 shows the mean 'width' grades assigned by each of the subjects to all 'processes' of all programme items. This gives an indication of the way in which each subject used the width scale. One can see that subjects 2, 4 and 7 tended to give lower values of width than the other five subjects.
Figure 8: Mean width grades given for each subject.

Figure 9 shows the mean ‘width’ grades assigned to each ‘process’ (a different combination of one of the “front” microphone techniques with one of the “rear” microphone techniques). Across all programme items, the subjects tended to rate the different processes in different ways. The combination of the OCT-inspired front microphone technique with the Spaced Cardioid rear microphone technique was judged to have produced the narrowest scene components across all programme items by all subjects. The Near-Coincident front microphone technique combined with the Hamasaki Square rear technique was judged to have produced the widest scene components across all programme items by all subjects. Other processes were spread fairly evenly in terms of width between these two extremes.

Figure 9: Mean width grades given to each ‘process’.
Pilot investigation

Figure 10 shows how each front technique varied in terms of width grades given by all subjects across all programme items. The 95% confidence intervals (CI) overlap slightly between the Fukada and OCT techniques, indicating that no statistically significant differences existed between the width grades given to stimuli incorporating the Fukada Tree front technique and those incorporating the OCT-inspired techniques. However statistically significant differences existed between all other combinations of front techniques. The recording angles (Herrmann and Henkels 1998) of the “Fukada”, “INA” and “OCT” techniques were deliberately calculated to be as close as possible (Kassier, Lee, Brookes and Rumsey 2005), resulting in theoretically similar sound stage widths. It is interesting therefore to note that, in terms of SC Width, there were statistically significant differences between the INA, Fukada and OCT techniques.

Figure 11 shows that significant differences existed between the width grades given to stimuli that incorporated the Hamasaki Square rear microphone technique and the Spaced Cardioid rear microphone technique.

All processes, all programme items. The vertical width scale displays a reduced portion of the scale for clarity.

Figure 10: Mean width grades by front microphone technique.
The above results indicate that the stimuli assembled for this study exhibit significant differences in terms of width. This is an important factor, as if the differences between the stimuli were too small, all subjects (whether trained, untrained or experienced) would find the task frustrating and unduly fatiguing.

4.2.3 Reducing the Number of Programme Items

The results from the experienced subjects’ width ratings also assisted the author in the task of selecting a reduced set of six programme items for use in the subsequent naïve subject training experiments.

Six of the twelve items were selected for use in the remainder of the study in order to reduce the complexity and length of the training experiment. It was predicted that grading eight versions of all twelve items would be an overwhelmingly difficult task for naïve listeners and would place an undue burden upon their busy academic schedules. Reduction of the programme items also allowed the experiments to take place comfortably within the three weeks available for the training experimentation.

Figure 12 shows the mean grades assigned to each programme item by all subjects. It can be seen that two items were rated as consisting of relatively narrow SC Width (the Trombone and Solo Soprano items), whereas two items were rated as consisting of relatively wide SC Width (the Brahms Piano and Accordion items). In order to select six appropriate items, the graph was visually inspected in order to find six items that covered a spread of the range of width values.
Pilot investigation

Figure 12: Mean width grades by programme type.

All microphone techniques, all subjects. The vertical width scale displays a reduced portion of the scale for clarity.

Figure 13 shows the spread of mean width grades for the items that were not selected for the training experiment. All three 'accompanied' items were not selected due to both their similarity of widths and the additional complexity involved with rating a Scene Component containing multiple sources.

Figure 13: Mean width grades of items not selected for the training experiment.
Figure 14: Mean width grades of items selected for the training experiment.

Other important factors governing the suitability of the stimuli are whether it is easy to discern differences between different 'processes' of the item, and how difficult it was to grade each of the processes consistently. In order to gauge these factors within the selected stimuli (to verify that they are indeed suitable for the task), an ANOVA test was performed for the Width grades assigned by all subjects to each of the programme items. To allow for inter-subject variability to be taken into account by the ANOVA, 'subject number' was incorporated into the ANOVA as a fixed factor. In order to gauge the variability in the grades assigned to each 'process', the variable 'stimulus' was also incorporated into the ANOVA as a fixed factor. Sample ANOVA tables are shown in Table 15 and Table 16.

Table 15: Sample ANOVA table for the Accordion Programme Item

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>60915.219</td>
<td>63</td>
<td>966.908</td>
<td>6.715</td>
<td>.000</td>
<td>.869</td>
</tr>
<tr>
<td>Intercept</td>
<td>424350.781</td>
<td>1</td>
<td>424350.78</td>
<td>2946.88</td>
<td>.000</td>
<td>.979</td>
</tr>
<tr>
<td>STIMULUS</td>
<td>34313.719</td>
<td>7</td>
<td>4901.960</td>
<td>34.041</td>
<td>.000</td>
<td>.788</td>
</tr>
<tr>
<td>SUBJECT</td>
<td>15197.844</td>
<td>7</td>
<td>2171.121</td>
<td>15.077</td>
<td>.000</td>
<td>.623</td>
</tr>
<tr>
<td>STIMULUS * SUBJECT</td>
<td>11403.656</td>
<td>49</td>
<td>232.728</td>
<td>1.616</td>
<td>.036</td>
<td>.553</td>
</tr>
<tr>
<td>Error</td>
<td>9216.000</td>
<td>64</td>
<td>232.728</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>494482.000</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>70131.219</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .869 (Adjusted R Squared = .739)
b. Programme Type = Accordion

(Mean Square Error is emphasised).
Pilot investigation

Table 16: Sample ANOVA table for the Clarinet Programme Item

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>39792.000</td>
<td>63</td>
<td>631.619</td>
<td>2.692</td>
<td>.000</td>
<td>.726</td>
</tr>
<tr>
<td>Intercept</td>
<td>188498.000</td>
<td>1</td>
<td>188498.000</td>
<td>803.401</td>
<td>.000</td>
<td>.926</td>
</tr>
<tr>
<td>STIMULUS</td>
<td>10527.625</td>
<td>7</td>
<td>1503.946</td>
<td>6.410</td>
<td>.000</td>
<td>.412</td>
</tr>
<tr>
<td>SUBJECT</td>
<td>7518.000</td>
<td>7</td>
<td>1074.000</td>
<td>4.578</td>
<td>.000</td>
<td>.334</td>
</tr>
<tr>
<td>STIMULUS * SUBJECT</td>
<td>21746.375</td>
<td>49</td>
<td>443.804</td>
<td>1.892</td>
<td>.008</td>
<td>.592</td>
</tr>
<tr>
<td>Error</td>
<td>15016.000</td>
<td>64</td>
<td>234.625</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>243306.000</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>54808.000</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. R Squared = .726 (Adjusted R Squared = .456)
b. Programme Type = Clarinet

(Partial ETA Squared of Stimulus is emphasised).

Figure 15 shows a scatter plot that summarises the data taken from ANOVA tables for all programme items. The ‘error’ figure is the root of the Mean Square Error (MSE) taken from the intersection of ‘mean square’ and ‘error’ as emphasised in Table 15.

The Mean Square Error is an indication of the amount of variance in the width grades assigned by this subject to the individual stimulus items (each of the processes of each of the programme items). The root of the mean square error allows this value to be expressed within the same scale as the original width values. The root of the mean square (RMS) error value (also referred to as ‘error’ in this thesis) is therefore an indication of how inconsistent the width ratings were for the individual stimuli by each subject.

Partial Eta Squared of ‘stimulus’ is taken from the intersection of ‘partial eta squared’ and ‘stimulus’ as highlighted in Table 16. Partial Eta Squared is a measure of the size of the effect in question. In this case, it indicates how much variability in the width grades can be attributed to the differences between each ‘process’ of the programme item in question. Partial Eta Squared of ‘stimulus’ can be considered a measure of how different the SC widths of each of the processes were within each item. This gives an indication of how easy it was to find differences between the widths of the different versions of each item.

If, for example, a subject rated every stimulus as having width ‘50’ in each session, they would achieve a very low RMS error score (zero, in fact) but also a very low value for Partial Eta Squared, indicating that the subject was very consistent, but did not distinguish between the widths of the various items under test.
From the scatter plot (Figure 15) it is possible to see that most of the selected items showed high partial eta squared values (indicating that it was relatively easy to determine SC Width differences for the various ‘processes’ of the programme item concerned), and also covered the range of error values. Both ‘piano’ items had similar characteristics, justifying the selection of one of them. The accompanied items (which were removed partially due to the additional complexity of rating multiple source scene components) showed similar characteristics to a number of the chosen items, indicating that they could be removed in favour of the selected items. The selection of the clarinet item (which had a low ‘partial eta squared’ value) could be revealing, as it could be considered a more ‘difficult’ item to find salient differences between its processes. It is interesting to note that the accordion item has both the lowest error and highest partial eta squared value. It is expected that it would be the ‘easiest’ item to grade. Overall, the scatter plot analysis indicated that selection of six of the programme items should cover the range of difficulty of the items whilst tending towards items that contain more salient width changes between their different processes.

4.2.4 Summary

A panel of eight experienced listeners was used to make judgements of the Scene Component Width of twelve short looped programme material items taken from the multiple microphone recordings described in Section 4.1 and in (Kassier, Lee, Brookes and Rumsey 2005). Significant differences in SC Width were found between items using each of the ‘front’ microphone techniques, and also between items using the two ‘rear’ microphone techniques, meaning that SC Width was indeed a salient spatial attribute change between the different versions of the
different programme items, and that the stimuli should be suitable for the ‘real-world task’. Six programme items were selected from the twelve for use in the later stages of the investigation:

- Solo Piano (more continuous music from a romantic era Sonata)
- Solo Accordion (a mixture of continuous and transient music)
- Solo Soprano (romantic era aria)
- Solo Violin (a baroque era solo sonata)
- Solo Harpsichord (a minimalist piece for keyboard played on the harpsichord)
- Solo Clarinet (classical clarinet concerto played solo, relatively continuous)

These were selected to cover the range of mean width values attributed by the experienced listening panel (in order to allow the naïve subjects to express as wide a variety of width values as possible in the tests). Solo items were preferred to items involving multiple source scene components in order to simplify the concepts that the naïve listeners were required to learn and use. The selected items were found to include various levels of ‘difficulty’ for the consistency of grading. Items featuring more salient differences between their processes (indicated by high Partial Eta Squared values) tended to be selected.

4.3 Subjective Attribute Rating: ‘Pre-Training’

In order to examine the effects of a spatial audio listener training programme upon the performance of naïve listeners within a ‘real-world’ experiment, a task has been established: rating of SC Width of various combinations of surround-sound microphone techniques recording the same acoustical event using six different programme items.

The format of the remainder of the pilot investigation was to recruit a number of naïve listeners, gauge their performance in the task without training, select and train a representative group of the listeners, and then calculate the improvement in performance of the trained and untrained groups.

The experiments were designed to take place within a three-week period during the spring semester in order that undergraduate students at the University of Surrey would be available to take part in the tests.

Due to funding restrictions, it was not possible to financially compensate the naïve listeners for taking part in the tests (unlike the pilot experiment reported by Neher (2004) where subjects were paid to participate). Intrinsic methods of motivation (see Section 2.3.2.1) therefore needed to be utilised in order to maintain attendance and encourage completion of the experiment. Intrinsic motivation should, however, result in greater learning (see Section 2.3.2.1).

A group of five undergraduate students on the Music course at the University of Surrey was used in the paid experiment reported in (Neher 2004). For this author’s pilot investigation, it was considered necessary that naïve subjects were drawn from backgrounds that would maximise their desire to participate. For this reason first year undergraduate students on the Music and Sound Recording (Tonmeister) course
and the new Music with Computer Sound Design (Music & CSD) course were approached. Students were recruited by means of a short introduction to the experiment that focussed upon the benefits of a listener training programme for the students’ critical listening skills, given during a break in a timetabled lecture for each group. Roughly half of each group provided their personal contact details as a means of expressing an interest in taking part in the study.

A balance needed to be struck between the need to have as many participants involved in the study as possible, and the time available for experimentation. It was estimated that a total of sixteen subjects could take part in the main ‘real-world’ experiment, with eight subjects participating in additional training. In order to maintain as balanced a group as possible, eight subjects (including six male and two female subjects) would be selected from each of the courses (Tonmeister and Music & CSD). This was in order to allow both genders from both courses to be represented in the trained and untrained groups. The available subjects were contacted via telephone and informed that they had been chosen to take part in the study. Once enough candidates from the respective groups had agreed and arranged their first session during the following week, the initial subjective attribute rating tests could begin.

The sample group can be described as both self-selecting (students were asked to express an interest, rather than be randomly sampled from a population) and quota-based (students who answered their phones and committed to the programme first were selected). These are seen as unavoidable limitations, as with an unpaid sample group, intrinsic motivation (such as the desire to improve one’s listening skills) must be harnessed in order to engage a sufficient number of subjects. It is expected, however, that the end-users of the training system will be intrinsically motivated (see Section 2.3.2.1). The sample group is likely to mirror them, at least within the respective age and experience groups that the subjects were drawn from.

The first part of the pre-training phase of the experimentation was to obtain two sets of gradings of SC Width of each of the ‘processes’ of each of the programme items from each of the subjects. A comparison of the grades assigned to each of the stimuli (each ‘stimulus’ is a specific ‘process’ of a certain ‘programme item’) by each subject would then allow an initial evaluation of their performance in terms of time taken for the task (fluency) and consistency in which they graded the stimuli. Not only was this important in creating a benchmark against which subsequent performance could be measured, it would also allow the selection of subjects into two groups with similar performance levels. One of these groups would then participate in the additional training sessions (before the final two iterations of the grading exercise), and one would act as a control group, being given no additional training.

4.3.1 Method

Sixteen subjects (referred to as Subjects 101-116 in order to not confuse them with the experienced subjects from Section 4.2) took part in an initial SC Width rating experiment. This involved the rating of the SC Width of eight different versions (processes) of six different programme materials involving solo musical instruments or voice.
Pilot investigation

The experimental set-up was similar to that described in Section 4.2.1 (although only the six programme items selected in Section 4.2.3 were used). Subjects were also required to sign a standard departmental listening test consent form after having read the instructions.

Each subject committed to participating in three 30-minute sessions during the first week of the experiment (the pre-training week), followed by two 30-minute sessions two weeks later (during the post-training week). They were informed that they may be selected to participate in additional training sessions (which should consist of no more than five additional 30-minute training sessions) during the week following the pre-training week (the training week).

Before beginning any of the experimental phases, all sixteen naïve subjects were first asked to complete a questionnaire on their listening habits. Subjects had to respond by indicating whether they had (in their view) experience in listening to four different types of music:

- Conventional stereo pop music
- Conventional stereo classical music
- Surround sound pop music
- Surround sound classical music

Subjects responded to a statement that they had experience in listening to each of the above types of music using a five-point ordinal scale (1=Strongly disagree; 2=Disagree; 3=Neither agree nor disagree; 4=Agree; 5=Strongly agree). On the whole, the responses from each of the courses were similar. Subjects tended to believe that they were experienced in listening to conventional stereo music, but not experienced in listening to surround sound music. In order to display the responses from each group, the five-point ordinal scale responses were converted to a scale from 1 to 5 and plotted as a bar graph shown in Figure 16.

![Figure 16: Relative Experience between Courses.](image)

Because the scale used was an ordinal scale, care should be taken when calculating and plotting means of the responses as the responses do not continuously cover the range of values. Notwithstanding, Figure 16 does show that the subjects' own estimation of their listening experience was similar regardless of course, and that they all did not consider themselves experienced in listening to surround sound. In addition, each subject was asked whether they had participated in controlled listening tests in the past. Thirteen of the sixteen subjects had not participated in listening tests, two others had done a previous test involving the width of noise in a previous
Pilot investigation

experiment, and one had done two other listening tests. These tests were all for final year Tonmeister students as part of their dissertations and were unlikely to involve intensive training. It was therefore assumed that the subjects could be considered ‘naïve’ in terms of their experience in controlled spatial listening tests.

The first three sessions consisted of a familiarisation and practice session and two 30 minute attribute grading sessions.

During the familiarisation stage, subjects were asked to think about the ‘width’ of the instrument or voice, taken as the angle subtended between the left and right edges of the sound source or sources (SC Width of the soloist). They were asked to formulate their own scale of width on a 100 point scale that they would use in the upcoming experiments. Subjects were given as much time as they needed on the familiarisation task (which varied from three to nine minutes).

After they were happy with the familiarisation page, the subjects were asked to move onto a ‘practice page’ which allowed them to begin using the scale that they had developed during familiarisation on a selection of the stimuli. The practice page contained three different versions (Fukada Tree & Hamasaki Square, Near Coincident & Hamasaki Square and INA-3 & Spaced Cardioid) of three different programme materials (Harpsichord, Solo Piano and Solo Clarinet). These were chosen because they were judged by the author to demonstrate a wide range of width values (the Accompanied Trumpet item from Section 4.2.1 was replaced by the Solo Piano item as it had not been selected for the training phases of the experiment). They were labelled A, B, C; K, L, M; and X, Y, Z respectively by programme material (the order of these triplets of sound files was randomised between subjects).

The subjects’ task was to evaluate the SC Width (referred to as ‘Width of Instrument or Voice’ in the instructions) of each of the items, then position the appropriate slider on the 0-100 point scale according to their evaluation of the width of the instrumental or voice. (The wording used in Section 4.2.1 ‘Width of Soloist or Ensemble’ was changed as the selected stimuli contained only solo instruments or voices to reduce the complexity of the task.) Subjects were informed that their results for the practice page would not be used in the analysis, and that they should take as much time as they needed (which ranged from two minutes to fifteen minutes).

Once the familiarisation and practice phases were completed, the subjects took part in the first two nominally 30-minute experimental sessions. These were completed within the first week of experimentation. Each experimental grading session consisted of six pages, each containing eight versions of one of the programme items, available as eight on-screen buttons. As in the practice page, the subjects’ task was to evaluate the SC Width of each of the items, then position the appropriate slider on the 0-100 point scale according to their evaluation of the width of the solo instrument or voice. The presentation of the programme items and position of the ‘process’ type behind each of the buttons was randomised for each subject and session.

As all six selected programme items were evaluated per session, the eight versions of each programme item were evaluated twice (once during the first session and once during the second session). The subjects SC Width grading evaluation for each stimulus was recorded. The time they took to complete each nominally 30-minute session was also recorded. The latter was in order to gauge if participation in a separate listening test programme would affect the time in which the subjects completed the task, and hence potentially be a measure of their fluency.
The resulting grades attributed to each stimulus by each subject were stored as text files on the SGI computer. These were then transferred onto another personal computer (PC) for subsequent processing and analysis. Results of the time taken for each experiment were recorded manually by the author during the tests.

4.3.2 Results

To get an overview of the data provided by the naïve subjects, bar graphs were plotted to show the mean and 95% confidence intervals of width grades for stimuli using specific ‘front’ microphone arrays (Figure 17), and for stimuli using specific ‘rear’ microphone arrays (Figure 18). Data for the width grades provided by the experienced listeners in Section 4.2 (for the six programme items used in the naïve tests) were also plotted to show the relative difference between the naïve and experienced groups.

Figure 17: Comparison of mean width grades by Experienced and Naïve listeners to stimuli incorporating different front microphone techniques (six selected programme items only).

From Figure 17, one can see that the grades assigned by the naïve and experienced listeners for the Near Coincident and INA items show no statistically significant differences. The grades assigned to the Fukada and OCT techniques, on the other hand were significantly lower for the experienced listeners than for the naïve listeners. The finer details show that there were other ways in which the naïve and experienced listeners graded the stimuli. For example, experienced listeners (for the stimuli selected for the naïve experiments) found similar widths between the Fukada Tree and OCT techniques, and between the Near Coincident and INA techniques. Naïve listeners, on the other hand, found statistically significant differences between the Near Coincident technique and the other techniques, but not between the other three techniques themselves.

It can be seen from Figure 17 that the means and CI of the naïve listeners span less of the 0-100 point width scale than those of the experienced listeners, confirming
Pilot investigation

Eisler's suspicions about the use of the extremes of attribute scales (see Section 2.1.1).

**Figure 18:** Comparison of mean width grades by Experienced and Naive listeners to stimuli incorporating different rear microphone techniques (six selected programme items only).

![Comparison of mean width grades](image)

The vertical scale displays a reduced portion of the scale for clarity. Values show means.

From Figure 18 one can see that significantly different grades were assigned to stimuli that involved the different 'rear' microphone techniques. Experienced and naive subjects both indicated that the four-channel Hamasaki Square microphone technique provided stimuli that were judged to have wider SC Width than that of the two-channel Spaced Cardioid configuration. It is surmised that the early reflection patterns within the additional channels of audio from the Hamasaki Square had combined with those from the left and right channels of the front microphone system to create wider scene components.

It is also interesting to note that naive listeners found relatively more difference in width between items with the different surround reproductions (the different rear microphone techniques) than with items involving different front microphone techniques. Conversely, the experienced listeners had found more difference in width between items involving different front techniques than those utilising different surround techniques. This can be seen in the difference between the maximum and minimum mean value attributed to the front technique stimuli and those of the rear technique stimuli. For experienced listeners, there was an 18 point difference (on the 100 point scale) between the front techniques and an 8 point difference between the surround items. For naive listeners, there was a 9 point difference between the maximum and minimum mean widths for front techniques, whereas there was a 15 point difference between rear techniques. Whilst not directly related, this sensitivity to changes in audio originating from the non-frontal arc (L-LS-RS-R in the 3/2 stereo set-up) could potentially be a similar effect to that which made naive listeners more susceptible than experienced listeners to spatial changes outside the frontal arc during preference testing of 5.1 surround sound material reported in (Rumsey, Zieliński, Kassier and Bech 2005).
Data regarding the subjectively evaluated SC Width of the solo instrument or voice for each subject were coded according to the session in which it was taken (Session 1 or Session 2). The grade assigned to each stimulus during the first grading session could therefore be compared to the grade assigned during the second grading session. Comparison of the actual grade values assigned to each stimulus can therefore provide an estimation of the consistency with which subjects assigned a width grade to each stimulus.

By performing individual ANOVA calculations on each of the subjects' width grades during the first two experimental sessions, a measure of the consistency by which subjects specified the grades for each stimulus, and a measure of the amount of sensitivity which they expressed when assessing differences between the items was possible. A sample resulting ANOVA table for one of the subjects (108) is shown in Table 17.

Table 17: Sample ANOVA table, Subject 108, Sessions 1 and 2 ('Pre Training' Sessions).

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
<th>Partial Eta Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>35944.406</td>
<td>47</td>
<td>764.775</td>
<td>3.031</td>
<td>.000</td>
<td>.748</td>
</tr>
<tr>
<td>Intercept</td>
<td>230790.094</td>
<td>1</td>
<td>230790.094</td>
<td>914.737</td>
<td>.000</td>
<td>.950</td>
</tr>
<tr>
<td>STIMULUS</td>
<td>35944.406</td>
<td>47</td>
<td>764.775</td>
<td>3.031</td>
<td>.000</td>
<td>.748</td>
</tr>
<tr>
<td>Error</td>
<td>12110.500</td>
<td>48</td>
<td>252.302</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>278845.000</td>
<td>96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>48054.906</td>
<td>95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- R Squared = .748 (Adjusted R Squared = .501)
- Subject No. = 108, Training = Trained, Pre/Post = Pre

Subject 108 eventually formed part of the group that later received additional training. Mean Square Error and Partial Eta Squared of 'Stimulus' are emphasised.

See Section 4.2.3 for an explanation of 'error' and Partial ETA Squared.

Table 18 shows the data resulting from the individual subject ANOVAs for the dependant variable 'width' over two iterations.

The data shown in Table 18 were plotted on scatter plot graphs (see Figure 19) to show the relative performances of each of the sub-groups of subjects (female and male subjects from each of the courses). Each graph shows how each subject (indicated by a round point with their subject number labelled to the right hand side) fared in terms of their RMS error value ('error') and their Partial Eta Squared value.

Because theories of differences in spatial awareness between males and females are both suggested (Geary 1995) and denied (Caplan, MacPherson and Tobin 1985), it was judged to be a safe course of action to maintain identical gender balances in the trained and untrained groups throughout the experiment. Wickelmaier & Choisel (2005) indicated that males tended to be better at spatial perception tasks than females.
Table 18: RMS Error and Partial Eta Squared values for the first two naive listener experimental sessions.

<table>
<thead>
<tr>
<th>Subject No.</th>
<th>Gender</th>
<th>Course</th>
<th>Error</th>
<th>P. ETA. Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Female</td>
<td>Music &amp; CSD</td>
<td>15.13</td>
<td>0.587</td>
</tr>
<tr>
<td>102</td>
<td>Female</td>
<td>Music &amp; CSD</td>
<td>17.22</td>
<td>0.625</td>
</tr>
<tr>
<td>103</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>13.14</td>
<td>0.686</td>
</tr>
<tr>
<td>104</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>12.08</td>
<td>0.808</td>
</tr>
<tr>
<td>105</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>14.95</td>
<td>0.741</td>
</tr>
<tr>
<td>106</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>12.11</td>
<td>0.638</td>
</tr>
<tr>
<td>107</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>14.66</td>
<td>0.678</td>
</tr>
<tr>
<td>108</td>
<td>Male</td>
<td>Music &amp; CSD</td>
<td>15.88</td>
<td>0.748</td>
</tr>
<tr>
<td>109</td>
<td>Female</td>
<td>Tonmeister</td>
<td>20.37</td>
<td>0.681</td>
</tr>
<tr>
<td>110</td>
<td>Female</td>
<td>Tonmeister</td>
<td>18.79</td>
<td>0.784</td>
</tr>
<tr>
<td>111</td>
<td>Male</td>
<td>Tonmeister</td>
<td>22.56</td>
<td>0.552</td>
</tr>
<tr>
<td>112</td>
<td>Male</td>
<td>Tonmeister</td>
<td>14.40</td>
<td>0.822</td>
</tr>
<tr>
<td>113</td>
<td>Male</td>
<td>Tonmeister</td>
<td>17.01</td>
<td>0.777</td>
</tr>
<tr>
<td>114</td>
<td>Male</td>
<td>Tonmeister</td>
<td>17.57</td>
<td>0.802</td>
</tr>
<tr>
<td>115</td>
<td>Male</td>
<td>Tonmeister</td>
<td>15.31</td>
<td>0.844</td>
</tr>
<tr>
<td>116</td>
<td>Male</td>
<td>Tonmeister</td>
<td>11.28</td>
<td>0.603</td>
</tr>
</tbody>
</table>

Subjects that are close to one another on the scatter plots have similar performances. Those positioned towards the top left of the graphs (for example, Subject 111 – a male Tonmeister student) have high error values (meaning that they are more inconsistent in the width grades they assigned to identical items), and low Partial Eta Squared values (meaning that they were relatively insensitive to width changes between stimuli). The lower the ‘error’ and the higher the ‘Partial Eta Squared’ values, the better the subject’s performance by these measures.

Figure 19: Scatter plots showing relative performances (RMS Error against Partial Eta Squared) of the 4 main subgroups of subjects (By Course, and by Gender).

Data points are labelled by subject number.
4.3.3 Selecting Candidates for the Training Programme

The data collected in the first two naïve subject experimental sessions were used to separate the subjects into two groups: one that would receive an additional training programme during the following week (the training week), and one that would not receive additional training.

In order to maintain the balance of the experimental design, it was decided that equal numbers of the subgroups displayed in Figure 19 would be placed in the training and non-training groups. For the two female subject pairs, one subject was chosen from each to be trained. For the male subject groups, pairs of subjects that showed similar performance (positioned closely together on the scatter plots in Figure 19) were identified within each group. This should ensure that the trained and untrained groups would have similar performance characteristics – it had been feared that random selection of the relatively small number of subjects involved in this test may have produced uneven performance characteristics between the groups.

The subject pairings were therefore:

- 101 & 102 (the female Music & CSD students)
- 103 & 106 (male Music & CSD students with very similar performances)
- 105 & 108 (male Music & CSD students with very similar performances)
- 104 & 107 (the remaining male Music & CSD students, 107 was less consistent and less sensitive than 104)
- 109 & 110 (the female Tonmeister students)
- 112 & 115 (male Tonmeister students with very similar performances)
- 113 & 114 (male Tonmeister students with very similar performances)
- 111 & 116 (the remaining male Tonmeister students, they had similar sensitivities, but 111 was more inconsistent)

In order to separate the pairs of candidates into the trained and untrained groups, the time that the subjects took for the first two experimental sessions was examined. Candidates were divided so that the mean time taken for the sessions by all subjects in the trained and untrained groups was as close as possible. This provided some control over the amount of time subjects in each group were taking to complete the task. It would also allow easy verification of the hypothesis that participation in the training programme would increase the fluency (and hence decrease the time taken) in performing the external task. If the hypothesis is correct, the mean time (and total time taken) for the final two experimental sessions should decrease by more for the trained group than for untrained group.

Just one external factor influenced the selection of the candidates. Subject 110 (one of the female Tonmeister candidates) had second thoughts about participating in the programme due to academic workload pressures, but offered to take part if she was not required to participate in the additional training during the training week. This pre-selected her for the untrained group and therefore Subject 109 for the trained group. Other subjects were assigned to each of the groups in order that the total time taken in each group was as close as possible. Figure 20 shows the means of the first two experimental sessions for each selected group. The mean time was 25 for each group. By way of verification, the sum total time taken for all of the first two
Pilot investigation

sessions for all subjects in each group was calculated and found to be within eight minutes of each other: 403 minutes for the trained group and 395 for the untrained group (the trained group taking, on average, 30 seconds longer per 30 minute session than the untrained group). This was as good as could be expected, given the inter-subject variability.

Figure 20: Bar graph showing mean time taken for the first two experimental sessions for the candidates selected to be in the trained and untrained groups.

![Error Bars show 95.0% CI of Mean Bars show Means](image)

Figure 21 shows the final assignments of the subjects into trained and untrained groups, and their relative performances.

Figure 21: Scatter plots showing performance of individual subjects (RMS Error against Partial Eta Squared) in the 4 main subgroups of subjects (By Course, and by Gender).

![Training](image)

Data points are labelled by subject number. 'Training' indicates which group the subjects were placed into.
Pilot investigation

It was also possible to calculate the mean RMS error (mean 'error') and mean Partial Eta Squared values for the trained and untrained groups in order to verify that the average performance of the groups was similar. As can be seen in Figure 22 and Figure 23, the mean performance of each group is not statistically different and is in fact very similar.

**Figure 22:** Bar graphs showing the initial mean 'error' of trained and untrained groups.

![Bar graphs showing mean 'error' for trained and untrained groups](image)

ERROR Bars show 95.0% CI of Mean
Bars show Means

**Figure 23:** Bar graphs showing the initial mean Partial Eta Squared values of trained and untrained groups.

![Bar graphs showing mean Partial Eta Squared for trained and untrained groups](image)

ERROR Bars show 95.0% CI of Mean
Bars show Means

**4.3.4 Summary**

Sixteen motivated but naïve subjects were recruited from a population of two first year undergraduate courses (Tonmeister, and Music & CSD). In order to gauge the
effect that a spatial audio training programme would have upon their performance in 'real-world' subjective tests, it was decided to split them into a group that would receive additional training and one that would not. To ensure that the groups had similar performance levels, all subjects were asked to participate in two experimental sessions where they assigned grades to the stimuli selected in Section 4.2.3 on a 0-100 point scale of width. Analysis of the consistency and sensitivity of these grades and the time in which they completed the nominally 30-minute sessions was taken into account in order to divide the subjects into two similar groups. One of these groups would participate in an additional training programme during the following week. Both groups would then take part in two further experimental sessions during the week following the training week.

4.4 Naïve Listener Training Programme

In Section 4.3.3, eight naïve but motivated subjects had been selected to participate in an additional training programme designed to enhance their perception of the scene-component width of individual source scene components.

According to Alessi and Trollip's phases of instruction (Alessi and Trollip 2001), effective and efficient learning is facilitated by:

- Presenting information
- Guiding the learner
- Practising
- Assessing learning

It was therefore decided to cover these four phases in the listener training programme.

Neher (2004) has provided a series of unidimensionally varying stimulus sets designed to demonstrate changes in single perceptual attributes. One of which he termed 'Source Width'. Moreover, Neher conducted a preliminary study where he investigated the effectiveness of a training programme that was devised to train subjects to correctly rank nine stimuli in terms of their 'Source Distance' attribute and five other stimuli in terms of their 'Source Width' attribute.

Neher recruited five subjects whom he paid to take part in the training programme. Three subjects were trained and two acted as a control group, receiving no additional instruction (although the method used to select which subjects would be trained and not trained is not reported). Training was separated into source distance training and source width training.

According to Neher's report, all five subjects took part in a test before the training phase, and once again thereafter. The task was to correctly rank the stimuli provided in terms of either source distance or source width appropriately. Whilst it is not clear in Neher's report whether or not the two untrained listeners took part in any form of familiarisation with the stimuli between his pre- and post-tests, it is assumed by this author that Neher's untrained listeners did not take part in anything other than the two ranking exercises.
Pilot investigation

To summarise his method, the training programme consisted of a familiarisation phase followed by three training phases: discrimination, pairwise ranking and multi-stimulus ranking. All phases were automated using Max/MSP patches. It appears that Neher interacted with the subjects during the familiarisation phase to “introduce the listeners to the physical principles governing the perception of each attribute so as to provide them with a firm background in the underlying modality” (Neher 2004). See Section 2.2.3.4 for further details.

This author obtained a set of Neher’s Max/MSP patches and two Source Width stimulus sets, and used them (with permission) as the basis for the pilot training programme.

The training programme implemented by Neher was essentially a set of increasingly difficult drills to be completed by each subject with an assessment phase consisting of the pre- and post-training tests. It would therefore cover the “practising” and “assessing learning” phases of instruction described by Alessi and Trollip (2001).

In order to cover the other phases, namely “Presenting information” and “Guiding the learner”, it was felt necessary by this author to incorporate an additional tutorial phase before the drills. This was in order to put the concept of the width of auditory sources and the drills into context, thereby enhancing motivation in the successful completion of the drills.

It is anticipated that any tutorial sections of future spatial audio listener training systems would be implemented within a self-administered computer software application. This would allow the training system to be distributed in a form that would enable listeners to train themselves without the need for a human instructor to be present. Due to the exploratory nature of the pilot research, however, the tutorial section was administered by the author using a Microsoft PowerPoint presentation and additional audio examples taken from the output of (Neher 2004). Print-outs of the 30 PowerPoint slides used in the tutorial are found in Appendix 7.2.

After the initial tutorial session, the subjects that were selected in Section 4.3.3 were asked to complete the “Source Width” drills undertaken by trained subjects in (Neher 2004). Prior to commencement of the tutorial, and upon completion of the training drills, the subjects were asked to participate in similar pre/post trials to the one Neher had given his subjects. This was in order to cover the “assessing learning” phase of instruction (Alessi and Trollip 2001). A determination of the accuracy with which the subjects ranked the stimuli with respect to source width in the pre and post trials was examined in order to investigate whether the subjects were any better at ranking the stimuli after successful completion of in the tutorial and training drills.

Each training session took place during the training week and lasted no longer than 30 minutes, which is considered to be the maximum amount of time suitable for drills (Alessi and Trollip 2001). Subjects agreed an initial training session time then scheduled each subsequent training session as their timetable permitted.

After the post-training test was completed, a questionnaire and feedback form was completed by each subject to gauge their own assessment of their performance and to elicit their feedback as to how the training programme could be improved.
4.4.1 Method

The training programme took place in Studio 3 at the University of Surrey’s Department of Music and Sound Recording. The room is equipped with a Sony OXF-R3 digital mixing console connected to five active loudspeakers (ATC SCM100As), at a distance of 2.3m from the listening position.

The training programme was administered via two computers: a portable PC that was used to display the Tutorial slides via an external monitor; and a DAW running Max/MSP v4.5.2 from Cycling ’74. The audio outputs of the DAW were connected to the mixing console via a digital audio interface. This allowed for the five loudspeakers to be driven by the five digital audio outputs of the Max/MSP patches.

The monitoring controls of the desk were adjusted to be of suitable level to hear salient width changes in the audio stimuli, whilst not being too loud to cause discomfort, as judged by the author. No level attenuation was introduced by the input stages of the desk, and the monitoring section was kept at 70 dB(SPL), as read from the console’s monitoring display. This was in order to set a standard level that could be easily repeated to reduce the chance of differences occurring between the training sessions.

The mixing console was set-up so that there were no visual indicators active and visible during the entire training programme. This was done to stop subjects from becoming distracted or relying on information gleaned in the number of channels and the level of each channel active during presentation of the various stimuli.

Subjects took part in the training programme individually, and sat on a stool with no castors at the listening position. They could control the training programme computer (running Max/MSP) using a USB keyboard and mouse. The keyboard and mouse (and subsequently a flat screen visual display panel) were placed upon a wooden plank suspended above the mixing console faders on two tubular metal bars. This allowed the subjects to remain facing forwards at the listening position whilst controlling the training drills and certain stages of the tutorial. It also protected the mixing console from being disturbed by the subjects accidentally.

Figure 24 shows the experimental set-up during the main training sessions. The computer screen, computer keyboard and mouse can be seen on the mixing console. The portable PC can be seen in the background, although it was only utilised during the tutorial phase of the training programme.

During the pre/post tests, subjects were presented with a screen showing five buttons and five sliders. The sliders had five available positions, 1-5. Subjects were instructed to click on each of the buttons to audition the five stimuli, and use the corresponding slider to indicate its rank order in terms of width (using each available rank just once). These were in fact the five source width stimuli provided by Neher. The time taken for the task was recorded within the Max/MSP patch, as was the rank order in which the subjects put the stimuli. The stimuli were randomised for each subject and pre/post session by the author from the example patch provided by Neher. The Max/MSP patches were also edited to change the term ‘source width’ used in Neher’s patch to ‘width’, which was consistent with the rest of this investigation.
Subjects sat in the listening position and viewed the flat-screen (which was rotated so that it could be viewed by the subjects). The video output from the laptop was mirrored to the flat panel for most of the tutorial. During the ‘audio examples’ phase (see Section 4.4.2), the flat panel was rotated to face the author and the input switched so that it displayed the output from the test computer to allow playback to be set-up from it. Figure 25 shows one of the trainees auditioning stimuli during the audio examples phase of the tutorial.
4.4.2 The Tutorial Phase

The tutorial phase of the training programme was implemented in Microsoft PowerPoint. Whilst it is expected that future training programmes would eventually use a tutorial implemented within a self-administered application, PowerPoint provided the author with a quick and easy method to produce a tutorial presentation that would not only allow the relevant material to be covered, but to allow interaction between the author and the subject to allow any misunderstandings to be rectified immediately. It was, after all, a learning process for the author as well as for the subjects.

In order to "present information" and "guide the learner" (Alessi and Trollip 2001), subjects were presented with a series of slides and a verbal explanation by the author. At various points throughout the tutorial, subject feedback was elicited. Print-outs of the 30 PowerPoint slides used in the tutorial are found in Appendix 7.2. A summary of the tutorial procedure given to every subject is described here. References to the various slide numbers in the tutorial refer to the appropriately numbered slides shown in Appendix 7.2.

Important in any tutorial is a statement of the objectives of the course as it enhances both learning and learner satisfaction (Alessi and Trollip 2001). Directly following the title page (Slide 1) is a statement of the objectives (Slide 2) of the course in non-behavioural form (Alessi and Trollip 2001).

Slide 3 through Slide 6 stimulate prior knowledge (Alessi and Trollip 2001), or present new knowledge. The subjects were asked what the 'surround sound' reproduction set-up added to conventional 2-channel stereo. This was revealed in Slide 4 along with a re-iteration that surround sound not only adds two surround channels, but also a frontal centre loudspeaker, enhancing the frontal arc as well as the soundfield surrounding the listener.

Prior to the presentation of Slide 5 then Slide 6, subjects were asked to provide examples of situations where the quality of multichannel audio would need to be evaluated, and how these systems could be evaluated. Subjects were prompted with an example to encourage them to think of their own before the stated examples were revealed. This was in order to demonstrate relevance, which according to Keller's ARCS theory, is an important factor in creating motivation (see Section 2.3.2.2). The examples provided were selected to show the subjects that what they would learn would be of use to them. The item 'sound processing algorithms' was included especially so that the Music & CSD students had something that was directly relevant for them.

Slide 6, Slide 7 and Slide 8 establish the need for a spatial audio evaluation system, and describe the concept of spatial audio attributes, and the need for a universal language to describe them. During Slide 8, subjects were asked to postulate why it would be a problem that no universal language exists to describe spatial audio attributes.

Slide 9 through Slide 13 explain the simplified scene-based paradigm used in this study. Subjects were asked to provide attributes to describe the sound source in Slide 11 (without the identified attributes and arrows showing). The attributes were revealed in a logical order: direction, distance, width and depth. Environmental
attributes were described in Slide 12, but the subjects were told that the environmental attributes would not form part of the current study.

According to Alessi and Trollip (2001), conceptual information should be taught by providing examples of simple instances and non-instances, progressing to difficult instances and non-instances. Slide 14 through Slide 26 demonstrated instances of certain spatial audio attributes (and therefore non-instances of others). The slides show two versions (labelled A and B) of the subject (the plan view of the head), in an environment (the white box) with a sound source (the shaded object). Between A and B the sound source changes, and the question asked of the trainee subjects was ‘what has changed between A and B?’ Answers were elicited from the subjects in terms of the attributes outlined in Slide 13. Upon successful answers, the name of the attribute and any arrows were revealed on the slide, confirming in which way the sound source had changed. For example, Slide 15 showed a unidimensional change in direction of the sound source (with width, depth and direction remaining constant). The slides progressed in complexity, showing changes in two of the attributes in Slide 19 through Slide 24, three in Slide 25 and four changes in Slide 26.

Because of the benefits of dual-coding (Clark and Paivio 1991), it would have been advantageous to incorporate audio playback with the visual representations. However, as PowerPoint has limited audio capabilities, this was not possible. Instead, audio examples were used directly after the visual analogy examples (Slide 27) to reinforce the concepts of the spatial audio attribute changes.

There was a short pause in the tutorial where the author rotated the flat-panel away from the view of the subject and switched the video input to display the output of the DAW. The author then loaded some of the other unidimensional stimulus sets provided by Neher (source distance; source width; ensemble width and ensemble depth), in order to demonstrate spatial audio attribute changes using audio examples.

Because learner control is also supposed to be an important factor in successful learning (see Section 2.3.2) as is active participation in the task (Alessi and Trollip 2001), the subject was given control of the playback. A computer keyboard was placed in front of the subject, and they were asked to use keys 1-9 to audition the range of values of the attribute in question. This stage is important, because it allows the trainees to actively apply the SSBP to audio examples.

Firstly, the source distance stimuli were loaded and the subjects were asked to say what was changing to the source, the correct answer being a change in distance.

Secondly, the source width stimuli were loaded and the subjects were asked to say what was changing to the source, the correct answer being a change in width.

Thirdly, the ensemble width stimuli were loaded. This consisted of a number of different voices that were positioned in a spread that got wider and wider about the centre speaker. Subjects were asked to consider that the voices formed a 'choir' and they should consider the choir to be a single entity. Subjects were then asked to say what was changing to the choir, the correct answer being a change in width. Further to that, they were asked to concentrate on a specific female voice within the choir (one that changed from a central position at one extreme to being positioned to the far left at the other extreme). Subjects were then asked to say what was changing to the female voice, the correct answer being a change in source direction.
Finally, the ensemble depth stimuli were loaded. This consisted of a string quartet, where the outer instruments (violin 2 and viola) appeared to get closer to the listener whilst the inner instruments (violin 1 and cello) appeared to get further away from the listener. Subjects were first asked to say what was changing with the cello, then with one of the outer instruments, the correct answer being a change in source distance (and potentially source direction). Subjects were then asked to say what was changing to the quartet as a whole, the correct answer being a change in depth.

Once the subjects had listened and responded to the audio examples, Slide 28, Slide 29 and Slide 30 were displayed, outlining the rest of the training programme, and demonstrating its relevance to the subjects.

4.4.3 The Drill Phase

The practice drills involved various combinations of five width levels of two stimuli: a cornet and a guitar, as validated in (Neher 2004). The Max/MSP patches for the familiarisation and drill phases were based upon the examples provided by Neher.

As a part of the drills phase, the subjects were presented with a familiarisation page allowing them to listen to the narrowest and widest of the cornet and guitar stimuli. As they auditioned the sounds, a visual representation of the sound that they were listening to was also displayed on screen, either to signify the narrowest stimulus (Figure 26), or the widest stimulus (Figure 27). This dual-coding approach (Clark and Paivio 1991) was expected to reinforce the concepts introduced and practiced in the visual analogy and audio examples sections of the tutorial.

Subjects were given as much time as they needed to familiarise themselves with the widest and narrowest cornet and guitar stimuli before beginning with the first drill – that of pairwise discrimination. An additional period of familiarisation (involving the widest, narrowest and middle stimuli) was also presented before the second drill – that of pairwise discrimination between three stimuli.
The drill phases ran more-or-less along the same lines as described in Appendix E of (Neher 2004). Subjects were presented with increasingly difficult sets of drills which would need to be successfully completed (by getting 80% correct) in order to progress through the programme. Once they had successfully completed all levels of difficulty of the drills, they were deemed to have been ‘trained’ and progressed to the post-training test.

Four complete sets of drills were created for each subject, with presentation orders randomised accordingly. Each drill type had two versions, one utilising the guitar stimuli and one utilising the cornet stimuli (the multi-stimulus ranking drill also had versions where the presentation of guitar and cornet pages was interspersed). All drills had ten pages (except the multi-stimulus ranking drill which had five). Successful completion involved providing the correct answer on eight of the ten pages (ranking the stimuli correctly on four of the five pages for the multi-stimulus ranking drills). The drills, in order of difficulty were as shown in Table 19:

The drill difficulty levels noted in Table 19 refer to the various types of programme items – mostly either cornet (nominally easier) or guitar (nominally more difficult). The multi-stimulus ranking drill has 3 levels: cornet, cornet and guitar mix and guitar. The cornet and guitar mix had alternating cornet and guitar stimuli pages.

<table>
<thead>
<tr>
<th>Drill Difficulty Level</th>
<th>Drill Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>Pairwise discrimination</td>
<td>Subjects were asked whether two stimuli were the same or different. The stimuli were drawn from the narrowest and widest of the five available source width stimuli.</td>
</tr>
<tr>
<td>3-4</td>
<td>Pairwise discrimination</td>
<td>Subjects were asked whether two stimuli were the same or different. The stimuli were drawn from the narrowest, the mid-widest and the widest of the five available source width stimuli.</td>
</tr>
<tr>
<td>5-6</td>
<td>Pairwise ranking</td>
<td>Subjects were asked which of two stimuli were the widest. The stimuli were drawn from the narrowest and widest of the five available source width stimuli.</td>
</tr>
<tr>
<td>7-8</td>
<td>Pairwise ranking</td>
<td>Subjects were asked which of two stimuli were the widest. The stimuli were drawn from the narrowest, the mid-widest and the widest of the five available source width stimuli.</td>
</tr>
<tr>
<td>9-10</td>
<td>Pairwise ranking</td>
<td>Subjects were asked which of two stimuli were the widest. The stimuli were drawn from all five of the available source width stimuli.</td>
</tr>
<tr>
<td>11-13</td>
<td>Multi-stimulus ranking</td>
<td>Subjects were asked to rank three stimuli (narrowest, mid-widest and widest) in order of width.</td>
</tr>
</tbody>
</table>

Progress throughout the drills was monitored via the ‘performance window’, shown in Figure 28. The performance window charted the subject’s progress through each drill, providing a running total of the right and wrong answers, and how many trials were remaining in the current drill.
As subjects achieved correct answers, their right answer tally increased and they were rewarded by seeing a cartoon character and hearing a positive message. If they answered incorrectly their wrong answer tally increased and they saw a cartoon character and heard a sound that informed them that the answer was incorrect. During the pairwise and multi-stimulus ranking drills, subjects who answered incorrectly were allowed a second chance to get the answer right. In the pairwise ranking drills, this allowed to the subject to listen again to the difference or similarity between the two stimuli so that they could learn from their mistake, no adjustment in the incorrect score was awarded. In the multi-stimulus ranking tests the subjects were given another chance to get the answer right after an incorrect attempt. If a subject gave four wrong answers, the trial would terminate and the drill would restart. Upon restart, the author would normally intervene to select a similar trial on the same level or an easier trial, in negotiation with the subject to find what they felt would be best for them.

A degree of control (see Section 2.3.2) was afforded to the subjects, to allow them to choose their path through the drills in consultation with the author.

Once each subject had progressed to and completed drills in the most difficult level, they were asked to complete the post-training multi-stimulus ranking task. This involved correctly ranking all five guitar stimuli. The subjects were given one attempt to do this and a record of the time taken for them to complete the task was kept.

4.4.4 Drill Performance

The subjects' performance through the drill activities outlined in Section 4.4.3 is discussed in this section.

Although Neher specified that multi-stimulus ranking using three stimuli was a more difficult task than pairwise ranking using five stimuli, this author was not convinced. This is because it appeared at least as difficult to perceive differences between adjacent stimuli in the pairwise task as to complete the multi-stimulus ranking task (where subjects were forced into assigning three separate width levels to three different sounds). For purposes of allowing completion of the trials, pairwise
Pilot investigation

ranking of five stimuli and multi-stimulus ranking of three stimuli were considered to be the most advanced tasks. The actual order in which the most advanced tasks was agreed on a subject-by-subject basis.

The time taken to progress through the drills varied from learner to learner, as can be expected. The shortest time taken was by Subject 113 who needed just three 30 min. sessions to complete the most advanced tasks. The subject who took longest was Subject 109, who needed eleven 30 min. sessions to complete the training tasks. The average number of sessions per subject was calculated as 6.25, meaning that the advice for future subjects could be that a similar training programme should take no more than seven training sessions (plus one for the tutorial).

Figure 29 charts each of the training subjects' progress through the drills in each training session. The 'difficulty' values denote the highest level drill completed in each session. During any given session higher tasks may have been unsuccessfully attempted, potentially resulting in lower level tasks being reiterated to boost confidence. The dotted reference line at difficulty '9' is the notional target. Tasks on or above this line are the most advanced ones, and successful completion of tasks at this level allows progression of the subject to the post-training ranking task. Difficulty level is not a linear scale, so these graphs should be taken as a performance indicator only.

Note that the apparent dip in 'performance' during the ultimate or penultimate session in subjects 104, 108 and 115 is due to them attempting and completing the multi-stimulus ranking task (difficulty 11, 12 and 13) before the 5-stimulus pairwise ranking task (difficulty 9 and 10).

Subject 109 took a lot longer than the other subjects to complete the drills, taking eleven 30 min. sessions in all (all other subjects were finished within seven 30 min. sessions). It is worth noting that Subject 109 successfully completed one of the
Pilot investigation

advanced drills within the fourth session. This was the multi-stimulus ranking task involving three cornet stimuli. During the drills, the author had inadvertently progressed Subject 109 directly from a difficulty ‘4’ task – pairwise detection involving 3 guitar stimuli, to the level 13 task – multi-stimulus ranking of 3 cornet stimuli. The subject completed this task flawlessly, although they took 20 minutes to complete it. Because Subject 109 seemed to be confident with the cornet stimuli, the next four sessions were spent attempting the lower difficulty guitar stimuli without much success. The remaining sessions were spent building confidence with cornet stimuli and attempting multi-stimulus ranking tasks involving guitar stimuli. During the final session, the subject had achieved a 100% success rate during the highest difficulty task and was then advanced to the final ranking test.

As will be seen in the next section, Subject 109 had actually improved markedly in ranking performance from the pre-training test to the post-training test. However, as indicated in feedback received via the post-training questionnaire (see Section 4.4.6), and from their haphazard performance during the drills (as seen in Figure 29), it is clear that Subject 109 would have benefited, and no doubt performed better in the drills if the drills had been attempted in order of difficulty. Care should be taken not to allow experimenter error to confuse the order of presentation to that extent in future experiments. Using a more automated method of presentation of drills (incorporating learner control) is expected to help.

Notwithstanding, the training drills were successfully completed by all subjects. Section 4.4.5 will analyse the ranking performance in the pre-training and post-training ranking tasks, and Section 4.4.6 will discuss the feedback received from the subjects by way of a questionnaire which they filled out immediately after they completed the training programme.

4.4.5 Pre/Post Test Ranking Performance

The pre- and post-training test involved the ranking of the five provided stimuli in terms of their width. Because there was an established ‘correct’ order for the stimuli, validated in (Neher 2004), it was possible to compare the subjects’ judged rank orders with the established order to which they should be assigned.

The method used in (Neher 2004) is to find the sum of the squared of the Euclidean distances between the correct rank order and those provided by the subjects. For this, the difference in the rank order number between the expected and subjective results was calculated (called the Euclidean distance). The Euclidean distance was then squared (to make any differences occur in magnitude only, not direction), and then summed across the five available stimuli. This gives a Sum of the Squared Euclidean Distances (SSED) for the particular rank order provided by the subject. Note that Neher referred to this as “SED” rather than SSED. It is unclear to this author why this was the case.

The size of the SSED (up to a maximum of 40 for five ranks) indicates how dissimilar the rank order provided by each subject was. An SSED of 0 indicates a perfect match between the correct and provided rank orders. Figure 30 shows the mean SSED for the trained subjects during the pre-test and post-test. As can be seen by the sizable reduction in mean SSED – from about 25 to about 3 – subjects had dramatically increased in their ability to correctly rank the source width of the five guitar stimuli.
Figure 30: Bar graph showing the mean SSED for *all* trained subjects, for the pre- and post-training tests.

![Bar graph showing the mean SSED for all trained subjects](image)

The maximum SSED possible was 40.

Figure 31 shows how this varied by individual subject. As can be seen from the graph, subjects 108 and 109 achieved a perfect SSED in the post-training trials. All subjects increased their SSED between the pre- and post-trials.

Figure 31: Bar graph showing the SSED of each trained subject, for the pre- and post-training tests.

![Bar graph showing the SSED of each trained subject](image)

The maximum SSED possible was 40.

It would have been useful to include an additional control group in the pre- and post-training trials (but not the drills). This would have the beneficial advantage of showing what effect practice alone would have upon these SSED figures. However,
Pilot investigation

it was decided not to include the untrained subjects in any form of additional training so as not to contaminate the overall training effect observable in the experimental ‘real-world’ grading sessions.

Figure 32 shows the mean time taken for the subjects to complete the pre- and post-training ranking task.

**Figure 32:** Bar graph showing the mean time taken (in seconds) for all trained subjects, for the pre- and post-training tests.

As can be seen in Figure 32, there was no significant change in the time taken to complete the task by the trained listeners. An investigation of the individual subject’s time taken in the task did not reveal any interesting trends.

It is suggested by this author that any fluency increase as a result of familiarisation with the task was being confounded by an increase in the attention being paid to the correct identification and ranking of the widths.

### 4.4.6 Questionnaire Analysis

Immediately following completion of the training programme, subjects were asked to complete a post-training questionnaire – mostly responding using a five-point ordinal scale (1=Strongly disagree; 2=Disagree; 3=Neither agree nor disagree; 4=Agree; 5=Strongly agree). The main purpose of the questionnaire was to obtain informal feedback from the subjects as to their opinions on how to improve the training programmes, as well as an assessment of its effectiveness. A couple of understanding questions were also included to test the recall of some of the themes discussed in the tutorial. The questions and their responses will be discussed below.

"I found the tutorial session (the first session) interesting"

Most subjects agreed or strongly agreed with this statement. One was ambivalent. The tutorial phase had therefore been successful in engaging the majority of trainees in the concept of spatial audio attributes and their description analysis.
"I found the tutorial session (the first session) useful"
All subjects agreed or strongly agreed with this statement (one strongly agreed). The tutorial was therefore seen by the subjects as being not only interesting, but beneficial.

"Concepts were well explained"
All subjects agreed or strongly agreed with this statement (two strongly agreed). Whilst it is not possible to tell from the response to this statement that subjects had understood the concepts correctly, it is possible to say that they were left with little or no ambiguity as to the concepts in their own understanding. The tutorial can therefore be assumed to have not confused the subjects.

"The tutorial session increased my motivation to attend the practice drills"
Six subjects were ambivalent to this statement, two agreed with it. With hindsight, and from one subject's feedback, this statement was badly worded. It was possible for subjects to be motivated to do the drills, and ambivalent to this statement (because they were already motivated, the tutorial did not increase or decrease their motivation levels). From the feedback to the previous questions, it can be assumed that subjects were motivated to take part in the training programme itself, even if the tutorial had not increased their motivation to take part in the remainder.

"Visual analogies helped me to understand the concepts (spatial audio attributes)"
All subjects agreed or strongly agreed with this statement (five strongly agreed). This was the highest proportion of strong agreements in response to any of the statements in the questionnaire. Visual analogies were therefore successful in aiding understanding and should be incorporated into future versions of the training programme.

"Audio examples helped me to understand the concepts (spatial audio attributes)"
Most of the subjects agreed or strongly agreed with this statement. One was ambivalent. Audio examples did not provide such a strongly positive response from the subjects, although they all agreed that they did not hinder understanding. It is expected that dual-coding (Clark and Paivio 1991) (the combination of audio and visual stimuli) in a future training programme will improve the reception of the audio examples.

"The tutorial helped to put the following practice sessions into context"
Five subjects agreed or strongly agreed with this statement, three were ambivalent towards it. The fact that no subjects disagreed with the statement can be taken along with other evidence discussed above to indicate that the tutorial was successful.

"I found the practice sessions interesting"
Most subjects agreed or strongly agreed with this statement, two were ambivalent and one disagreed. This author may have expected there to have been more subjects who thought that the drills were not "interesting", however this was not the case and the drills were generally well received. Not allowing drill sessions to last more than 30 minutes (Alessi and Trollip 2001) can be seen as partly helping to maintain
Pilot investigation

interest in the drills, as can use of visual and auditory feedback (cartoon characters and voices).

"I found the practice sessions useful"

Most subjects agreed or strongly agreed with this statement (one strongly agreed), one was ambivalent. Even though they were not found universally interesting, all-but-one of the subjects found the drills useful. Alessi and Trollip (2001) suggest that whilst drills may have their detractors, many learners (especially adults — in this case young adults) acknowledge that drills are what they need to achieve quick mastery of specific skills.

"Overall, the practice sessions were (Too easy – Easy – Just right – Difficult – Too difficult)"

Seven subjects found the drills to be “just right”, one found them “difficult”. This is strong evidence that learner control (see Section 2.3.2) was successful in helping to pace the trials to the appropriate difficulty level for each subject. The subject who stated that the trials were “difficult” was Subject 109, most likely because of the particulars of their specific drill session order, as described in Section 4.4.4.

"I had enough time to complete the exercises"

Subject 109 disagreed with this statement (see Section 4.4.4), all other subjects were either ambivalent (2 subjects), agreed (3 subjects) or strongly agreed (2 subjects). Even though the training programme took place in the relatively confined timeframe of a single week, the majority of subjects felt that they had been given adequate time to complete the training programme to their satisfaction.

"How do you think you have benefited from the training?"

In response to this question, most subjects indicated that they believed that their own listening skills had improved. Half of the subjects cited an increase in the ability to detect differences in the width of sound sources. Other comments received related to having a greater understanding of spatial audio and being able to use the paradigm to describe perceived auditory changes.

"My spatial audio listening skills have improved as a result of the training"

All subjects agreed with this statement (one strongly agreed). This unanimous endorsement of the training scheme is further evidence that subjects had grown in confidence through participating in the programme.

"My listening skills in general have improved as a result of the training"

Half of the subjects agreed with this statement, three were ambivalent, with one subject neglecting to answer this question. It is not clear if subjects were expected to detect an improvement in their general listening skills or not. There was no indication that subjects had found the concentrated training in width perception to be detrimental to their general listening, and indeed many had believed it to be helpful.

"This week’s training would have helped me with the tests I did last week in TB7"

This question was aimed at finding out whether subjects believed that the training scheme would have been useful in the ‘real-world’ task. (‘TB7’ is the room designation for the listening room at the University of Surrey)
Pilot investigation

All subjects agreed with this statement (two strongly agreed). This unanimous endorsement of the training scheme is further evidence that subjects had grown in confidence through participating in the programme. Whether or not this translated to actual improvements, however, will be discovered in the analysis of the post-training experimental sessions found in Section 4.5.

"Why learn about a spatial audio attribute description scheme?"

This question was included as a method to test recall of the information presented during the tutorial phase. Seven of the eight trained subjects responded by talking about the need for a universal language when describing spatial audio attributes between different people. The eighth subject (a Tonmeister student) mentioned that it would benefit him personally in his later studies.

"Why gain practice at detecting and rating spatial audio attributes?"

The expected answer from the tutorial information (see Slide 30 in Appendix 7.2) would be that it would improve their abilities and increase confidence and fluency in undertaking spatial audio listening tasks. The actual responses to the questions varied quite widely. On the whole, the Music & CSD students talked in terms of improved perceptual abilities (one also mentioned quicker response times). The Tonmeisters tended to talk in terms of the benefits to them in real-life situations (such as audio mixing sessions). This apparent difference between the responses from the subjects from different courses could indicate a difference in the perceived relevance of the training programme to the two groups.

"How could the training be improved?"

Subjects were told that responses to this question were the most important, as they would shape future training systems. Every subject found something to include in this section, the greatest consensus of opinion was regarding the stimuli used in the training system. Three subjects called for a greater variety of sound samples from the two that were used (guitar and trumpet). Subject 109 stated again that they would have liked to attempt all levels of difficulty with the cornet samples before progressing to the guitar samples. Another subject criticised the guitar sample as having a widely varying dynamic range, stating that it was difficult to separate changes in level with changes in width when switching between different stimuli.

Other comments related to the information available. One subject called for details of the terms used in the paradigm to be provided in a way that they could take away from the test room. This could be implemented as an information sheet or downloadable computer application. Another subject asked for additional help in what to listen for when the differences between stimuli was not obvious. This latter point is potentially difficult to implement, as it has been thus-far assumed that the subjects would make their own sense of the differences between the audio stimuli, rather than relying on any help from the experimenter.

The final grouping of comments related to the training programme implementation. One subject wanted to have access to the familiarisation stimuli before every session. This could easily be accommodated in future training system implementations. Another subject noted that the samples should loop (which was either not implemented or not functional in the Max/MSP patches provided by Neher). Again, this should be implemented in the next training investigation as it will reduce the distraction of having to continually click to restart audio reproduction. One subject
also called for the ability to listen back to erroneous answers in every drill variety, rather than just the pairwise and multi-stimulus ranking tasks. This could be implemented in the next training system to allow immediate feedback and corresponding learning to occur (Alessi and Trollip 2001). In addition, one subject said that he would have preferred regular scheduled times to be provided at the first meeting, rather than the system employed where subjects would schedule their following session upon completion of the session they were taking part in. Each subject should be given the option of scheduling regular times, or arranging each session during the previous session. Based upon the results presented in Section 4.4.4, a more accurate description of the length of the training system would be possible, and should be welcomed by future trainees.

It is also worth noting that three subjects took the time to provide positive comments about how the training system was “well done”, “overall very good” and that the “gradual steps through the levels of difficulty were good”, even though they had just been asked to describe how the system could be improved. This can be taken together with the other positive feedback provided in the questionnaire responses to conclude that the experience had, on the whole, been a positive one for the majority of the subjects. The dramatic reduction in the SSED in rank-ordering of the stimuli shown by all subjects and the group as a whole described in Section 4.4.5 is evidence of the actual improvement of their performance during the training period.

4.4.7 Summary

A listener training programme based upon a modified version of that implemented in (Neher 2004) was used to train eight motivated subjects in the detection, discrimination and ranking of stimuli that varied in terms of the perceived width of a sound source.

Analysis of the SSED (a measure of the wrongness of the rank order provided by the subjects) has revealed that, over the course of the training week, all subjects improved in their ability to rank the five controlled stimuli.

The increases shown by the individual subjects was far in excess of that shown by the subjects in (Neher 2004). The potential source for this difference could be that Neher’s subjects were paid to participate and thus probably did not do so out of intrinsic motivation (see Section 2.3.2.1), as was the case with the subjects used in this study. Another difference is the use of a tutorial session to put the learning into context and demonstrate simple and more complex instances of the concepts involved. This was in contrast to Neher’s methodology that, although it includes a mention about an explanation of terms for the subjects, did not use a structured tutorial before the drills.

On the basis of the improvement in SSED between the pre- and post-training ranking task, the training scheme has been judged to have been successful for all trained subjects.

It is noted that, even though the subjects improved in their ability to rank order the stimuli, they were not shown to change significantly in the time they took to complete the task. This suggests that estimations of fluency for such tests are not appropriate as indicators of successful learning.
An analysis of the feedback left by subjects about the training programme, and the author’s own observations throughout the experiment has allowed for certain improvements to the training programme to be suggested.

The tutorial phase appeared to be very well received. It was perceived by subjects to be both interesting and beneficial. They also appeared to not be confused by the presentation. The use of visual analogies was successful in engaging the training subjects and was perceived to be successful in demonstrating the concepts in the tutorial. Dual-coding (Clark and Paivio 1991), i.e. auditory and visual presentation would have been an improvement to the audio-only examples used in the tutorial.

Maintenance of a 30-minute time-limit for the drills, and the use of cartoon characters in the feedback can be seen as contributing factors to the generally positive reception of the drill phase. Regarding the cartoon characters, the author believes the feedback for incorrect answers implemented by Neher in the pilot study drills was too harsh. It would have been useful and interesting to ask the subjects their opinion of the feedback. A more neutral and/or encouraging feedback system should be used for incorrect answers in future.

Learner control (see Section 2.3.2) of the drill presentation order is likely to be the main reason why the drills were found to be of ideal difficulty level. Subjects had more-or-less decided to progress in difficulty level when they were confident to move on. This intrinsic self-pacing has manifested itself in a satisfying learner experience in the majority of subjects. The only subject to find the drills “difficult” was Subject 109 who did not have as much control over the presentation as the other subjects, and expressed as much in their feedback comments. Self-pacing should therefore be incorporated in the drill phases of future training schemes.

Differences in the perceived relevance of the training system between subjects on the two courses was indicated by the Tonmeister students’ readiness to fantasise about the ways in which the training would help them in real-life situations (see Section 2.3.2.1). This apparent difference in the perceived relevance of the training system between subjects on the different courses may have further ramifications.

Other specific improvements for the next training system implementation can be summarised as:

- An increased variety of training stimuli.
- Additional materials (such as an information sheet) to be given out...
- User-controlled access to the familiarisation stimuli before every drill.
- Incorporate monitoring of whether subjects require ‘additional help’ to that provided.
- Looped playback of samples.
- The ability to listen back to incorrect drill items in order to learn from mistakes during every trial. (This could be implemented by using a new ‘neutral’ feedback for incorrect answers with a chance to listen again.)
- Subjects would be given the option of planning out training session timetables or arranging them on a session-by-session basis.
4.5 Subjective Attribute Rating: ‘Post-Training’

In order to ascertain the effect that the training programme described in Section 4.4 had upon performance in the ‘real-world’ subjective task, all sixteen subjects were invited to take part in another two experimental sessions (the ‘post-training’ sessions) similar to those undertaken in Section 4.3 (the ‘pre-training’ sessions). Any performance differences found between the way in which the additionally trained and untrained subjects graded the stimuli during the ‘pre-training’ sessions and the ‘post-training’ sessions would therefore be as a result of having participated or not participated in the intervening training programme.

As the training programme for SC Width (for individual sources) described in Section 4.4 had been deemed to be successful, it was expected that this would transfer to the ‘real world’ task, allowing the trained group to show an improvement in performance greater than that shown by the untrained group.

As Bech had found, repetition alone was found to be beneficial to performance in subjective testing (see Section 2.2.1.3). This suggests that the ‘untrained’ group should also improve in the ‘post-training’ sessions versus the ‘pre-training’ sessions, even though they had not taken part in the additional training programme.

Because of the inclusion of a control group (see Section 2.2.4), the difference in improvement between the trained (experimental) and untrained (control) groups will therefore be a measurement of how the training system affected the subjective preference.

4.5.1 Method

Two experimental sessions were scheduled for each listener. Where possible, these took place at the same times and days during the week as the initial two experimental grading sessions had taken place two weeks previously for each subject. Control was therefore achieved for the time interval between the ‘pre-training’ and ‘post-training’ sessions, and between the two sessions within the ‘pre-training’ session and the ‘post-training’ session. This should minimise any time-dependent performance effects that may otherwise have confounded the results.

Directly before the first session began, each subject was provided with the opportunity to familiarise themselves with the stimuli once again using the familiarisation page encountered during the familiarisation and training phase described in Section 4.3.1. Whilst auditioning the stimuli, subjects were required to think about assigning a value to the width of the various solo items on a 0-100 point scale. The subjects were given as much time as they needed for the familiarisation, which ranged from 1-4 minutes for the Music & CSD students, and 3-10 minutes for the Tonmeister students. There was no significant reduction in the time taken for the familiarisation phase in the pre- and post-training sessions between the trained or untrained subjects, but the mean time taken for all Music & CSD students did decrease significantly before and after the training (see Figure 33).
Figure 33: Bar graph showing the mean time taken to complete the familiarisation phases during pre- and post-training for the different course groups.

This indicates that the two 'course' groups were not acting in a similar manner. It could have been that the Tonmeisters were taking more care to fully prepare themselves, or it could in fact have been that the Music & CSD students had become more confident (rightly or wrongly) in the task than the Tonmeister students. Without further investigation, it would not be possible to say for sure, but from the background of the students, and the fact that the effect was only seen on all subjects grouped together from each course regardless of 'training', it seems more likely that the Tonmeisters were being less blasé about the familiarisation phase during the second set of results.

Once the subjects were happy with the familiarisation phase, they moved on to the third experimental session, and after one or more days' rest, the fourth experimental session.

As before, each subject rated the SC Width of the various randomised 'processes' of each of the six programme items (also randomised) on a 0-100 point scale. The width grades assigned to each stimulus by each subject were recorded, as were the times taken for each subject to complete each nominally 30-minute experimental session.

After the fourth and final experimental session had taken place, subjects were given a final questionnaire to fill-out. Their thoughts on how the stimuli were created were also canvassed, and this was explained to them if time allowed and if they expressed an interest (where time allowed, all subjects expressed an interest in how the stimuli were created).

The results of the 'post-training' sessions for both the trained and untrained groups could then be compared against their performance in the pre-training sessions. This would gauge the improvement attributable to the intervening training programme against repetition alone.
Pilot investigation

4.5.2 Results

The results from the pre-training and post-training sessions are analysed in order to determine if the trained subjects had improved in their performance in the 'real-world' task.

4.5.2.1 Analysis of Overall 'Error' and Partial Eta Squared

As in Section 4.3.2, a measure of the consistency and sensitivity performance of the subject groups was sought by calculating 'Error' and Partial Eta Squared values for each subject via an ANOVA of the width grade assigned to each stimulus. These values could be compared with their own previous performance, and the naïve listeners’ performance could be compared to that of the experienced listeners.

Figure 34 shows the results of the analysis for all listeners in the untrained and trained naïve groups during their 'pre-training' and then 'post-training' sessions, alongside that calculated for the experienced listeners. At first sight, the result is counter-intuitive: there appears to be more inconsistency in the data provided by the trained subjects during their post-training sessions than during their pre-training sessions.

![Figure 34: Bar graph showing the mean pre- and post-training RMS error values for each subject for all programme items.](image)

Results are split to show the difference between the various subject groups (untrained and trained naïve subjects, and the experienced listeners). Note that the experienced listener data only include the values taken from the six programme items selected for the naïve tests. As the experienced listeners only rated two iterations of the six programme items selected, their results show up in the 'pre-training' column only.
However, none of the differences between the means of the groups are significant. It appears that skills learned in the additional training programme have not transferred to an improved performance at consistent grading of SC Width. It was hypothesised, however, that the individual differences in the subject’s performance were confusing these raw mean values.

The overall consistency of the experienced listeners was also found to not be statistically significantly different to that of either the trained or untrained naïve listeners in either their pre- or post-training sessions. This supports the claim that the ‘real-world task’ was actually a very difficult one, even for experienced listeners.

In order to verify the naïve subject results, difference grades (diff. grades) were analysed for the naïve subject pre-and post-training session data. Diff. grades were calculated by subtracting the two values of SC Width assigned to each stimulus by each subject from one another. Because the stimuli could be rated wider or narrower during consecutive gradings, diff. grades could therefore be positive or negative. This meant that they would cancel each other out to some extent when averaged. In order to remove this possibility, and obtain a value for the magnitude by which the grades deviated from one another, the modulus was taken to obtain an unsigned ‘Absolute Diff. Grade’ for each stimulus and each subject. See Figure 35.

Figure 35: Bar graph showing the results of the diff. grade analysis.

Bars show the mean magnitude of the differences between every stimulus in the pre-training and post-training sessions for the trained and untrained subjects.

This showed a similar result to that in Figure 34, indicating that the mean magnitudes of diff. grade rose after training for the trained group, and decreased (with practice) in the untrained group. However, this was not a statistically significant result.

Further analysis of diff. grades, especially on a subject-by-subject basis became increasingly difficult to interpret because of differences between subjects’ overall and relative performances.
Analysis of the Partial Eta Squared values was then conducted in order to see if they would reveal evidence of any improvements in subjective performance attributable to the training programme. Figure 36 shows the results of the analysis for all listeners in the untrained and trained naïve groups during their 'pre-training' and then 'post-training' sessions, alongside that calculated for the experienced listeners. As with the 'Error' values shown in Figure 34, changes between the naïve listener's pre-training and post-training sessions are not statistically significant. Also similar to the 'Error' values shown is the fact that the overall sensitivity of the experienced listeners is not statistically significantly different to that of either the trained or untrained naïve listeners in either their pre- or post-training sessions. This could be attributable to the individual differences between the subjects in each group averaging-out the data values.

![Figure 36: Bar graph showing the mean pre- and post-training Partial Eta Squared values for each subject for all programme items.](Image)

Results are split to show the difference between the various subject groups (untrained and trained naïve subjects, and the experienced listeners). Note that the experienced listener data only includes the values taken from the six programme items selected for the naïve tests. As the experienced listeners only rated two iterations of the six programme items selected, their results show up in the 'pre-training' column only.

As analysis of overall 'Error' and 'Partial Eta Squared' values in isolation proved inconclusive, it was decided to take another look at the performance of each subject using scatter plot graphs similar to those already used before in Figure 19.

### 4.5.2.2 Scatter Plot Analysis

In order to get an idea of the performance achieved by the experienced subjects – and hence a potential 'target' performance for the naïve subjects – a scatter plot graph similar to those shown in Figure 19 was plotted for the data from the six programme items selected in Section 4.2.3. The graph is shown in Figure 37.
The 'best' performance is seen in the data for subjects 5, 6 and 8 who show the lowest error (best consistency) and highest partial eta squared (best sensitivity). Note that the data only includes the values taken from the six programme items selected for the naïve tests. Data points are labelled by subject number.

As before, the lower the 'Error' score and the higher the 'Partial Eta Squared' value, the better the performance of the listener. From Figure 37, one can see that there are, as suggested in 4.5.2.1, indeed very different performance measures for the individual experienced subjects. Three subjects appear to show very good performance: subjects 5, 6 and 8. Subject 7 has a low error score, but has a partial eta squared value similar to that of subjects 1 and 4. Subjects 1, 2 and 3 can be considered the worst performing of the experienced candidates using these measures.

Regarding the naïve subjects, scatter plots were drawn to show the performance of each subject in the trained and untrained groups, both pre- and post-training sessions. The results for each subject group, pre- and post-training were plotted on separate
Pilot investigation

figures: Figure 38, Figure 39, Figure 40 and Figure 41. The scatter plots show at times very large inter-subject differences in performance measures within the different groups.

**Figure 39: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for trained listeners. Post-Training Sessions.**

![Figure 39 Scatter Plot]

Data points are labelled by subject number.

Expected trends from such graphs would be that the subjects should move towards the bottom right corner of the graph between the pre-training and post-training sessions (whether trained or untrained). This is expected because of the beneficial effects of repetition as the subjects become more and more familiarised with the task. The subjects who completed the additional training programme were expected to have an increased improvement in performance versus the original attributable to any transfer of the training into the width rating experiment.

**Figure 40: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for untrained listeners. Pre-Training Sessions.**

![Figure 40 Scatter Plot]

Data points are labelled by subject number.

However, as can be seen from the results, there was no clear improvement of the subjects as a whole. In fact, individual performances can be spotted moving in a haphazard manner within each of the subject groups. For example, in Figure 38 and
Pilot investigation

Figure 39, compare the performances of subjects 115 and 116 in the pre-training test with the post-training test. Subject 115 improved in both consistency and sensitivity to almost rival the best of the experienced subjects during their post-training session. Subject 116 on the other hand, had an unchanging Partial Eta Squared value (a measure of sensitivity) whilst their error values rose steeply. So much so, that they changed from being the most consistent subject during the pre-training test to being the most inconsistent subject during the post-training test.

Figure 41: Scatter plot graph showing the performance measures ('error' and 'partial eta squared') for untrained listeners. Post-Training Sessions.

The data for the untrained listeners also showed a variety of different changes in performance metrics across the various subjects. For example, subjects 111 and 105 are identifiable as having both increased in consistency (reduced 'Error') and increased in sensitivity (increased 'Partial Eta Squared'). Subjects 107, 110 and 114 on the other hand, all appear to be getting less consistent and less sensitive.

It is interesting to compare the scatter plots of the naïve listeners (shown in Figure 38, Figure 39, Figure 40 and Figure 41) with that of the experienced listeners (shown in Figure 37). As was suggested by the analysis of the overall values of Error (Figure 34) and Partial Eta Squared (Figure 35), the averaged performance of all subjects between the groups is similar, but a range of individual performances exist between the various subjects. From Figure 37 it is noticeable that certain experienced subjects (5, 6, 7 and 8) found it easier to use similar grades for each of the stimuli whilst detecting a variety of differences between them. It is also noticeable that other experienced subjects (1-4) found the task more difficult and produced performances that were more comparable to those of the majority of naïve subjects than their experienced colleagues.

Because of the difficulty in examining trends in each of the subject groups’ performance, a method was devised to show the relative improvement or reduction in performance between each subject in each group to be displayed. It is described in the following section.
4.5.2.3 Consistency and Sensitivity Change Metric Analysis

For this analysis, two additional terms were introduced: Consistency Change Metric and Sensitivity Change Metric.

Consistency Change Metric is defined as the beneficial change in RMS Error between the pre- and post-training sessions. It was calculated by subtracting the post-training 'Error' from the pre-training 'Error'. This would give a positive value if the post-training 'Error' was smaller than the pre-training 'Error'. Therefore if the subject’s consistency improved (their 'Error' decreased) between the pre-training and post-training sessions, they would obtain a positive Consistency Change Metric score. The converse is also true, and a negative Consistency Change Metric would indicate a reduction in consistency.

Sensitivity Change Metric is defined as the beneficial change in Partial Eta Squared between the pre- and post-training sessions. It was calculated by subtracting the pre-training Partial Eta Squared value from the post-training Partial Eta Squared value. This would give a positive value if the pre-training Partial Eta Squared value was smaller than the post-training Partial Eta Squared value. Therefore if the subject’s sensitivity improved (their Partial Eta Squared value increased) between the pre-training and post-training sessions, they would obtain a positive Sensitivity Change Metric score. The converse is also true, and a negative Sensitivity Change Metric would indicate a reduction in sensitivity.

The benefit of analysing the data in this manner is that it removes the inter-subject differences in overall performance, allowing an investigation into the changes in performance only. It also allows the magnitude of these changes to be compared between subjects.

To investigate the changes in each subject’s performance between the pre-training and post-training sessions, scatter plot graphs were created that show Consistency Change Metric against Sensitivity Change Metric for each subject, separated into the trained and untrained groups (see Figure 42). Note that, as these graphs display changes in the subjects’ performance, the inter-subject variability in overall performance has been removed and any inter-subject differences shown on the graphs are attributable to differences between the way in which they improved (or got worse) in performance between the two sets of grading sessions. This means that a subject whose performance remained constant across the trials would not show a change in either metric, and hence would be plotted at the origin on both axes. From Figure 42, one can see that Subject 108 can be seen to show very little change in performance compared to the other subjects. Subjects who improved both in consistency and in sensitivity can be seen in the upper right quadrant of the scatter plot graphs, showing a beneficial change in both measures. Subjects who got worse both in consistency and in sensitivity can be seen in the lower left quadrant of the scatter plot graphs.

By visual inspection of Figure 42, it is possible to get an idea of trends in the data. For subjects in the ‘not trained’ group, a majority appear to reduce in sensitivity and a majority appear to reduce in consistency. Four subjects (107, 110, 112 and 114) are found in the lower-left quadrant whereas two (105 and 111) are found in the upper-right quadrant.
As for the trained subjects, a majority appear to increase in sensitivity, with approximately equal numbers increasing and reducing in consistency. Three subjects (109, 113 and 115) appear in the upper-right quadrant, whereas only one subject (104) can be designated as clearly in the lower-left quadrant.

As these results seemed to vary by course, another set of scatter-plot graphs were plotted, this time with the subjects also separated by course. These are shown in Figure 43.
Pilot investigation

Figure 43 shows that the way in which the subjects drawn from the different undergraduate courses varied considerably. For the Music & CSD course students, no clear patterns emerged on the scatter plots for either the trained or untrained groups. The four untrained subjects (101, 105, 106 and 107) either reduced or increased in consistency or sensitivity in different ways. The trained subjects (102, 103, 104 and 108) whilst tending to reduce in consistency, did not show overall increases or reductions in sensitivity.

The Tonmeister students, however, showed a clearer pattern in the ways in which the performance of the trained and untrained subjects changed. Three of the four untrained subjects (110, 112 and 114) reduced in both consistency and sensitivity, being found in the lower-left quadrant of the scatter plot. The other subject (111) was found in the upper-right quadrant of the scatter plot, having increased in both consistency and sensitivity. On the other hand, three of the four trained subjects (109, 113 and 115) showed an increase in sensitivity and consistency, being found in the upper-right quadrant of the scatter plot. The other subject (116) showed little or no change in sensitivity, but the largest reduction in consistency evident in the data (more of this subject later in the report).

The trends visible in the Tonmeister subjects’ data in Figure 43 appear to indicate that the Tonmeister students who had participated in the training programme had tended to improve in terms of their own sensitivities and a majority of them improved in terms of their own consistencies. However, there was no statistically significant overall increase or reduction in consistency or sensitivity in the measures used so far (shown in Figure 34 and Figure 36).

In order to verify the validity of the trends in Figure 43, bar graphs showing the mean Consistency Change Metric and Sensitivity Metric were plotted and are shown in Figure 44 and Figure 45.

As can be seen in Figure 44, the 95% confidence intervals overlap with the origin, indicating that there is no statistically significant improvement or reduction in grading consistency for the trained or untrained listeners in either the Music & CSD student groups or the Tonmeister student groups.
Figure 45 shows that there is no statistically significant improvement or reduction in the sensitivity of the Music & CSD student subjects (both the trained and untrained group).

However, the Tonmeister students’ data again shows a different phenomenon: the untrained listener group showed no increase or reduction in sensitivity, but the 95% confidence intervals on the trained Tonmeister subjects’ bar graph show that the increase in sensitivity, although small, is statistically significant for the group. See Figure 46.

To summarise, Figure 44, Figure 45 and Figure 46 show the means of the individual subjects’ improvement or reduction in the measures of consistency or sensitivity used thus far. A positive value indicates an improvement in the individual subjects’ performance in the specified measure between the values obtained in the pre-training sessions and the post-training sessions. Because the confidence intervals of many of the measures for the trained and untrained subjects in both courses overlapped with the origin, there was no statistical improvement or reduction in the measures for
Pilot investigation

these groups. The trained Tonmeister subject group, however, showed a small, but statistically significant increase in their sensitivity as a result of the training scheme.

In other words, although when analysed together, the consistency or sensitivity of all subjects did not improve when trained or through repetition alone, it has been shown that there was an improvement in sensitivity on a subject-by-subject basis for the trained subjects from the Tonmeister course.

It can therefore be stated that the training programme did not affect the consistency by which the subjects graded absolute SC Width in a significant way. The training programme did, however, improve the sensitivity score of the group of the four subjects drawn from the Tonmeister course.

That the training programme had an effect on the subjects drawn from only one of the courses needs to be investigated further. It could be argued that, whilst expressing similar experience levels in the pre-test questionnaires (Figure 16), the actual listening experience of each of the subjects, and their motivation to become better technical listeners was different.

It has to be admitted at this point that, as only 16 subjects were used in total, the individual sample size of four subjects in the trained Tonmeister group is small, so any conclusions need to be treated with care. The result does, however, inspire a further analysis of the resulting data, and further investigation into the cause of this effect.

4.5.2.4 SSED and Range Analysis

The training programme implemented in Section 4.4 was assessed and targeted towards the successful ranking of five stimuli in terms of SC Width. The training programme was also shown to be successful, as all subjects and individual subjects reduced their SSSED from the pre-test to the post-test before and after the training programme. Due to these facts, an analysis of the way in which subjects ranked the different ‘processes’ (each individual stimulus) within each programme item was investigated. The objective was to discover whether or not the training programme affected the way in which subjects ranked the stimuli during the two experimental sessions (pre- and post-training).

In order to work out the differences in the ranking of the stimuli of each programme item by subject, the Sum of Squares of Euclidean Distances (SSED) was again used. Each subject was analysed separately. Here the grades assigned to each ‘process’ within each programme item in each subjects’ experimental sessions were ranked from 1-8. The SSED was then calculated between the first and second experimental session to obtain a value for SSED during ‘pre-training’, then calculated for each programme item between the third and fourth experimental session to obtain values for SSED during ‘post-training’ for each programme item. The maximum SSED possible for 8 ranked items is 168. The experienced listeners SSED values were calculated for the six programme items selected in Section 4.2.3.

SSED in these cases is therefore a measure of the consistency of ranking of the stimuli within each programme item. A low SSED value will therefore indicate that a subject has ranked the stimuli within the programme item in a very similar manner in both sessions. High values of SSED (up to 168) would indicate more difference in the way in which the stimuli were ranked between the two grading sessions.
This is somewhat different to the SSED analysis used in Section 4.4.5, because for the training stimuli there was an established 'correct' order, whereas with the SSED analysis used here, neither of the two rank orders are considered 'correct', one is simply compared against the other.

It is worth noting that the way in which the SSED values were calculated here did not take into account tied scores. Tied scores on the 0-100 point scale were expected to occur very infrequently together on each page of 8 stimuli. A completely accurate analysis method would have been to assign an identical rank number to each stimulus that was graded identically; the rank assigned being the mid-point value between the actual ranks that were covered. For example: a rank of '1.5' would be awarded for two identically ranked items that occupied the top two ranks. However, the amount of extra time needed to rank these values to this degree of accuracy was expected to reach the point of diminishing returns, as even if identical values occurred they would be assigned adjacent ranks, minimising the Euclidean Distance between them in any case. Any undue increase in SSED due to identically ranked items was also considered to be a fair penalty to any subject who was not able to distinguish between the widths of all of the processes.

A bar graph was plotted to show the mean SSED achieved by the trained, untrained and experienced subjects for all programme items and is shown in Figure 47.

Results are split to show the difference between the various subject groups (untrained and trained naïve subjects, and the experienced listeners). Note that the experienced listener data only includes the six programme items selected for the naive tests. As the experienced listeners only rated two iterations of the six programme items selected, their results show up in the 'pre-training' column only.

No statistically significant overall changes in mean SSED grades can be seen from Figure 47 pre-training or post-training for the naïve subjects. It is expected, however, that individual differences between the performances of the subjects, and
the relative difficulties of the different programme items could be obscuring any finer detail within the data.

The experienced listener's mean SSED grades were not statistically significantly different to those of the post-training grades of either the trained or untrained naïve subjects, but these data show that the untrained listeners pre-training SSED were statistically significantly higher than those of the experienced listeners. This indicates that although there was no statistically significant difference between the trained and untrained subjects' SSED scores either pre-training or post-training, the untrained group was shown to be less consistent than the experienced subjects at ranking the stimuli using SSED as a measure. Although care was taken to select subjects of similar abilities for each of the trained and untrained groups, this was done using the consistency and sensitivity measures previously investigated.

The results were also examined by course to see if the different undergraduate groups were behaving in a different manner (as had been evident in Figure 45). But, as can be seen in Figure 48, no statistically significant changes were observable between the subject groups within each course or indeed between courses.

**Figure 48: Bar graphs showing mean SSED for Untrained and Trained Naïve Subjects by Course.**

Due to the way in which the SSED values were calculated, it was also possible to analyse the extent to which the subjects were utilising the 0-100 point scale. A new term was introduced to describe this: Range. Range in this study is defined as the difference between the maximum and minimum grade assigned to the stimuli within each programme item. Range values were calculated for each programme item and for each subject. If a subject was using more of the scale than they previously were (due to enhanced detection of differences between the stimuli, for example), then the value of range would increase for that subject. As discussed in Section 2.1.1, Eisler theorised that expert listeners would use more of the scale than naïve listeners.

In order to see how range was being affected by participation in the additional training programme, a bar graph showing the mean range of each of the subject groups during pre- and post-training sessions was plotted (see Figure 49).
Pilot investigation

Figure 49: Bar graph showing mean Range Pre- and Post-Training for each subject for all programme items.

Results are split to show the difference between the various subject groups (untrained and trained naïve subjects, and the experienced listeners). Note that the experienced listener data only includes the six programme items selected for the naïve tests. As the experienced listeners only rated two iterations of the six programme items selected, their results show up in the 'pre-training' column only.

The graph shows that, on average, the trained group was using more of the scale after training than before. They were also using more of the scale than the experienced listeners. Closer inspection of the range data showed that it was being dominated by one subject — subject 116, who was using very much more of the scale during the post training sessions than before. This can be seen in an analysis of the mean range pre- and post-training by subject plotted in Figure 50. Subject 116 is shown demonstrating a markedly different range change between pre-training and post-training sessions than the other subjects. In fact, without subject 116’s data included, the trained group showed a statistically insignificant range difference between the pre- and post-training sessions.

Figure 50: Bar graphs showing mean Ranges for each naïve subject (numbered above each graph) for all programme items during the pre and post tests.

Note Subject 116’s very much increased Range.
Pilot investigation

The training programme appears to have been responsible for Subject 116’s different use of the scale. During the post-training sessions they appear to have been grading at least one stimulus as having width zero and one stimulus as having width 100 in every programme item (because of the size of the 95% confidence interval). With all other subjects, the changes in range were not as pronounced. Subject 116’s increased use of the grading scale could also be the reason behind their marked decrease in grading consistency shown in Figure 43. As they used much more of the scale, the individual errors between width grades assigned to individual stimuli would become much larger in magnitude when compared to those achieved when they were using far less of the scale. That their ‘sensitivity change metric’ (which was based upon values for ‘partial eta squared’ in the ‘stimulus’ grades) was shown to not change in Figure 43 is perplexing as they were undoubtedly using more of the scale. Perhaps the increased error involved was confounding any noticeable improvement in sensitivity, or the variability in the different sensitivities within the various ‘programme items’ needed to be taken account in the previous analysis.

In order to investigate the results plotted in Figure 49 in more detail, a similar method to that described in Section 4.5.2.2 was used to investigate the relative improvement or reduction in SSED and Range between each subject. Two more terms were introduced: \textit{SSED Change Metric} and \textit{Range Change Metric}.

\textit{SSED Change Metric} is defined as the beneficial change in SSED between the pre- and post-training sessions. It was calculated by subtracting the post-training SSED from the pre-training SSED. This would give a positive value if the post-training SSED was smaller than the pre-training SSED. Therefore if the subject’s consistency of ranking the stimuli improved (their SSED decreased) between the pre-training and post-training sessions, they would obtain a positive SSED Change Metric score. The converse is also true, and a negative SSED Change Metric would indicate a reduction in the consistency by which the subjects ranked the stimuli within specific programme items.

\textit{Range Change Metric} is defined as the beneficial change in Range between the pre- and post-training sessions. It was calculated by subtracting the pre-training Range value from the post-training Range value. This would give a positive value if the pre-training Range value was smaller than the post-training Range value. Therefore if the subject’s use of the scale increased between the pre-training and post-training sessions, they would obtain a positive Range Change Metric score. The converse is also true, and a negative Range Change Metric would indicate a reduction in the use of the scale.

The benefit of analysing the data in this manner is that it removes the inter-subject differences in overall performance, allowing an investigation into the changes in performance only. Because SSED and Range values were calculated for each subject and for each programme item, the SSED Change Metric and Range Change Metric analysis also takes into account individual differences between the various programme items.

Figure 51 shows scatter plot graphs of SSED Change Metric against Range Change Metric for each subject, separated into the trained and untrained groups by course. Each point on the graphs shows how different subjects’ performance for each of the six programme items improved or reduced in terms of consistency of rank ordering (SSED) and use of the scale (Range).
Note that, as in Figure 42, these graphs display changes in the subjects' performance between the first two experimental grading sessions, and the second two sessions. The inter-subject variability and the variability between programme items have been removed. Any inter-subject differences and subject-specific inter-programme item differences shown on the graphs are attributable to differences between the way in which the subjects improved or got worse in performance between the two sets of grading sessions. This means that a subject whose performance remained constant across the trials would not show a change in either metric, and hence would be plotted at the origin on both axes.

Whilst it is relatively difficult to interpret Figure 51 due to the number of points plotted on each of the graphs, it is possible to see some trends in the data.

Each individual point shows how each subject (numbered) rated each programme item (shown as different shapes).

It appears that both sets of untrained subjects have a scattering of data points in all four quadrants of their scatter plots, showing that SSED Change Metrics and Range Change Metrics were improving and reducing in roughly equal numbers for these groups. Noticeable are the improvements in certain programme items by subjects 101 and 111 which stand out over the other data points.

As far as the trained listeners are concerned, it is possible to see a general positive shift in the data points along the 'range change metric' axis in both trained listener graphs, most notably on those points attributed to Subject 116 (who has already been shown to have increased their use of the scale by much more than the other subjects – see Figure 50). As far as SSED change metric is concerned, it appears that there is a general positive trend for the Trained Tonmeister subjects, as what seems to be a majority of data points can be found in the upper right quadrant, but there appears to
Pilot investigation

be at least equal numbers of positive and negative changes amongst the Trained Music & CSD subjects.

In order to quantify the changes in SSED and Range brought about by participation in the training programme, bar graphs showing SSED Change Metric for trained and untrained subjects (Figure 52 and Figure 53) and Range Change Metric (Figure 54 and Figure 55) were plotted.

A bar graph of a change metric will show if there was an overall increase or decrease in the measure concerned between the pre-training sessions and the post-training sessions. If the 95% confidence intervals overlap with the origin, it can be said that there is no statistically significant increase or decrease in the measure concerned.

Figure 52 shows the overall SSED Change Metric for the trained and untrained subjects. As can be seen, the confidence intervals overlap with the origin in both cases. This means that no statistically significant change occurred in the SSED score for either the subjects who participated in the training programme or those who relied on repetition alone.

Figure 53 confirms that this is also the case for the various subject groups when split by course.
This shows that participation in the training programme did not significantly increase the ability of the subjects to rank the stimuli in terms of SC Width consistently between subsequent trials of the 'real-world' task. Importantly, this was also the case for those who did not take part in the additional training programme but simply repeated the experimental sessions.

An analysis of the Range Change Metric, on the other hand, shows significant changes are evident in both trained and untrained subjects.

Figure 54: Bar graphs showing the effect of the training programme on the Range Change Metric.

![Bar graphs showing the effect of the training programme on the Range Change Metric.](image)

Figure 54 shows the Range Change Metric of the trained and untrained subjects. As can be seen from the graph, the untrained subjects showed a statistically significant decrease in Range Change Metric (they used less of the scale during the post-training sessions than during the pre-training sessions), whereas the trained listeners increased their Range Change Metric in a statistically significant manner (they used more of the scale during the post-training sessions than during the pre-training sessions).

The exclusion of Subject 116’s data (which featured a very much larger increase in Range Change Metric) does not change the significance of this overall effect.

In order to see if the different course groups reacted differently, the data was plotted by course in Figure 55.

Figure 55: Bar graphs showing the effect of the training programme on the Range Change Metric by Course.

![Bar graphs showing the effect of the training programme on the Range Change Metric by Course.](image)
From Figure 55 one can see that statistically significant changes were observed for three of the four groups.

The untrained Music & CSD subjects used no more or less of the scale in the post-training sessions than they did during the pre-training sessions. However, the untrained Tonmeister subjects showed a statistically significant \textit{decrease} in Range Change Metric.

This contrasts with statistically significant \textit{increases} in Range by both sets of trained subjects. (The exclusion of Subject 116's data from that shown in Figure 55 causes the confidence levels to overlap with the origin, but this is understandable as this reduces the sample size by a quarter for that group).

The analysis of SSED Change Metric and Range Change Metric has been revealing. It has shown that there was \textit{no improvement} in the ability of subjects to rank the stimuli of individual programme items in a consistent manner over subsequent tests whether the subjects took part in the additional training programme or not. It has also shown that there was a statistically significant increase in the subjects' use of the scale after participation in the training programme.

\subsection*{4.5.2.5 Analysis of Time Taken for the Task}

In order to test whether or not the time taken for subjects to complete the nominally 30-minute experimental sessions (and hence their fluency) had changed as a result of the training programme, each subject's time taken per session was analysed. Figure 56 shows the mean time-taken per subject by training group. The data is separated into the first two (also called the 'pre-training' sessions for the naive subjects), and the second two (also called the 'post-training' sessions for the naive subjects). Data from the four sessions undertaken by the experienced subjects has been included, separated nominally into the first two and second two sessions to correspond to the naive subjects' 'pre-training' and 'post-training' sessions.

As shown in Figure 56, the time that the various groups took to complete the sessions was not statistically significantly different, either between the first and second two sessions, or between the various groups.
In case inter-subject differences were obscuring the results, a ‘Change Metric’ style analysis on the naïve subject data was undertaken in a similar way to those used previously in this report. Here *Time Change Metric* is the beneficial change in the time taken per session. This was calculated by taking the average of the second two sessions’ times and subtracting it from those from the first two sessions. A positive ‘Time Change Metric’ would therefore be obtained if the average time for the second two sessions had reduced to be less than that of the first two sessions. The results for the trained and untrained groups are shown in Figure 57.

![Figure 57: Bar graphs showing the effect of the training programme on the Time Change Metric.](image)

As can be seen (and is also evident in separate plots by Course), there was no statistically significant increase or decrease in the time taken to complete the tasks. This is the case even though the data has been analysed on a subject-by-subject basis.

### 4.5.2.6 Post Experiment Questionnaire

The naïve subjects were all asked to complete a final questionnaire. Subjects responded to four statements using a five-point ordinal scale:

- At first I found the task (judging width) difficult:
- By the end I found the task (judging width) difficult:
- I am now able to use the width scale more consistently than when I started (e.g. wide sounds have always been graded wide):
- I am now able to detect differences between the widths better than when I started:

The responses were given ordinal values (1=Strongly disagree; 2=Disagree; 3=Neither agree nor disagree; 4=Agree; 5=Strongly agree) and plotted as histograms in Figure 58, Figure 59, Figure 60 and Figure 61. The histograms have been plotted for the trained and untrained listeners and show the frequency with which each response was elicited.

Figure 58 shows the response to the statement ‘At first I found the task (judging width) difficult’. It shows that the majority of the subjects found the task difficult at
the beginning. Noticeably, a couple of the untrained group stated that they did not find the task difficult.

**Figure 58:** Histogram showing response for the trained and untrained subjects to the statement “At first I found the task (judging width) difficult”.

Responses were on a five-point ordinal scale: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree

Figure 59 shows the response to the statement ‘By the end I found the task (judging width) difficult’. The responses have now shifted into the negative portion of the five-point ordinal scale. Untrained listeners are now ambivalent to the statement, indicating that they found it easier than at first. However, a majority of the trained subjects were prepared to disagree with the statement (after training they did not find the task difficult). Only one of the trained subjects was prepared to state that they had found the task difficult after the training programme.

**Figure 59:** Histogram showing response for the trained and untrained subjects to the statement “By the end I found the task (judging width) difficult”.

Responses were on a five-point ordinal scale: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree
Figure 60 shows that fourteen of the sixteen subjects believed that they had become more consistent in the use of the grading of the stimuli. The untrained subjects were just as convinced as the trained subjects.

Figure 60: Histogram showing response for the trained and untrained subjects to the statement "I am now able to use the width scale more consistently than when I started (e.g. wide sounds have always been graded wide)."

![Histogram](image)

Responses were on a five-point ordinal scale: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree

Figure 61 shows that all subjects believed that they had become more able to detect differences in width between the stimuli. Again, the untrained subjects were just as convinced as the trained subjects.

Figure 61: Histogram showing response for the trained and untrained subjects to the statement "I am now able to detect differences between the widths better than when I started".

![Histogram](image)

Responses were on a five-point ordinal scale: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree
Pilot investigation

So, interestingly enough, subjects who simply repeated the task without taking part in the extra training programme seemed just as convinced that repetition of the task had helped them to improve their consistency and sensitivity as those subjects who had done the extra training.

However, the trained subjects appeared to be more convinced that the task had become easier after the training programme. This implies that participation in the additional training had made the trainees more comfortable with the task of rating SC Width than their untrained counterparts.

4.6 Pilot Investigation: Overall Summary

A study was conducted to establish whether or not trained spatial audio listening skills could be transferred from one task and stimulus set to another task and stimulus set. Sixteen listeners were tested and separated into two groups of equivalent performance in a spatial audio attribute rating task. One of the groups underwent a formal training programme – a modified implementation of the one used in (Neher 2004) – which trained the detection and ranking of differences in a spatial audio attribute (Individual Source Width) using a separate set of contrived stimuli, which were provided in (Neher 2004). The other group did not take part in any additional training. There was an established ‘correct’ order in which to rank the items, so it was therefore possible to measure the correctness of each trainee’s response. The trained group showed a significant improvement in the way that they ranked the audio stimuli used in the training scheme (t=2.524, p < 0.05) after training. This means that the modified training system was able to achieve near transfer, in that there was an increased performance on the trained stimuli. Because there was no control group for the training phase (as there had been in (Neher 2004) - albeit with just two subjects), there is no way to rule out that such an increase could have occurred by taking the pre- and post-training tests without the intervening training.

Both groups were then retested on the spatial audio attribute rating (far transfer) task. The only transferred training effect observed was in the way the subjects used the 0-100 point scale to rate the items. The trained subjects used a significantly greater range of the scale to express their judgements after training, whereas the untrained subjects used a significantly smaller range of the scale to express their judgements. No change was seen in either group relating to their consistency or fluency.

The pilot study can be classified as a ‘Test Type 2’ transfer experiment (see Table 13 in Section 2.3.1.2), with the exception that the pre-test was actually the entire transfer test taken before any groups were separated and half of them were trained. The results show that students from the two courses reacted in different ways to one another. Whilst there was no significant improvement in consistency of rating or ranking in any subject group, there was an increase in sensitivity (as measured using partial eta squared) in the trained listeners from the Tonmeister course. In agreement with a hypothesis proposed by Eisler (see Section 2.1.1), this study also seems to indicate that trained listeners become more confident at using more of a given rating scale, because the range of values used to describe the widths within each programme item increased. Untrained listeners from the Tonmeister course actually
used less of the scale during the 'post-training evaluation' experiments than previously.

The lack of improvement of consistency in the trained listeners may be attributed to the use of different response paradigms for the training and evaluation tasks. A grading scale was employed in the evaluation task, whereas rank ordering was used in the training system. Anecdotal feedback from most subjects (including the experienced listeners) indicated that the evaluation task was considered to be a very difficult one. This was backed up by the fact that half of the experienced listeners (used to evaluate the stimuli before the naïve listeners were tested) produced performances similar to many of the untrained naïve listeners in terms of consistency and sensitivity. It is hypothesised that the task would become less demanding and closer to the training scheme if subjects were instead asked to rank the eight stimuli for each programme item. Skills learned in the rank ordering tasks of the training system are expected to transfer to increased consistency in the rank ordering of stimuli of each programme item. However, if the task is too easy, transfer might not be demonstrable due to the fact that the target task required no additional training.

The observed lack of far transfer of skills from the training environment to the 'real world' task is an important issue. It is possible that the lack of transfer occurred because the rating task was too difficult (and indeed, even experienced listeners struggled to be consistent and sensitive when responding). Another likely factor was a potentially demotivating aspect of the training programme which involved negative feedback being given for incorrect answers in the form of a cartoon character and comic sound effect. Furthermore alternative levels of transfer might have been achieved but had not been investigated. It is possible that results could have been different if the recordings detailed in Section 4.1 had been made in a different recording environment, but further research would be needed to verify this.

One of the problems reported by the participants in this study was the lack of variety of stimuli. As suggested in Section 2.3.1.3, this is something that is addressed in the study detailed in Chapter 5.
5 TRANSFER INVESTIGATION

A new Spatial Audio Attribute Listener Training System (SAALTS) was developed to provide a generalisable training programme (see Section 2.3.1.8). SAALTS is detailed in Section 5.4.

An investigation was conducted to evaluate SAALTS based upon two main issues. The first was the inconclusive nature of the results from the pilot experiment: Whilst the training system showed dramatic results using its own stimuli, transfer to a different situation seemed severely limited. Various forms of transfer needed to be studied in order to be able to discover transferred skills that were potentially hidden in the previous study.

The second motivation was the reliance on repetitive practice in previous studies and standards (see Sections 2.2.1.1 and 2.2.1.3). If SAALTS is to be shown to be useful it will not only need to produce improvements and be transferable, but it should also compare favourably with repetitive practice.

In order to gauge the effectiveness of a spatial audio attribute training regime, it was necessary to compare it against two control groups, one that repetitively practiced the task and one that did no additional training or practice. Comparison with untrained subjects allowed the overall training effect to be quantified. Comparison with repetitive practice would allow the training system to be measured against a previously established method (see Sections 2.2.1.1 and 2.2.1.3).

Because potential transfer effects, especially those concerning near transfer, could have been missed during the pilot (see Chapter 4), a range of transfer tests were devised in order to evaluate a new training system.

5.1 Transfer Investigation Outline

The overall method was to recruit and screen a group of subjects using an initial spatial attribute evaluation task so that three equivalent-skill subject groups could be created. The relative performances of three groups of subjects before and after some experimental manipulation would then be investigated. The three groups were separated into a group that would utilise a spatial training system, another that would repetitively practice an initial screening task, and a control group that would not undergo an additional training or practice regime.

The task of rank-ordering simulated spatial audio stimuli was found to be an effective way of evaluating training in a previous study by Neher (see Section 2.2.3.4) and in the pilot study detailed in Chapter 4. Rank-ordering tasks therefore formed the basis of the transfer study and are explained in Section 5.3. The creation of stimulus sets for the ‘ranking’ tasks is detailed in Section 5.2.1.

The training and practice experimental groups were given an identical length of training, in that they each performed six additional half-hour sessions in which to train or practice. This controlled the amount of additional time between the two experimental groups. Six experimental sessions was also suggested by Bech (see
Section 2.2.1.3) and 6.25 experimental sessions was the average number of sessions
taken to complete the training programme provided in the pilot experiment (see
Section 4.4.4). The Spatial Audio Attribute Listener Training System (SAALTS) is
detailed in Section 5.4.

After the two experimental manipulations (the training and practice phase) all three
groups were tested once again using the ranking task in order to compare their
performance in near transfer tasks.

To test for far transfer, two different transfer scenarios were used.

Firstly the post-test ranking task was repeated using a different set of stimuli (the six
stimulus sets generated in Section 5.2.1 that had been held in reserve). Examining
the performance (between groups rather than pre-post) in these tasks would therefore
indicate how effectively training and practice would transfer to stimuli other than
those already encountered.

Secondly, a set of stimuli were created in a different manner to those created for the
ranking tasks to make them more ecologically valid. This resulted in stimuli
(described in Section 5.2.2) where many different attributes of the sound
reproduction changed. The ability of subjects to discern and describe a particular
sensory characteristic in a “sea” or “fog” of other sensory impressions is more
important than sensory acuity (Meilgaard, Civile and Carr 1991). The evaluation
task was also altered accordingly. Examining the performance in these tasks (in
terms of rating sensitivity and consistency in a similar manner to that described in
Sections 4.2, 4.3 and 4.5) would therefore indicate how effectively training and
practice would transfer to stimuli and situations other than those already
encountered. These ‘rating’ tasks are described in Section 5.5.

If training or practice were shown to improve performance with such stimuli
(whether using the near or far transfer test paradigms) then this would be powerful
evidence for their wider usefulness.

Therefore the following hypotheses were tested:

- Both the trained and practice groups will show improved performance in the
  near transfer test and far transfer tasks over the untrained group, and over
  their previous performance. (Because practice and near-transfer training will
  aid the initial test).

- The practice group will show improved performance over the trained group
  for the near test and stimuli, because they practiced on a task and stimuli
  closer to the initial task than the training tasks were.

- The trained group should show improved performance over the practice
  group for the other transfer tests. Because more decontextualised training
  and varied examples will lead to greater far transfer (see Sections 2.3.1.3 and
  2.3.1.8). Encouraging trainees to reflect upon what they have learned is also
  considered to be beneficial for transfer (see Section 2.3.1.7).

Section 5.1 details the transfer investigation whose outcomes are described in
Section 5.6.

The investigation followed a pre/post test methodology. Subjects were pre-tested
using a ranking session (detailed in Section 5.3). This is known as ‘Task 1’ of the
transfer investigation. The subjects’ resulting performance was used to separate them into three groups of approximately equal skill. One of these groups was trained using SAALTS over six sessions, each lasting half an hour, whilst another repetitively practiced the ranking task over six 30-minute sessions (also known as ‘Tasks 2-7’). The third did no additional practice or training. Thereafter the three groups were tested once again using the ranking task (‘Task 8’) in order to compare their performance in near transfer, again using a ranking task with different stimuli (‘Task 9’). Finally all three groups took part in the ‘further’ transfer ‘rating’ session (‘Task 10’) detailed in Section 5.5. A group of experienced listeners also took part in ‘Task 10’ in order to provide a performance benchmark for the task. An overview is shown in Figure 62.

Figure 62: Transfer Investigation Overview.

Tasks 1-10 took place in the listening room at the University of Surrey. This room conforms to ITU-R recommendation BS. 1116 (ITU-R 1994-1997) and features five active loudspeakers (Genelec 1032A). The loudspeakers were placed 2.2m from the listening position in the 3/2 stereo configuration (ITU-R 1992-1994). The tests, practice and training were administered via a DAW running Max/MSP 4.5 from Cycling '74. The computer was situated in an adjacent room connected to a keyboard, mouse and a video monitor in the listening room via extended cables. During the training phase, a portable PC was connected to the video monitor in order to display the tutorial presentation to the trainees. Figure 63 shows the layout of the listening room.
5.2 Stimulus Creation

In order to provide a varied training programme, and to reduce potential boredom for all subjects in the tests (trainees in the pilot experiment had called for greater variety of stimuli – see Section 4.4.7), it was necessary to generate a varied set of stimuli for the ranking tasks. Section 5.2.1 details this process.

In addition, different stimuli were required for the far transfer task that would test for transfer of the trained skills to a different task and environment. Section 5.2.2 describes a procedure similar to the one outlined in Section 4.1 where simultaneous complex spatial audio signals were captured.

5.2.1 Ranking/Training Task Stimulus Set Creation

Variety was provided for this investigation over the pilot investigation detailed in Chapter 4 by utilising two spatial audio attributes, and twelve different programme items. Variety aids in the decontextualisation of the stimuli and hence boosts transfer (Section 2.3.1.7), whilst allowing for an expanded range of task difficulty.

Neher had provided a multi-channel audio processing platform which can create stimulus sets from mono source recordings.

Neher’s Individual Source Width stimuli had been utilised in the pilot study (Chapter 4), and had actually been the most problematic attribute simulation according to (Neher 2004). Individual Source Distance was relatively straightforward requiring potentially less training. However the two ensemble attributes Ensemble Width (EW) and Ensemble Depth (ED) were distinct, but could be potentially simulated using new, similar source material to provide new stimulus sets with a degree of familiarity between the attribute simulations.
Source material was selected from a set of commercial recordings (Power FX 1999) featuring ensembles of various instruments recorded separately in an acoustically dry environment. Twelve different programme items were selected, each with four sources. Table 20 details the selected programme items.

### Table 20: Programme Item Listing

<table>
<thead>
<tr>
<th>Programme Item ID</th>
<th>Original Track No./Name</th>
<th>Source 1 (Far Left)</th>
<th>Source 2 (Left)</th>
<th>Source 3 (Right)</th>
<th>Source 4 (Far Right)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2: Reflective</td>
<td>Drums</td>
<td>Bass Guitar</td>
<td>Acoustic Guitar</td>
<td>Electric Guitar</td>
</tr>
<tr>
<td>B</td>
<td>3: Psychedelic Hitchhiker</td>
<td>Bass Guitar</td>
<td>Electric Guitar</td>
<td>Electric Piano</td>
<td>Drums</td>
</tr>
<tr>
<td>C</td>
<td>4: Intrigue</td>
<td>Electric Guitar 1</td>
<td>Electric Guitar 2</td>
<td>Drums</td>
<td>Bass Guitar</td>
</tr>
<tr>
<td>D</td>
<td>5: Highway Fire</td>
<td>Acoustic Bass</td>
<td>Violin</td>
<td>Electric Guitar</td>
<td>Cello</td>
</tr>
<tr>
<td>E</td>
<td>6: For the Byrds</td>
<td>Electric Guitar</td>
<td>Acoustic Bass</td>
<td>Electric Piano</td>
<td>Violin</td>
</tr>
<tr>
<td>F</td>
<td>8: Cosmic Traveller</td>
<td>Electric Guitar 1</td>
<td>Drums</td>
<td>Electric Guitar 2</td>
<td>Cello</td>
</tr>
<tr>
<td>G</td>
<td>13: Days End</td>
<td>Cello</td>
<td>Electric Guitar</td>
<td>Violin</td>
<td>Drums</td>
</tr>
<tr>
<td>H</td>
<td>14: Dew Drops</td>
<td>Acoustic Bass</td>
<td>Violin</td>
<td>Acoustic Guitar</td>
<td>Cello</td>
</tr>
<tr>
<td>I</td>
<td>15: Dirge</td>
<td>Electric Guitar</td>
<td>Bass Guitar</td>
<td>Cello</td>
<td>Violin</td>
</tr>
<tr>
<td>J</td>
<td>17: Sunday Drive</td>
<td>Acoustic Guitar</td>
<td>Bass Guitar</td>
<td>Violin</td>
<td>Electric Guitar</td>
</tr>
<tr>
<td>K</td>
<td>18: Contemplation</td>
<td>Cello</td>
<td>Acoustic Guitar</td>
<td>Bass Guitar</td>
<td>Violin</td>
</tr>
<tr>
<td>L</td>
<td>19: Hero Theme</td>
<td>Acoustic Guitar</td>
<td>Bass Guitar</td>
<td>Drums</td>
<td>Electric Guitar</td>
</tr>
</tbody>
</table>

The various source tracks were selected and auditioned to find musical phrases. Four mono sound files of the appropriate length were created for each programme item.

The source files were then loaded into Neher’s processing platform, which simulated the nine levels of SC Depth and SC Width. The resulting 3/2 stereo (ITU-R 1992-1994) outputs were recorded and further edited (to accurately position start and end points for looping) and saved as multichannel wave format sound files (nine for each attribute within each of the twelve programme items).

Neher’s *ensemble width* simulations normally require five sources (two that move to the left, two that move to the right and one that remains in the centre). Rather than finding five-source items, or adding a fifth source to the selected programme items, the simulations were generated without a centre source. Sources 1 and 2 of the selected programme items were therefore placed in the *left* side positions of the simulation. Sources 3 and 4 were placed in the *right* side positions. Informal listening by the author and his colleagues found that the simulations demonstrated the perceptual illusion of the widening of an ensemble of sources appropriate for SC Width changes despite the lack of a centre source. Figure 64 shows a plan-view of the *ensemble width* simulation.
Neher's *ensemble depth* simulations require four sources. Sources 1 and 4 of the selected programme items were therefore placed in the *outer* positions (moving *toward* the listener with increasing *ensemble depth* in the simulation). Sources 2 and 3 were placed in the *inner* positions (moving *away* from the listener with increasing *ensemble depth* in the simulation). The 'flattest' stimulus simulated the sources...
Transfer Investigation

arranged in a line. The ‘deepest’ featured the largest distance between the outer source and the inner sources. Figure 65 shows a plan-view of the ensemble depth simulation.

Of the twelve programme items, six were chosen to feature in the training system (A, B, E, G, J, L) and a further six were chosen to be held in reserve for some of the transfer tests (C, D, F, H, I, K). Both sets had a similar mix of musical styles and instrumentation.

The simulations of the two ensemble attributes utilise changes in both ISC Distance and ISC Direction. The use of these more complicated attribute simulations include these subsidiary attributes which can be focused upon and discussed in the training system. This allows for all source-related attributes (see Section 3.2.4) of the SSBP to be covered. Additionally the two attributes should provide a range of difficulty as well as variety within the training.

5.2.2 ‘Real-World’ Stimulus Creation

The situation and stimuli should be different to the trained environment to provide a far transfer task – see Section 2.3.1. The ability of subjects to discern and describe a particular sensory characteristic in a “sea” or “fog” of other sensory impressions is more important than sensory acuity (Meilgaard, Civille and Carr 1991). Stimuli where spatial attributes were changing in a less controlled way to those simulated in Section 5.2.1 were therefore required.

Simultaneous multiple microphone recordings had proved to be a convenient way of creating a series of switchable and complex multichannel stimuli in the pilot experiment (see Section 4.1). A similar method was employed to create the varying stimuli needed for the experiment.

If elements of an ensemble needed to be recorded in various positions, a highly repeatable performance would be essential. Any small changes in the timing or feeling of the performance would be recognisable when switching between stimuli recorded at different times. This problem was solved through the use of a repeatable acoustic playback system. To allow a degree of continuity with the previous training and practice sessions, the original mono source recordings used to create the stimuli for the training system were used.

The experimental recording session took place in Studio 1 at the University of Surrey’s Department of Music and Sound Recording. The studio is 14.5m wide, 17m long and is approximately 6.5m high. It is primarily used for the recording of classical music. The studio is acoustically a typical concert hall, with a reverberation time of approximately 1.5 seconds.

Figure 66 shows the layout of the recording session. The four original sound sources shown in Table 20 were replayed via four loudspeakers (Genelec 1032A) arranged in various configurations toward the front of the studio. The recording set-up consisted of three triplets of microphones positioned facing towards the front of the studio. The three techniques were chosen from the techniques already described in Section 4.1. This consisted of a Fukada triplet (using AKG C451 cardioid microphones), an OCT-inspired technique (using an AKG C414 B-ULS cardioid as the centre microphone and two AKG C414 B-XLS hypercardioid microphones as the side microphones), and an INA-3 technique (using AKG C414 B-ULS cardioid
The three triplets were mounted on a bespoke microphone stand that centred all triplets, and were raised to a height of 220 cm from the ground. A spaced cardioid technique was used to capture ambience, and was implemented using two B&K 4011 microphones at a height of 3.4 m from the ground. These were positioned towards the rear corners of the studio, facing the corners (to reject as much direct sound as possible). All microphones were connected to a DAW via similar microphone pre-amplifiers and level-aligned using a constant level sound source held a fixed distance from the capsule of each microphone. Four outputs of the DAW were connected to the four loudspeakers in order to replay the source files. This enabled the DAW to simultaneously replay the four source files whilst recording the eleven microphone channels. This system offered reliable repeatability as multiple record/replay passes could be completed with almost identical acoustical conditions.

The loudspeakers were repositioned to create different physical widths and depths of ensemble between recording passes. There were six different loudspeaker configurations recorded to recreate six different actual width and depth configurations for the four-source ensemble. Table 21 shows the loudspeaker configurations. These were chosen to provide a good range of ensemble widths and depths whilst fitting within the confines of the studio and within the recording angles of the ‘front’ microphone triplets. They also followed the plan-view of the ED and EW simulations shown in Figure 65 and Figure 64.

Because of the flexibility of the record/replay system it was also possible to change the position of the microphones in order to create a further level of variation. Therefore two positions for the triplets were used – the ‘front’ position which was centred 4.1 m from the ‘centreline’ of the loudspeakers (8.1 m from the front wall), and the ‘back’ position which was centred a further 2.9 m away (total 11 m from the front wall. Figure 66 shows the relative positions of the triplet in the studio.
The signals of the triplets could be combined with the spaced cardioid signals to create unique five-channel recordings for each combination of triplet type, triplet position and loudspeaker configuration.

**Table 21: Loudspeaker Configurations (for the ‘area’ shown in Figure 66)**

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Inner Channel Width</th>
<th>Outer Channel Width</th>
<th>Inner Channel Depth</th>
<th>Outer Channel Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>±18 cm</td>
<td>±64 cm</td>
<td>0 cm</td>
<td>0 cm</td>
</tr>
<tr>
<td>Wide</td>
<td>±64 cm</td>
<td>±224 cm</td>
<td>0 cm</td>
<td>0 cm</td>
</tr>
<tr>
<td>Widest</td>
<td>±173 cm</td>
<td>±420 cm</td>
<td>125 cm</td>
<td>125 cm</td>
</tr>
<tr>
<td>Deep</td>
<td>±64 cm</td>
<td>±224 cm</td>
<td>232 cm</td>
<td>218 cm</td>
</tr>
<tr>
<td>Deepest</td>
<td>±64 cm</td>
<td>±224 cm</td>
<td>125 cm</td>
<td>125 cm</td>
</tr>
<tr>
<td>Wide &amp; Deep</td>
<td>±173 cm</td>
<td>±420 cm</td>
<td>125 cm</td>
<td>125 cm</td>
</tr>
</tbody>
</table>

Microphone configurations were symmetrical about the centre line along the length of the studio, and about a ‘centreline’ of the loudspeakers parallel with and 4m from the front wall of the studio.

‘Widths’ refer to the position of the loudspeakers with respect to the centre line along the length of the studio.

‘Depths’ refer to the position of the loudspeakers with respect to the ‘centreline’ of the loudspeakers parallel with and 4m from the front wall of the studio. Inner channels moved towards the front wall, outer channels moved away from the front wall.

There were, in all, three triplets, two positions for the triplets, and six loudspeaker configurations. This created 36 unique five-channel recordings for each of the twelve quasi-simultaneous programme items.

Figure 67, Figure 68 and Figure 69 show photographs of the experimental recording session.

**Figure 67: Photograph of front triplets in the ‘front’ position.**
To establish how best to use the generated stimuli, an informal listening session was conducted by the author in the listening room. Although changes in the apparent depth of the ensembles were apparent, these were not as obvious as the changes in the apparent width of the ensembles. A wide variety of ensemble width changes
were found across the various stimuli, so this was selected as the most salient change to be employed in the grading experiment.

Out of the entire collection of recordings nine versions of four programme items (B, D, H, L) were selected. These were made up of recordings from the three triplets in the 'front' position, with three loudspeaker configurations chosen ('start', 'deep' and 'widest'). This provided an array of different types of width and depth changes.

5.3 Ranking Sessions

As explained in Section 5.1 rank-ordering of simulated spatial attributes was chosen as the main test type for the study. Because there was an established rank order for the stimuli it was possible to compare each of the responses with the correct order so that the subjects' test performance could be evaluated. This option was not available during the transfer tasks in the pilot investigation (see Sections 4.3 and 4.5), so the use of ranking tasks would provide an improvement in the measurement of training effects. Rank-ordering tasks were therefore chosen to form the basis of the transfer study. Stimuli created in section 5.2.1 were used as the programme items for these tests. All stimuli were looped when played back.

5.3.1 Familiarisation

Each session began with a chance to use the familiarisation page (subjects in the pilot test had requested that feedback be available at the beginning of each session – see Section 4.4.6) in order to re-acquaint the listeners with the stimuli, and give them a chance to prepare themselves for the task – their so-called set (see Section 2.3.1.9).

Figure 70 shows the familiarisation phase screen. Subjects were free to listen to the nine versions of the two attributes for each of the six programme items to be used in the test.

In order to avoid having to explain the SSBP to all subjects, and hence bias the control groups with information destined for the training group, the terms Ensemble Width (EW) and Ensemble Depth (ED) were used during the ranking tasks. Each subject decided the order in which they would attempt the tasks.

Note that, although the terms "Ensemble Width" and "Ensemble Depth" were used in the ranking tasks, this is not because the SC terms as found in the Simplified Scene-Based Paradigm (see Chapter 3) were not appropriate. The use of EW and ED was an expedient way to not need to explain the SSBP to all subjects, and was only possible because all of the selected stimuli contained ensembles of sources. If both single sources and ensembles were used, then it is likely that the SC concept would need to have been explained to all subjects before they started.

The stimulus sets that Neher had provided for ED (a jazz ensemble) and EW (voices) were also made available under programme item 'T'. The EW stimulus set was recreated by this author without a central source so as to match the other four-source programme items used in the transfer study. Subjects were instructed that they should consider programme item 'T' a reference for ED and EW. The multichannel sound reproduction system implemented in all tests in the transfer investigation allowed for seamless switching between simultaneously playing files.
5.3.2 Ranking Tasks

During all ranking sessions (except for the initial session), the familiarisation phase was followed by two sets of ranking tasks, one for each of ED and EW.

Figure 71 shows a screenshot of a page from the ranking tasks (in this example, the first page of an ED ranking page). Nine sliders with nine positions were used to assign a rank order to the nine stimuli replayed via the buttons labelled QWERTYUIO.

There was a notional time-limit of two minutes per page. This was done for various reasons:

- Speed of response is considered to be a sign of proficiency at a task – someone who performs a task as well as someone else but in a shorter amount of time can be assumed to be better at the task. By monitoring the time it takes for subjects to complete the task that important source of potential transfer will not be lost.
- Quesnel had found that if speed of response was not explicitly required, it did not improve – see Section 2.2.2.4.
- Time-scheduling of tests would become impossible if subjects were left to complete the tasks in their own time. A time limit of two minutes per page, with a total of twelve pages to complete per session and about five minutes of familiarisation, would give a total of about half an hour with which to complete the session.

The time-taken per page was monitored with the use of a ‘traffic-light’ timer (which can be seen in the top right corner of the screen in Figure 71). During the first
minute that the subject ranked items on the page the lights were steady green (the top block of nine lights) indicating that the test was running and that all was well. During the second minute of the page the lights changed to yellow (the middle block of nine lights) indicating that the subjects should think about finishing up their rank order and move on to the next page. At the end of the second minute the lower block of nine lights flashed red repeatedly indicating that the time was up and that the Subjects should finish up and move on as soon as possible. The test software would not force the subjects onto the next page at this point, and the subjects were told in advance that they were not to worry unduly if they over-ran on one of the pages as the chances were that they would make up the time on one of the other pages during the test. The ‘traffic-light’ system was chosen because it was instantly recognisable and the meaning of the three phases was easily explained.

Figure 71: Ranking task screenshot.

Once each of the nine ranks had been assigned, the computer allowed the subject to move to the next page (the ‘next’ button would be displayed only after all nine stimuli had been assigned a unique rank to disallow accidental or incorrect results). After six pages of one of the attributes had been completed, the subjects could move on to the other attribute. Once both attributes had been completed the session was over. Ranking and timing information and the human-computer interaction log were saved along with the subject’s profile by the computer.

5.3.3 Initial Ranking Session (‘Task 1’)
During the very first session after the familiarisation stage, the subjects had a chance to practice ranking five of the nine items (stimuli 1, 3, 5, 7 and 9 of the various stimulus sets created in Section 5.2.1) against the clock, and with feedback (shown in Figure 72). There was a notional time-limit of one minute (because there were fewer stimuli involved than the main ranking tasks). Feedback was provided to give the subjects an idea of how they were doing, and to give them the confidence to attempt the main task.
The last stage of the initial practice phase was a ‘test-conditions’ practice of ranking all nine versions of Neher’s validated Ensemble Width and Ensemble Depth stimuli (on two separate pages) without feedback, and with a notional limit of two minutes (see Section 5.3.2 for a discussion about this time limit). The benefit of this was that it allowed each subject to experience the time-pressure and complexity of the main task using validated stimuli, but not to pre-bias any particular subject to any particular test programme item. This proved to be a very valuable phase, as it allowed the clarification of the test procedure to at least one subject who had become confused and had not spotted that they had ranked two items in the same position.

5.3.4 Selection of Subject Groups

Recruitment of subjects was handled carefully. The author approached the student groups in person and (in order to maximise the motivation to participate in the study) explained the benefits of participation. These involved the gaining of critical listening experience, a chance to individually use the listening room and experience in the running of listening tests which would benefit them in their later studies.

The investigation was run with two separate intake phases of subjects – the first intake phase had a total of 18 subjects, the second had an intake phase of 33 subjects. Three subjects dropped out either during or shortly after ‘Task 1’ (citing study pressures), meaning that the total number of subjects available for the investigation was 48 individuals. All subjects were first-year students who were studying Sound Recording modules. There were a total of 42 males and 6 females.

Performance measures of rank accuracy and time taken per page were utilised in each intake phase in order to classify each subject’s performance. Three groups
needed to be separated in each intake phase to form the experimental and control groups.

It was important that all three groups had equivalent skills, not only to ensure that no single group had an abundance of highly or lesser skilled individuals, but also to allow changes in performance as a result of transferred learning to be easier to detect.

The performance of the subjects in ‘Task 1’ was examined by evaluating how their rank-ordering data matched the expected order, and how long they took to complete the tasks. The intention was to separate the subjects into three groups, each of which had approximately equal performance characteristics.

SSED and total time taken was calculated for each of the tasks (six pages of ED and six pages of EW) for each of the subjects. The summed totals of both SSED and timing were used as overall performance measures for each subject.

Groups were created in each intake phase by first attempting to balance total SSED then total time taken of three groups. For reasons explained in Section 4.3.2, gender balance between the groups was also achieved.

Once the three groups were created in each intake phase, they were randomly assigned to be the control group (Group 1), the practice group (Group 2) or the training group (Group 3).

Members of the control group (Group 1) were informed that they had not been randomly assigned to receive additional training, and were asked to report back for the next phase (which for them was “Task 8”) in three weeks.

Subjects in the practice and training groups (Groups 2 and 3 respectively) were individually informed the ‘good news’ (a motivational technique) that they had been randomly selected to take part in additional sessions. The practice group members were invited to six additional iterations of the familiarisation and ranking task phases outlined in Sections 5.3.1 and 5.3.2 (Tasks 2-7). The training group members were invited to participate in six additional sessions involving the training system described in Section 5.4.

5.4 Spatial Audio Attribute Listener Training System (SAALTS)

The Spatial Audio Attribute Listener Training System (SAALTS) used by the training group during the additional six sessions consists of three main elements:

- Tutorial
- Active learning using the Spatial Audio Attribute Toolkit (SAAT)
- Self-administered training drills with feedback

SAALTS is set within the context of the Simplified Scene-Based Paradigm for spatial audio scene description (see Chapter 3). It also implements changes recommended after the pilot study (see Chapter 4) and conforms with Alessi & Trollip’s model for successful instruction (Alessi and Trollip 2001), which has four elements (information presentation; learner guidance; practice and assessment).
Transfer Investigation

To match the six additional ranking sessions completed by the practice group, the training group was given six half-hour sessions. The first of these consisted of a tutorial (outlined in Section 5.4.1) followed by the introduction and use of the Spatial Audio Attribute Toolkit (SAAT) described in 5.4.2. The next four additional sessions allowed the trainees to use the SAAT and the training drills (detailed in Section 5.4.3). The final additional session was actually a copy of the practice group’s ‘Task 7’ ranking task (consisting of a familiarisation and ranking phase detailed in Sections 5.3.1 and 5.3.2), in order to provide an assessment target stage within the training system, and to compare the performance of the practice and training groups at the end of their additional sessions.

5.4.1 Tutorial

Presenting information was achieved through an individual tutorial administered by the author using a computer-based graphical presentation. The tutorial followed a similar procedure to that outlined in Section 4.4.2 and Appendix 7.2 (differences noted in the appendix). The need for a universal spatial audio description language was explained, the Simplified Scene Based Paradigm was presented and the concept of Scene Components was clarified. Visual analogies (see Section 2.3.1.6) were used to elicit responses from the trainees and Neher’s validated stimulus sets were used as audio examples (see Section 2.3.1.5). During playback of ensemble stimulus sets, trainees were asked to describe how each scene component changed considering both the individual sources as well as the ensemble that they were in (multi-source scene component). This mindful abstraction is particularly useful for far transfer (see Section 2.3.1.7).

5.4.2 Spatial Audio Attribute Toolkit (SAAT)

Guiding the learner was performed during the tutorial, and also during the Spatial Audio Toolkit phase and the drills phases.

The concept of scene components that can be individual sources or groups of sources had been described in the tutorial, and subjects were then given the chance to interact with the programme items used in the ranking tasks.

The Spatial Audio Attribute Toolkit (SAAT) is shown in Figure 73 and Figure 74, and is in the style of a constructivist learning environment (Alessi and Trollip 2001). This allows the learners to experiment with the various stimuli for SC Width (Figure 73) and SC Depth (Figure 74). All stimuli were looped when played back.

Programme items are selected in the SAAT using the “Stimulus” drop-down menu. Individual scene components (ISC) can be muted or soloed by selecting/deselecting the appropriate buttons – labelled QWER and YUIO in the two toolkits. The overall SC Width or Depth setting for each ensemble can be selected via a 9-point slider. Below each source button is a non-interactive slider that indicates the attitude of each of the ISC within the overall scene component (OSC). For the SC Depth page, the outer ISCs are portrayed as moving down the screen (labelled closer to the listener), as they appear in the simulation. The inner ISCs are portrayed moving up the screen (labelled further away from the listener). For the SC Width page the four ISCs are portrayed moving to the left or right of horizontal sliders (the sliders were marked from 1-9 to indicate the OSC width level).
The SAAT is actually an interface layer controlling Neher’s processing platform. The original single-source files used to create the new stimulus sets in Section 5.2.1 are sent through the processing platform and can be muted, soloed or changed at will. Presets for the nine different levels of the attribute could also be selected. This degree of control over the reproduction (especially the individual mute and solo functions) was only possible using real-time manipulation of the original source files within the processing platform. Using the SAAT, the subjects were able to not only imagine that the OSC was made up of four ISCs, but they were actually also able to listen to the ISCs individually, and build up the OSC using its constituent ISCs.

Being given control is a powerful motivating force for learners (see Section 2.3.2), but the main idea behind the SAAT is to provide a learning environment where the trainees can perform discovery learning (Alessi and Trollip 2001) initially guided by the author, but then eventually constructing their own knowledge using the toolkit.

The SAAT was presented to the learners immediately after the tutorial during their first training session, and was always made available at the beginning of subsequent learning sessions.
During the following learning sessions the learners would start by using the SAAT to re-acquaint themselves with the stimuli and concepts. They were left to use the system for as long as they felt comfortable before spending the remainder of the thirty minute sessions on the self-administered drills, explained in Section 5.4.3.

**5.4.3 Training Drills**

The *practice and assessment* phases of Alessi & Trollip's model for successful instruction (Alessi and Trollip 2001) are handled via a *self-guided test regime*.

Users were given full control over the difficulty and the type of test task as well as the attribute and stimulus set used. This was both because user control is an important factor in motivation (see Section 2.3.2), and because by giving the user control over their training, the type of experimenter-based error that occurred in Section 4.4.4 (relating to Subject 109's training course) would be avoided. Users were encouraged to find the appropriate difficulty level for themselves and if tasks were appearing too difficult, to try other tasks at a lower difficulty level.

Trainees would log into the computer (which maintained a user database), and administered their own tests using the interface shown in Figure 75. The administration page allowed the trainees to select the next drill to attempt, selecting the 'Stimulus' (programme item), 'Attribute' (either *SC Depth* or *SC Width*), 'Level' (of difficulty, from 1-4), and finally whether it was a 'Discrimination' or 'Pairwise Ranking' drill that was required. The 'Progress' button opened the panel of lights shown at the foot of the screen in Figure 75 - this reminded each subject which levels of difficulty they had already completed for the various combinations of options. They were instructed to challenge themselves to complete as many drills at as high a level as they could over the remaining training sessions.

Drills were loosely based upon those implemented in Section 4.4.3. However, due to the vast increase of available stimulus sets described in Section 5.2.1, far more choice was available for the trainees.
There were two types of drill: Discrimination (“are these the same or different?” – see Figure 76) and Pairwise Ranking (“which of these is wider/deeper?” – see Figure 77). Both involved the comparison of two items drawn from a randomly selected pool (Salisbury 1988). There were four different difficulty levels for each test. Difficulty Level 1 randomly selected from a pool containing just the most extreme stimuli in the set (stimuli 1 and 9). Difficulty Level 2 added stimulus 5 (the midpoint), Difficulty Level 3 added stimuli 3 and 7, and Difficulty Level 4 included all nine stimuli in the pool.

Stimuli were auditioned by clicking on the buttons labelled Q and W (see Figure 76 and Figure 77), or pressing the corresponding key on the keyboard. The “Select...” drop-down menu was used to provide a response. The green and red lights next to the drop-down menu indicated whether the answer was correct or not. The text below the two vertical sliders displayed the total number of trials attempted, the number of correct and incorrect responses and the total percentage of correct answers.
The two ‘pass’ criteria for the training drills were that at least 20 trials were attempted (in order to ensure that a given level of variety was attempted), with an 80% correct percentage (see Sections 2.2.2.6 and 2.2.3.4). The two vertical sliders tracked the trainee’s progress through the drill. The slider on the left would increase from trial 1 through to trial 20 (where it would stay as this criterion would have been fulfilled). The slider on the right displayed the current percentage of trials answered correctly. The pass criterion (80%) was shown as a visual marker.

If a trainee answered a question incorrectly the red button next to the ‘Select...’ drop-down menu would flash and the progress totals would update accordingly. However, the drill would pause to allow the trainee to listen back to the stimuli before selecting the correct answer. When a correct answer was given to a question that had already been answered incorrectly, the running performance totals did not increment – the next trial was simply displayed.

The ability to receive immediate feedback and be able to listen to the correct answer was one of the recommendations of the pilot study (see Sections 4.4.6 and 4.4.7), and is suggested in the literature (Alessi and Trollip 2001).

Several motivational devices were implemented in order to maximise interest and willingness to participate in the tests. As well as being given control, each user had their own ‘board’ of proficiency indicators (green lights), one for each difficulty level of each element of the tests (shown below the main window in Figure 78).
Subjects were challenged to complete as many tasks as they could, switching on as many lights as possible in the time available to them. The criteria for completing a task were that at least 20 trials needed to have been attempted, and 80% needed to have been correctly answered. It was possible for someone who had not achieved the pass mark after the 20th trial to continue the test until they increased their overall score to 80%. The 80% passmark has been carried over from previous studies (see Sections 2.2.2.6 and 2.2.3.4), and could be adjusted in the future if necessary. Progress was tracked with a numerical and graphical display showing the number of trials attempted and the percentage of correct answers given. Upon completion the user was rewarded with a window displaying a smiling face (Olive had also used such a feedback mechanism – see Section 2.2.2.6) and the corresponding proficiency indicator light was switched on (as shown in Figure 78).

Once both criteria had been fulfilled, the trainee was presented with positive feedback in the form of the ‘success screen’ shown in Figure 79.

![Figure 79: Screenshot of the success screen.](image)

CONGRATULATIONS!

YOU DID IT!

Speed of response should be enforced in order for it to improve through training (see Section 2.2.2.4). Therefore the familiar ‘traffic light’ system was utilised to mark the start of each test (green), the half-way point (yellow) and end point (flashing red). During training drills a trial was marked as incorrect if the user did not respond within 20 seconds (when the red lights began flashing). This was in order to enforce a similar time-pressure as was present in the ranking tasks (see Section 5.3.2).

The Spatial Audio Toolkit and self-administered drills were implemented using the Max/MSP programming language. Each user was automatically assigned a unique number along with their name. Their progress (which drills they had completed) was saved every time they completed a task. In addition most interactions that they had made with the software were logged for subsequent analysis. The training system was designed to be modular, so additional tests can be accommodated easily, and criteria can be adjusted if necessary.
5.5 ‘Real-World’ Stimulus Rating Sessions

Stimuli created using the process described in Section 5.2.2 were utilised in a far transfer task that would examine how subjects consistently and sensitively rated these stimuli in terms of ensemble width (EW was used because the rating tasks were performed by all of the subject groups).

To allow for three iterations of the test during a single 30 minute session, four programme items were selected (giving two minutes per page). Two were drawn from the training stimuli, and two were taken from the far-transfer ranking stimuli. All four featured distinct musical styles. All stimuli were looped when played back.

5.5.1 Familiarisation & Practice

Each subject took part in one far-transfer rating test. After reading through the test instructions (which also contained definitions of Ensemble Depth and Ensemble Width), the subjects were given the opportunity to familiarise (see Section 2.2.1.1) themselves with the stimuli and begin to place rate them using the 0-100 point scale. Each of the four programme items could be selected using a drop-down menu (which would randomly link the nine stimuli for that programme item to the QWERTYUIO buttons), subjects could use the nine sliders to assign a grading to each of the stimuli on the page. The now very familiar ‘traffic-light’ timer reset every time a new programme item (labelled ‘stimulus’) was selected. The familiarisation and practice screen is shown in Figure 80.

Figure 80: Screenshot of the rating task ‘familiarisation and practice’ screen.
Subjects were asked to listen through to all of the stimuli and formulate their own means of establishing what they perceived to be the width of the ensembles. To calibrate them to the widest and narrowest perceived \( EW \) they were told to listen through to all of the stimuli from the four programme items and practice rating each using the 0-100 point scale. As each slider moved, a number would be displayed beneath it corresponding to its rating on the 0-100 point scale.

Once subjects were happy to move on, they clicked the ‘finished’ button and began the rating task.

### 5.5.2 Rating Task

The rating task consisted of twelve pages, each with nine stimuli to rate on a 0-100 point scale. Within the task there were three randomised blocks of four pages (one page for each of the programme items to be evaluated), creating three iterations for each stimulus. Subjects were warned not to be perturbed if the same programme item appeared on subsequent pages (between two blocks of four pages, where the same programme item ended the previous block and began the next block). They were instructed to pay attention to the incrementing page number on the top right of the screen, and to the ‘traffic light’ timer which warned when one and two minutes had passed on each page. Figure 81 shows the rating task test screen.

![Figure 81: Screenshot of the rating task test screen.](image)

The ‘next’ button would appear, allowing progression to the next page once every slider had been displaced from the start position at the base (which guarded against accidental progression). A rating of zero could be given by moving the slider from the base and then back. This was explained to the subjects during the familiarisation phase. Once all twelve pages were completed, the computer saved the grading data.
along with a log of most of the interactions that they had made with the interface for subsequent analysis.

5.6 Analysis of Results

For Tasks 1-9, absolute differences between the subject-provided ranks and the expected (correct) ranks for each stimulus were used as one performance measure. Absolute rank differences (ARD) measured the rank ordering accuracy of the subjects – not the same as ‘SSED’ as used in the pilot experiment (see Chapter 4). Another performance measure was the time taken for each page of rank orders. The task had an explicit – important according to Quesnel (see Section 2.2.2.4) – but not enforced time limit of 120 seconds. Subjects were free to continue before the full 120 seconds had passed, or they could wait and finalise their answers. Overrunning was not directly penalised, but will appear as a higher mean of time taken (subjects were made aware of this drawback).

A number of questions were asked of the resulting data:

1. Did Groups 1, 2 and 3 have statistically equivalent performance during the pre-test prior to any practice or training?
2. Did each group change (improve) their performance between the pre-task and near-transfer task?
3. Did the control, practice and training conditions cause there to be differences in the way each group performed during the near-transfer task?
4. Did each group change (improve) their performance between the pre-task and far-transfer ranking task?
5. Did the control, practice and training conditions cause there to be differences in the way each group performed during the far-transfer ranking task?
6. Were there improvements for practice and/or training during the final additional task (Task 7)?
7. Did practice or training have more impact on improving performance in the final additional task (Task 7)?
8. Did the practice group show changes in performance throughout the practice tasks?

Performance in the ‘far transfer’ rating task (Task 10) will be addressed in Section 5.6.9.

Figure 82 and Figure 83 display the performance measures for the various groups for the ED and EW attributes for tasks 1, 8 and 9.

In Figure 82, Groups 2 and 3 (practice and training) show a marked reduction in mean absolute ranking difference (ARD) during tasks 8 and 9 with respect to their performance in Task 1, and have dramatically improved over the control group. Group 1 (control) shows a much more gradual reduction, and does not match the other groups performance during Tasks 8 and 9.

In Figure 83, Groups 2 and 3 (practice and training) show a marked reduction in time taken during ED Tasks 8 and 9 with respect to their performance in Task 1. Group 1
Transfer Investigation

(control) shows a much more gradual reduction, and does not match the other groups performance during Tasks 8. Groups 2 and 3 appear to improve during the EW tasks, but this improvement is less dramatic than in the ED items. The 120 second time limit is marked with a line.

Figure 82: Rank Accuracy

Graphs showing relative performances of rank order accuracy of the 3 different groups ('training' groups 1, 2 and 3).

Figure 83: Time Taken

Graphs showing relative performances on time taken of the 3 different groups ('training' values 1, 2 and 3).

5.6.1 Question 1 (Initial Inter-Group Performance Equivalence)

Did Groups 1, 2 and 3 have statistically equivalent performance during the pre-test prior to any practice or training?

In order to verify that each group had a similar initial performance level, univariate ANOVAs were performed on the Absolute Ranking Differences (ARD) and Time
Taken per Page (Time) for Task 1 to check for any initial differences between the groups.

The ANOVAs confirmed that there were no significant differences between the subject groups during the pre-test for either ARD or Time, and for either ED or EW. The groups were therefore equivalent at the outset, showing that they had been satisfactorily assigned into groups, and that initial differences were not the reason for the observed differences between the subject groups for the transfer tasks. Importantly, this shows that any differences between the groups during Tasks 8, 9 and 10 were a result of the training and/or practice.

5.6.2 Question 2 (Intra-Group Near Transfer)

Did each group change (improve) their performance between the pre-task and near-transfer task?

Paired samples T-Tests for each group, and with the ED and EW rank orders, were used to establish if each group had improved in performance between Task 1 and Task 8.

The training and practice groups improved dramatically compared with the control group, which indicates that training and practice have helped performance in the task.

5.6.2.1 ED Tasks

All three groups improved their ranking performance, but only the practice and training groups got quicker.

This shows that, for ED tasks, the subjects who did not do additional practice or training improved the way in which they ranked the stimuli. However additional training and practice improved both the training and practice group’s speed in completing the task. (This may indicate that the confidence of these subjects in their answers was improving, resulting in faster completion times.) Because of the improvement in both the training and practice group’s accuracy, it is unlikely that the faster times were a result of becoming bored or tired. If they were getting bored or tired, the group’s accuracy would reflect random rankings, thus worsening scores.

5.6.2.2 EW Tasks

All three groups got quicker, but only the practice and training groups improved their ranking accuracy.

This shows that, whilst getting quicker, the control group did not improve their accuracy. This is a sign that they were perhaps frustrated with the task and found it difficult. As previously discussed, the control group may have become tired and simply began to guess the ranks for the stimuli.

In fact, the practice group, who had a worse ARD mean during their pre-test, showed significant improvements for ranking EW in the near-transfer task. This is further evidence that practice improved performance and did not hinder it by making the subject bored of the task.
Furthermore, the training group showed similar improvements in their ranking accuracy and confidence.

5.6.3 Question 3 (Inter-Group Near Transfer)

Did the control, practice and training conditions cause there to be differences in the way each group performed during the near-transfer task?

This is similar to Question 2, but regards the applicability of training and practice to a greater extent. Both were found to have helped with performance versus the control group.

5.6.3.1 Ranking Performance

An ANOVA was carried out on ARD during Task 8, between the different training groups and for ED and EW. It showed significant differences in the subject groups for both ED and EW.

Bonferroni post-hoc tests showed that the practice and training groups were now significantly better than the control group (a genuine experimental effect). However, the practice and training groups were not significantly different to one another.

This shows that for ranking accuracy, both training and practice had a significant, positive effect on ranking accuracy over controlled conditions.

5.6.3.2 Time Performance

For ED and EW, a univariate ANOVA was performed on Time for Task 8, between the different training groups. It showed significant differences in the subject groups for ED, but no difference between groups for EW.

Bonferroni post-hoc tests showed that the practice and training groups were now significantly quicker than the control group for ED (a genuine experimental effect). However, the practice and training groups were not significantly different themselves on ED times.

This shows that for ED tasks the practice and training regimes were allowing those groups to get more accurate and faster than the control group. For EW tasks, the groups were not getting quicker, which may have been as a result of no improvement in confidence, or an understanding that EW was more difficult to accurately rank than ED (which was anecdotally the case), and that the allowable time was being used for additional checks. Whilst the practice and training groups were genuinely improving versus the controlled group, they were not getting any quicker.

5.6.4 Question 4 (Intra-Group Far Transfer)

Did each group change (improve) their performance between the pre-task and far-transfer ranking task?

Independent samples T-Tests for each group, and with the ED and EW rank orders were used to establish if each group had improved in performance between Task 1 and Task 9.
5.6.4.1 ED Tasks
All groups showed improvements in both ranking and time. All conditions were shown to have a positive effect versus the initial task. The control group’s improvement in time taken here, rather than between Tasks 1-8 can be attributed to the extra practice at the task that they received during Task 8 before they attempted task 9 (between Tasks 1 and 8, however, they had no intervening practice).

5.6.4.2 EW Tasks
All groups got quicker, but only the practice and training groups got more accurate. This follows a similar pattern to that in Section 5.6.2.2.

5.6.5 Question 5 (Inter-Group Far Transfer)
Did the control, practice and training conditions cause there to be differences in the way each group performed during the far-transfer ranking task?
It was found that the practice and training groups were able to transfer gained experience to a different task.

5.6.5.1 Ranking Performance
An ANOVA was carried out on ARD during Task 9, between the different training groups and for ED and EW. It showed significant differences in the subject groups for both ED and EW.
Bonferroni post-hoc tests showed that the practice and training groups were now significantly better than the control group (a genuine experimental effect). However, the practice and training groups were not significantly different themselves.
This shows that for ranking accuracy, both training and practice had a significant, positive effect on ranking accuracy over controlled conditions. It also mirrors what happened in Task 8 (see section 5.6.3.1).

5.6.5.2 Time Performance
An ANOVA was carried out on Time during Task 8, between the different training groups and for ED and EW. It showed no difference between the groups for ED or EW.
This shows that all groups had similar time performance levels during Task 9.

5.6.6 Question 6 (Intra-Group Improvement: Practice vs. Training)
Were there improvements for practice and/or training during the final additional task (Task 7)?
Paired-samples T-Tests were used to verify that improvements had occurred in both training and practice groups between Tasks 1 and 7. All conditions improved, meaning that the practice and training regimes had caused significant increases in ranking accuracy, and decreases in time taken.
5.6.7 Question 7 (Inter Group Improvement: Practice vs. Training)

Did practice or training have more impact on improving performance in the final additional task (Task 7)?

Independent samples T-Tests were conducted on the ARD and Time data for the training and practice groups for ED and EW trials in Task 7. The practice and training groups showed similar performance in everything but the rank accuracy of the ED attribute, where the practice group showed a significant improvement over the training group. This could be explained as an effect of familiarity with the task, as by the time the subjects took part in Task 8, the training and practice groups were equivalent in terms of ranking accuracy (see section 5.6.3.1).

5.6.8 Question 8 (Practice Performance)

Did the practice group show changes in performance throughout the practice tasks?

Data was available that charts the progress of the practice group through Tasks 1-7, which provides a useful study of the effect of repetitive practice in a spatial audio evaluation task. This could be compared with Bech’s study on repetitive practice in loudspeaker quality tests outlined in Section 2.2.1.3.

A univariate ANOVA was carried out on the practice group’s results during the pre-task and the practice sessions. Every condition (ED/EW and ARD/Time) were shown to improve significantly. These improvements can be described as linear (according to Tests of Within-Subjects Contrasts).

Closer inspection of Figure 84 and Figure 86 shows that rank accuracy performance levelled off for ED and EW after Task 5 (the fourth practice session). This could point to potential time-savings using targeted practice regimes over more generalised training systems, if interest is in improvement in performance of one specific task. A combination of a general training scheme with additional practice could also prove to be optimal.

Figure 85 and Figure 87 seem, however, to show a more-or-less constant reduction of Time in keeping with their linear description.

In summary, both performance measures improved throughout practice, but a case could be made for stopping the practice sessions after four practice sessions (not including the pre-task) without adversely affecting ranking performance. This concurs with Bech’s finding that asymptotic performance (in his case with loudspeaker tests) was achieved with about four practice sessions (see Section 2.2.1.3).
Figure 84: ED ARD over the practice sessions

Estimated Marginal Means of MEASURE_1

Attribute: ED

Figure 85: ED Time over the practice sessions

Estimated Marginal Means of MEASURE_1

Attribute: ED
Transfer Investigation

Figure 86: EW ARD over the practice sessions

Estimated Marginal Means of MEASURE_1

Attribute: EW

Figure 87: EW Time over the practice sessions

Estimated Marginal Means of MEASURE_1

Attribute: EW

- 205 -
5.6.9 Task 10 (‘Further’ Transfer)

The resulting data for Task 10 were analysed using methods including those outlined in Section 4.5.2. However no statistically significant differences were found between the four subject groups (Control, Practice, Trained and Experienced). In addition, the experienced listeners in Group 4 were not particularly consistent or sensitive during Task 10.

The most likely reason for these outcomes is that the rating task in Task 10 (described in Section 5.5.2) was too difficult, as had been the case in the pilot investigation (see Section 4.5).

In planning the investigation and selecting the task and stimuli for the rating task (see Section 5.2.2), it was considered preferable to create a task that was potentially too difficult than one that was potentially too easy. For example Merimaa & Hess’ study was most likely ‘too easy’ for reasons detailed in Section 2.2.3.3, and Neher’s training experiment outlined in Section 2.2.3.4 could also be accused of this (one of his training subjects got a perfect score in the pre-training tests). The ability to discern attributes in confusing situations is more important than acuity according to Meilgaard et al. (1991), and even if the task was overly difficult it could have left room to demonstrate performance differences between the groups, and hence potential transfer effects. In addition it was suggested in Section 2.3.1.3 that practising an easier task may sometimes facilitate better performance in a subsequent task than training on the task itself.

Another possible reason that the far transfer rating task did not demonstrate any differences between the groups was that it was too far. In changing the stimuli, task and particularly the scale used in the tests it is possible that too many concepts were being changed, so-much-so that the original training was not relevant any more. Ellis had recommended that the training and target tasks are as similar as possible to maximise transfer (see Section 2.3.1.5). However, if the rating task was also implemented in a pre/post manner, there would have been considerable biasing of the participants as to the nature and goal of the study.

Notwithstanding, the experienced listeners did not distinguish themselves significantly from even the control group, which means that further transfer of either the SAALTS or repetitive practice regimes has not yet been demonstrated to situations and stimuli that are different from the learned one.

It is possible that results could have been different if the recordings detailed in Section 5.2.2 had been made in a different recording environment, but further research would be needed to verify this.

5.7 Transfer Investigation Summary

A Spatial Audio Attribute Listener Training System (SAALTS) was developed based upon previous audio training studies (outlined in Section 2.2) employing motivational techniques (covered in Section 2.3.2) in order to optimise transfer (see Section 2.3.1). SAALTS addresses a number of concerns raised by the pilot training investigation described in Chapter 4.
Chapter 5 has described a study that compares the performance of subjects trained using SAALTS with that of repetitive practice (recommended in the standards and in previous studies) and a control group. A variety of transfer tasks were utilised to assess both near and far transfer.

A pre/post test methodology ensured that transfer effects could be isolated between experimental and control groups, and the relative impact of the training and practice regimes was able to be studied.

As a result of the previous (pilot) training investigation, a number of recommendations for SAALTS were delineated in Section 4.4.7. The need for an increased variety of training stimuli was addressed by the utilisation of two spatial audio attributes, the creation of twelve new ranking stimulus sets for each of them (Section 5.2.1) and the creation of four complex SC Width rating stimulus sets (Section 5.2.2). All stimuli described in Section 5.2 were looped when played back in accordance with the recommendations. Users were given full control of their learning experience (although consultation with the author was always available), and were given access to familiarisation elements during each session. The feedback system within SAALTS minimised the demotivating negative responses found in the system implemented in Section 4.4, and allowed incorrect trials to be auditioned without time pressure in order to learn from their mistakes. The recommendation that information sheets be made available for further study was not implemented in this evaluation study to reduce potential cross-contamination of the subject groups, however subjects were allowed a more flexible way to plan out their timetable of sessions in accordance with the recommendations. To reduce the course-differences found in the pilot investigation (see Chapter 4), all subjects involved in the transfer investigation were enrolled on courses studying 'Tonmeister' sound-recording modules at the University of Surrey.

Because of the increased effect size of the extended subject pool, and success in the division of subjects into similar-skill, randomly assigned experimental groups, various transfer effects were able to be shown that were not clear in Chapter 4.

Near transfer was shown to be superior for the practice and training groups over the control group in both ranking accuracy and time taken. The training and practice groups had similar levels of near transfer though. This is surprising as it was hypothesized that targeted practice on a specific task with specific stimuli would prove to be more beneficial for the task itself than a more generalized training programme. SAALTS, however, produced near transfer that rivalled the practice regime.

Far transfer to different stimuli was also shown to have occurred for the practice and training groups to a similar extent (and once again, both improved much more than the control group). It was hypothesized that a more generalised training programme would be better at transfer away from the initial situation. The far transfer task can be seen as 'relatively near' in that only the stimuli were changed, and to stimuli that were similar and changed in similar ways. Bech suggested that repetitive practice can aid transfer to different stimuli (see Section 2.2.1.3).

Far transfer to situations and stimuli different to the learned ones was not demonstrated in this experiment. It was suggested that the far transfer rating task was too different from the ranking training to have demonstrated a transfer effect. However, since a group of experienced listeners was unable to demonstrate a
Transfer Investigation

distinguishable performance level from a group of naïve listeners, it is most likely that the transfer rating task was simply too difficult for all groups to have a chance to consistently and sensitively judge the stimuli.

Time taken is a tricky performance measure to analyse, as subjects who were 'recklessly' performing tasks as quickly as possible might show fast, or improving time-taken, whilst potentially getting worse in terms of their ranking accuracy. Time-taken data should therefore be analysed with the other performance measures taken into consideration.

Whilst the control group's subjects frequently got quicker without improving their accuracy, both the training and practice groups always improved their accuracy, and other than EW far-transfer tasks, always got faster at responding.

The benefits of using either practice or training over no additional regime for spatial audio evaluation are clear: subjects are not only better judges, they appear to be more confident in their own judgements.

The fact that SAALTS was not as targeted towards the particular task as the dedicated practice regime but yet achieved similar near and far transfer could have significant consequences. If a certain training scheme was made general enough to apply to many different stimuli or tasks, subjects could be trained effectively for a variety of listening tasks in a fraction of the time it would take to allow them to repeatedly practice each task individually.

A potential explanation for the apparent far transfer shown by the practice group during Task 9 is that this was dominated by procedural learning (of the task) over the perceptual learning (of the concept of the spatial attribute in question). There is evidence in Figure 84, Figure 86 and Figure 87 that an asymptotic level of performance was being approached by the training group, as described by Bech (see Section 2.2.1.3) and Drennan & Watson (see Section 2.2.1.7). If this were the case, then it is possible that there are other transfer tests that are not so 'far' as Task 10, but feature a task other than that on which the practice group had gained their procedural knowledge. Such tests offer the potential of demonstrating enhanced transfer for the training group (which had been trained in a more generalisable way) over the practice group.

Regarding time issues, any time savings that can be achieved with minimal detriment to sensory acuity are a benefit to scientific research (with frequently limited resources and time constraints) and industry (where “time is money”) alike. If practice sessions can be stopped early once an asymptotic performance level has been achieved (see Section 2.2.1.3), or a single general training programme is used instead of several practice phases, potential time savings could be enormous.

As a result of this study, it is possible to recommend SAALTS as a viable method for producing near transfer and certain degrees of far transfer for spatial audio attribute assessment, because it produced results that were comparable with repetitive practice. Whilst far transfer to a different situation and stimulus set was not demonstrated, the literature suggests that a more generalised method such as SAALTS would produce better and further transfer than repetitive practice of spatial audio attribute tasks. An optimal combination of practice of spatial audio evaluation tasks and training using a system such as SAALTS could potentially be found to maximise near and far transfer.
These conclusions can be reliably limited to spatial audio ranking tasks involving 'ensemble depth' and 'ensemble width' attributes, however there is nothing to suggest that other elements of the SSBP could not be incorporated into SAALTS successfully. The ensemble depth simulations, for example, utilise individual source distance changes of the individual sources within the ensembles, since ensemble depth was successfully trained, it can be expected that SC Distance (of a single source, using the SSBP nomenclature – see Section 3.2.4) could also be successfully implemented in SAALTS.

Further work would need to be undertaken to fully assess the generalisability to other types of task and other attributes. As repetitive SAALTS and repetitive practice have been shown to be successful when compared with a control group, it would also be possible, as a result of this research, to conduct future investigations without a third (control) group. This would allow more of the available subjects to be utilised in the comparison between future implementations of SAALTS and repetitive practice.

Some of the content of this chapter has been published previously in (Kassier, Brookes and Rumsey 2006a, 2007).
6 CONCLUSIONS & FURTHER WORK

This Thesis described the background, motivation, development and testing of a Spatial Audio Attribute Listener Training System (SAALTS).

Before spatial training could begin, a suitable description language for spatial audio was required. Previous studies involving spatial audio evaluation were covered under Section 2.1. Although many of these studies utilised spatial audio attributes, only Rumsey’s Scene-Based Paradigm (detailed in Section 2.1.8) provided a rigorous approach for describing spatial audio scenes. However, Rumsey’s Scene-Based Paradigm contained a number of terms that would have caused complications when trying to implement them in a training system. For example, ensemble attributes would have conceptually overlapped with any individual source attributes. These issues were discussed in detail in Chapter 3, which describes the development of a Simplified Scene-Based Paradigm (SSBP). The SSBP is a descriptive language that is based on previous studies but optimised for use in training systems. The previously-used concept of Scene Components is utilised to solve the issues relating to overlapping source-based spatial audio attributes. The SSBP focuses on non-overlapping dimensional attributes drawn from previous studies, and can be recommended for the description of a wide range of spatial audio scenes for normative, product evaluation or training investigations. The Simplified Scene-Based Paradigm (SSBP) contains the following elements:

- Source-related attributes:
  - Scene component direction
  - Scene component distance
  - Scene component width
  - Scene component depth

- Environment-related attributes:
  - Environment width
  - Environment envelopment

Once this paradigm was established, the training programme could be developed and optimised based upon previous studies. A method to establish how to check that spatial audio attribute training would be useful outside the original training context was also considered.

General training studies, the various timbral ear training systems, and the four published studies containing specifically spatial audio-based training were examined in Section 2.2 with a view to informing a method of training for spatial audio attribute listening skills.

Two important concepts in the learning literature were isolated as being of particular interest to this study. Firstly, Section 2.3.1 covered issues relating to Transfer of Training. Secondly, Section 2.3.2 covered the concept of motivation, and combined the recommendations of two theories of motivation in learning to inform the training system.

A pilot study to ascertain the effectiveness of a spatial audio attribute training system based upon the training of ranking tasks, and its transferral to tasks involving the
rating of spatial audio attributes is described in Chapter 4. As a result of the pilot study, it can be concluded that it is possible to train naïve listeners in the concept and judgement of spatial audio attributes as outlined in the SSBP, but that this training only transferred as an increase in the range of the scale that was used by a (potentially more motivated) sub-set of the trained listeners. Whilst the pilot study utilised a control group (as had Kirk and Neher – see Sections 2.2.1.2 and 2.2.3.4 respectively), this control group did not control the training received within the training phase, as it only took part in the spatial audio rating tasks. The feedback system was also deemed to be too de-motivating.

A further training investigation (reported in Chapter 5) was designed to incorporate a control group for the training condition, and also to allow comparison between a generalised training system and the method of repetitively training the task which had been suggested by the standards and in previous work. Informed by the pilot study and the literature on transfer and motivation, the resulting Spatial Audio Attribute Listener Training System (SAALTS) employs the following elements:

- A tutorial explaining the SSBP and its importance
- Active Learning using the Spatial Audio Attribute Toolkit (SAAT)
- Self-guided drills featuring a feedback system involving “learning from your mistakes”, no de-motivating effects, and a persistent progress indicator that encourages over-achievement

The transfer investigation, like Kirk’s study (see Section 2.2.1.2), compared the performance of two groups with different experimental treatments (in this case, the practice and training groups), with a control group.

Forty-eight subjects, recruited in two experimental phases, were pre-tested using a task based upon the ranking of nine levels of spatial audio attributes that had been carefully simulated using an existing spatial audio processing platform. Performance measures for each subject were analysed and the subjects were separated into three equal skill groups during each experimental phase. These groups were randomly assigned to each of the experimental conditions. As these groups had equivalent skills at the outset, any changes between them during later ranking tests would be due to the intervening treatments.

The training group was trained using SAALTS whilst the practice group performed further iterations of the initial rank ordering task. The last of the six additional training sessions for the training group involved an iteration of the initial ranking task in order to provide an evaluation target phase for SAALTS, but also in order to compare the extra training sessions directly with the practice group.

The practice group seemed to have assimilated the majority of procedural and perhaps perceptual knowledge after four practice sessions (not including the pre-practice iteration). This concurs with Bech’s finding (see Section 2.2.1.3) that asymptotic performance (in his case with loudspeaker tests) was achieved with about four practice sessions.

All three subject groups were tested with an iteration of the initial ranking task (in order to test for near transfer), another iteration with a new stimulus set (in order to test for far transfer to different stimuli), and a separate rating task (to test far transfer
to different situations and stimuli). Both near and far transfer were shown to have been provided by both the practice and training regimes.

Far transfer to different situations and stimuli was not demonstrated in the transfer experiment. Whilst it is possible that no far transfer occurred, it is more likely that the selected transfer task (rating stimuli recorded using a novel and repeatable multi-microphone and sound source reproduction system) was too difficult, as even experienced listeners were not able to distinguish themselves from naïve and untrained listeners performing the task.

As a result of the research described in the transfer investigation, it can now be concluded that training in spatial listening can improve performance in spatial audio attribute evaluation tasks, and that this training transfers to similar tasks with similar stimuli. It was also found that the performance of SAALTS was comparable to a repetitive practice regime for the target task.

These conclusions can be reliably limited to spatial audio ranking tasks involving Scene Component Depth and Scene Component Width attributes, however there is nothing to suggest that other elements of the SSBP could not be successfully incorporated into SAALTS.

Regarding the use of Scene Component Depth and Scene Component Width attributes, Neher’s simulations of the two ensemble attributes utilise changes in both ISC Distance and ISC Direction. The use of these more complicated attribute simulations allowed these subsidiary attributes to be focused upon and discussed in the tutorial phase (see Section 5.4.1) and isolated in the SAAT (see Section 5.4.2), allowing for all source-related attributes (see Section 3.2.4) of the SSBP to be covered. As was discussed in Section 5.2.1, ISC Distance was informally felt to be potentially too easy to rank in isolation. ISC Width was found to be a very difficult attribute to simulate (Neher 2004), and none of Neher’s subjects was able to rank the examples correctly – even after training (see Section 2.2.3.4). Judging from the results of the transfer study OSC Depth was easier to rank than OSC Width, but the attribute tasks were both challenging and achievable after practice and/or training. A suitable compromise appears to have been arrived at where the selected attributes SC Width and SC Depth have allowed for all source-related attributes to be discussed and identified, the tasks have been challenging enough to require training and/or practice, and not too difficult that they were within reach after practice and training. Additionally the two attributes provided a range of difficulty within the training as well.

The Spatial Audio Attribute Listener Training System (SAALTS) described in Section 5.4 was found (in Section 5.6) to produce similar near and far transfer during ranking sessions as that achieved by a group that repetitively practiced the target task. Both regimes were found to be beneficial as they out-performed a control group.

It is recommended therefore, that in order to save time and maximise the potential for future spatial audio transfer experiments two groups should be utilised: a training and a practice group. In this way the practice group would become a type of control group with respect to the training group. Future experiments could be targeted towards finding the differences between the performance enhancements afforded by the training and practice regimes, or towards finding the optimal combination of practice and training to maximise task-based learning whilst maintaining
Conclusions & Further Work

generalisability. A further investigation direction could examine the asymptotic performance increase found in the practice groups, especially if this is mirrored in the training groups or if these can improve performance beyond that of the asymptotic level of the practice groups.

Far transfer from ranking to rating tasks with different stimuli was not demonstrated in either the pilot or the transfer investigations. Piloting of far transfer rating tasks in order to optimise the difficulty level is strongly recommended.

The study presented in this thesis concentrated on two spatial audio attributes: SC Width and SC Depth. Future studies should begin to implement the other spatial attributes within the Simplified Scene-Based-Paradigm (SSBP). Certain attributes promise to be straightforward, for example SC Distance. The SC Width of single-source Scene-Components has already proved to be a difficult attribute to simulate and judge in (Neher 2004) and Chapter 4, however, the multi-source simulations used in the ranking tasks in Chapter 5 could continue to be used instead to simulate SC Width in future versions of SAALTS.

Section 3.2.4 suggested that the possible ambiguity relating to whether scene component width should be considered to be perpendicular to the egocentric line or in a circular arc about the listener (or both/either), could be resolved using a further normative experiment. In addition, experimental verification of the SSBP would be useful, but the author has not yet found a viable test paradigm.

With hindsight the inclusion of a within-training evaluation-phase ("Task 7") for the training group was not as useful as had been hoped. The recommendation for future implementations of SAALTS transfer experiments would be not to have an additional evaluation test before the near transfer session ("Task 8").

If procedural learning was dominating in Task 10, it would be possible to devise other transfer tests that are not so 'far' as Task 10, but feature a task other than that on which the practice group had gained their procedural knowledge. Such tests offer the potential of demonstrating enhanced transfer for the training group (which had been trained in a more generalisable way) over the practice group.

Implementation of tasks other than discrimination and pairwise ranking (both of which are suitable for the training of listeners to make a series of comparative judgements) within the training system would be useful. Of potentially great benefit would be the inclusion of a grading scale (similar to the 0-100 point scale used in the pilot experiment and Task 10 of the transfer investigation. Crude training stimuli could be easily created using uneven levels of the nine different attribute levels in the Spatial Audio Attribute Toolkit (SAAT) element of SAALTS.

In summary, further work should concentrate three phases. The first is the identification of the extent of the far transfer potentially afforded by SAALTS over repetitive practice, but not sought in the transfer investigation described in Chapter 5. The second is in attribute simulations and verification of training for other elements of the SSBP using SAALTS. Finally, the ultimate goal would be to create a system that had the optimal combination of repetitive practice elements and generalisable training that would benefit academic normative evaluation studies and commercial product evaluation or training tasks.
7 APPENDICES

The appendices contain information not contained within the main thesis, but considered to be useful enough to be added to it.

7.1 Reproduction from (Kassier, Lee, Brookes and Rumsey 2005)

The following pages contain information regarding the microphones used and their set-up published in Sections 3 & 4 of (Kassier, Lee, Brookes and Rumsey 2005) and reproduced here (with altered section and figure numbers) for convenience.

7.1.1 Selected Microphone techniques

In order to determine the feasibility of the methods chosen for location recording, it was important to maximise the number of different simultaneous multichannel recordings possible when recording 24 channels of audio. 24 recorded channels is the maximum number that can easily be accommodated on location (a 24-track hard disk recording unit or 3 synchronously ganged 8-track tape recorders).

A method was devised to allow 16 different 5-channel recordings to be created simultaneously using 24 channels of audio. By recording four different 3-channel ‘front’ microphone arrays and four different ‘rear’ microphone arrays (two with 2-channels, two with 4-channels), it was possible to combine one of the ‘front’ techniques with one of the ‘rear’ techniques to create sixteen different configurations for reproduction using 3/2 stereo.

Limitations in the quantity and directivity of microphones available during the recordings made it impossible to use 24 similar microphones or microphones of the specified directivity for all of the techniques selected. Attempts were made to use the same microphone types within each array, as far as availability allowed. Comparison of the techniques themselves will need to be treated with caution. Ideally, the recordings should be made with similar microphones of the specified directivity (for example, 24 Schoeps CMC-5U microphones with appropriate directivity capsules) for the comparisons between techniques to be as fair as possible.

The recording angle of a microphone array is the angle (subtended around the centre line) between the left and right most edges of the sound stage that appear in the left and right edges of the reproduction (Herrmann and Henkels 1998). If the recorded sound stage subtends a wider angle about the microphone array than the array’s recording angle, the sources outside the angle will all be reproduced (bunched) in the left or right speakers respectively.

The four ‘front’ triplets and four ‘rear’ recording arrays used in the recordings are described below. The recording angles of the first three front techniques were similar to each other (108°, 118°, and 120°), while that of the last technique was much larger than the others. The ‘front’ arrays were centred at the same location to allow the stereophonic scope of the phantom images produced by the first three
techniques to be similar in order to make a controlled comparison between the techniques.

7.1.1.1 Front Microphone ‘Triplet’ Techniques

All three-channel ‘front’ techniques described here use a triplet of microphones that are subsequently used routed to L, C & R in the 3/2 stereo configuration.

Fukada Tree

The Fukada Tree technique is a modification of the “Decca Tree” stereophonic recording technique. The Fukada Tree replaces the omni-directional microphones of the Decca Tree with cardioid directivity pattern microphones in order to reduce the non-direct sound energy in the front channels (Fukada, Tsujimoto and Akita 1997). The configuration of this technique is shown in Figure 88.

![Figure 88: Fukada Tree](image)

The centre microphone faces forward, outer microphones face away from one another, 90° from the centre-front line. The widely spaced outer pair should produce a large interchannel time difference, providing a good sense of ‘spaciousness’ and ‘openness’ (Theile 2001), while the centre microphone should provide the ‘articulation’ of the stereo image (Streicher and Everest 1998). There is, however, a potential problem in localisation of sound sources, as there is a strong precedence effect triggered between L & C, or C & R due to the long distance between each microphone. Therefore, it is difficult for the Fukada Tree to achieve a balanced distribution of the phantom sources although there are three solid localisation areas (around the three front loudspeakers) that can be obtained with this technique.

The Fukada Tree was implemented in this study using three AKG C414 B-ULS microphones set to cardioid directivity as specified. The recording angle of the Fukada Tree is 108°.

OCT-Inspired Technique

Theile (2001) proposed a front microphone technique called Optimal Cardioid Triangle (OCT), which is optimised with regard to interchannel crosstalk. It is suggested by Theile that the crosstalk between channels must be reduced as much as possible in order to obtain accurate localisation characteristics. OCT employs a cardioid centre microphone just 0.08m in front of two outer supercardioid directivity microphones. The outer microphones are faced towards the sides in order to obtain
maximum channel separation. The recording angle is adjustable depending on the spacing between the outer microphones, and this flexibility can be important for recording engineers to have freedom of microphone array placement, to control direct/indirect sound balance and also to create sound colour (Theile 2001).

The OCT-inspired technique used in this experiment is shown in Figure 89. It employs a cardioid capsule for the centre microphone and hypercardioid capsules for the outer microphones. Three AKG C414 EB microphones were used to implement the OCT-inspired technique in this experiment.

For this experiment, the distance between the outer microphones was chosen to be 0.7m in order to achieve a recording angle of 118°.

**INA-3 Technique**

The INA-3 technique (Herrmann and Henkels 1998) is based upon the ‘critical linking’ technique, proposed in (Williams and le Du 1999, 2000). ‘Critical linking’ intends to attach the left (L-C) and right (R-C) segments of the reproduced frontal sound image without overlap, and thus aims to provide a balanced and continuous presentation of the reproduced sound image across L-C-R in the 3/2 stereo configuration. This ‘critical linking’ is achieved by using either ‘electronic offset’ or ‘microphone position offset’. The electronic offset is created by adding a certain value of intensity difference or time difference to the time and intensity function. The microphone position offset is achieved by changing the physical position of the microphones to adjust the time and intensity differences of the array. It is suggested that the array must be placed so that the outer microphones point to the edges of the recording stage in order to obtain the full spread of the stereo image, provided that the centre microphone points to the centre (Herrmann and Henkels 1998).

Figure 90 shows the configuration of INA-3 technique used in this experiment.
The angle between the outer microphones (and hence the recording angle) for the INA-3 array used for the current experiment was 120°. The INA-3 technique was implemented in this study using three AKG C451 microphones with cardioid directivity capsules (as specified).

Near-Coincident-Inspired Technique

Klepko (1997) proposed a near-coincident front triplet, which consists of three microphones placed in line with a distance of 0.175m between each adjacent microphone. The centre microphone should be directed forwards with the outer microphones angled at ±30° from the centre-front line. See Figure 91.

Klepko (1997) suggests the need to avoid producing a strong phantom centre image between the left and right channels since there is already an additional centre microphone. For this reason the outer channels employ a super-cardioid directivity pattern microphone, while the centre channel uses a cardioid microphone. However, as (Theile 2001) points out that this technique suffers from a serious interchannel crosstalk problem despite the use of supercardioid microphones. Theile also affirms that there is a large and inevitable overlapping between the recording area L-C and C-R because the recording angles of each microphone pair are wide due to the small angle between the microphones and the use of a more directional polar pattern on one channel. The most dominant effects of interchannel crosstalk are an increase in source width and a decrease in locatedness according to (Lee and Rumsey 2005). The wide recording angle may also result in a narrow stereo image when using a normal microphone distance from the stage.

The near-coincident-inspired technique was implemented in this study using three Neumann KM84 cardioid directivity microphones (as no supercardioid microphones were available for the outer channels). This would have probably given rise to
stronger interchannel crosstalk, thus reproducing wider and more poorly localised sound sources than would have been the case if supercardioid directivity microphones had been available.

7.1.1.2 4-Channel ‘Rear’ Microphone Techniques

In four-channel ‘rear’ microphone techniques, four microphones are used to record diffuse reverberation. The signals from these microphones are generally reproduced using the L, R, LS and RS speakers in a standard 3/2 stereophonic loudspeaker configuration.

Hiyama et al. (2002) showed that four loudspeakers in the L, R, LS & RS positions in 3/2 stereo could reproduce a spatial impression close to that of a twelve loudspeaker encircled configuration (when reproducing decorrelated reverberation). The use of four channels should therefore be beneficial when compared to the two-channel techniques employed in five-channel main microphone arrays.

IRT-Cross-Inspired Technique

Theile (2001) proposed a four-channel rear microphone array called ‘IRT-Cross’. It consists of four (normally cardioid) microphones arranged in a square of side 0.2m to 0.25m wide, with each microphone at the corner of the square pointing away from the centre. This array is optimised for recording ambience, but can be disadvantageous with regard to crosstalk from the direct sound (because the front two microphones facing towards the front corners may not have a sufficiently suppressed direct sound pick-up). Theile (2001) suggests that the spacing between the microphones can be decided depending on the recording situation and the desired characteristics of spatial image, although he recommends the distance of 20-25cm. Closer microphone spacings provide a more balanced distribution of enveloping sources, whilst wider spacings provide more diffused reverberation. Extreme spacing of either too close or too wide causes a loss of envelopment (Theile 2001). The polar pattern of the microphones can be also chosen depending on the situation.

The implementation used in this experiment used a microphone spacing of 30cm to allow the microphones used in the recordings to fit together in the cross arrangement. See Figure 92.

Figure 92: IRT-Cross-Inspired Technique

Due to a shortage of available microphones of suitable directivity and quality, the IRT-Cross-inspired technique implemented in this experiment used a pair of Oktava Mk-012-01 cardioid pattern microphones pointing towards the front left and front
right, and a pair of AKG C460B microphones with CK61-ULS cardioid pattern capsules pointing to the back left and back right.

**Hamasaki-Square Technique**

Another four-channel rear microphone array is the ‘Hamasaki-Square’. It employs dipole microphones pointing to the left or right of the centre-line so that their dead-axes are facing forward. This is in order to reduce the crosstalk from the direct sound as much as possible. The distance between each microphone was originally suggested to be 1m (Hamasaki 2000), but this was later adjusted to 2-3m based upon calculation and subjective listening tests (Hamasaki and Hiyama 2003). Their calculation of cross-correlation-coefficient between two omni directional microphones in the reverberant field showed that the distance of 2m provided decorrelation above 100Hz, which seem to fulfil the requirement for the perception of spatial impression. They also conducted a subjective listening test in order to compare the spatial impression between each pair of 1m, 2m and 3m distances, and found that most of the listeners participated in the test preferred 3m to 2m, and 2m to 1m. The array is usually placed far away from the sound stage and at a high position in the recording space in order to obtain the maximum ratio of reverberant to direct sound. Theile (2001) suggests that this array is a better option for achieving good spatial impression compared to the IRT-Cross. The pair of microphones furthest towards the front are routed to channels L and R or panned between L-LS and R-RS, and the pair of microphones furthest towards the rear are routed to channels LS and RS. The degree of L-LS or R-RS panning is dependent on the amount of desired spatial information in the front loudspeakers, and also seems to rely on the headroom of spatial image in the front array that is used in combination.

The Hamasaki Square configuration implemented in this experiment is shown in Figure 93.

![Figure 93: Hamasaki-Square Technique](image)

It uses four Schoeps CMC-5U microphones with dipole directivity capsules. The positive lobes of the dipoles faced away from the centre of the array. The recorded signals were routed to L, R, LS & RS in the 3/2 stereo reproduction.

**7.1.1.3 2-Channel ‘Rear’ Microphone Techniques**

2-channel rear techniques normally route the signals from two microphone channels to the LS and RS loudspeakers in the 3/2 stereo loudspeaker configuration.
Appendices

**Dummy Head Technique**

Klepko (1997) proposed using a dummy-head binaural microphone in order to provide a 'continuous' lateral spatial impression. He affirms that the problems of high frequency acoustical crosstalk that are present when the binaural signals are reproduced through the loudspeakers are solved naturally when the dummy head is used for the rear channels. As the rear loudspeakers are placed almost at the sides of the listener, the listener's head acts as a diffracting barrier to frequencies above 1kHz, which carry the most effective HRTF cues. The maximum crosstalk rejection is achieved when the rear loudspeakers are positioned exactly at ±90° of the listener, where the maximum differences between the ear signals are produced. In the listening test using the dummy head microphone coupled with the 'near-coincident front' triplet described above, Klepko found that continuous and clear spatial image were created between ±30° and ±90°. Klepko stated that the distance between the front triplet and the dummy head was 1.24m, but the reason of the spacing is not explained. Despite the acoustical crosstalk rejection between the rear channels by the head shadow effect, the interchannel crosstalk relationship between the front channels and the dummy head is likely to be poor. Since the dummy head is facing the front and the distance from the front array is relatively short, the crosstalk from the direct sound will have large intensity and short time delay (about 0.38ms). This might be critical with regard to accurate localisation of the front image. In this respect, if the dummy head technique was used as a separate rear microphone array, it might be a more reasonable way to place the microphone further back from the front array in order to reduce the direct energy as much as possible.

In this experiment, a dummy head (Cortex MK2) was positioned with the 'rear' microphone techniques, about 7m from the centre of the 'front' arrays. It was faced forwards.

**Spaced Cardioid Technique**

In this experiment, a technique was also used consisting of two cardioid microphones (Brüel & Kjær 4011) directed away from the direct sound. Each pointed towards the respective rear wall corners of the studio. They were also positioned as far back and as far apart from one another as was possible to capture as much reverberant sound as possible. This was done in order to reject as much of the direct sound as possible. They were placed 8m apart, with each microphone 4m from the centre-front line of the studio.

7.1.2 **Experimental Set-up**

The recording sessions took place in Studio 1 in the Department of Music and Sound Recording at the University of Surrey.

The studio is 14.5m wide, 17m long and is approximately 6.5m high. It is primarily used for the recording of classical music. Figure 94 shows the positions of the centre-points of the 'front' and 'rear' techniques, the dimensions of the Studio 1 floor and the area where the recorded sounds were positioned.
Figure 95 shows the relative positions of the microphones within the studio (facing backwards).

The ‘front’ arrays were positioned so that their outer (left-right) microphones were in a line, and that their centre microphones also formed a line. They were all mounted on a multiple microphone array centred on the centre-line of the studio, a distance 7m from the front wall (to allow recorded sources between the array and the front wall to be picked up within the recording angles of the front arrays and with a good direct-to-reverberant sound ratio). Figure 96 shows the ‘front’ arrays being set-up.
As for the rear arrays, they were ‘centred’ on the dummy head which was placed 7m behind the front arrays (about 3m from the far wall). The two four-channel techniques (IRT-Cross-inspired technique and Hamasaki Square) were positioned so that the dummy head was at their centres, and the Spaced Cardioid technique microphones were spaced 4m from the centre points (8m from one another), measured perpendicular to the centre-line of the studio. The rear microphones were raised as high as the microphone stands would allow. Figure 97 shows the rear microphone arrays in more detail. The IRT-Cross can be seen as four microphones positioned high up above the dummy head. The Hamasaki Square consists of the four microphones positioned about the dummy head on separate stands. The right spaced cardioid microphone can be seen on the far left of the picture, angled towards the right rear corner of the studio.

The 24 microphone signals (four ‘front’ triplets, two four-channel surround techniques and two two-channel surround techniques) were connected to the inputs of an analogue mixing console (Neve V series). They were recorded using three ganged 8-track digital recording tape recorders (Sony PCM-800). Each microphone and the channels of the dummy head were level aligned using a small portable tone generator held at 15cm from the capsule (or pinna on the dummy head). Recording levels were not altered during any of the recordings (except the Harpsichord – see below – where the level of all channels was boosted 5dB). The recordings were later transferred to a digital audio workstation (Digidesign Pro Tools HD) from the digital tapes to allow editing and the creation of separately mixed multichannel sound files for use in subjective testing. The transfer was done over analogue connections and sampled into Pro Tools at 16 bits and at 44.1kHz.
Appendices

Figure 97: ‘Rear’ Microphone Arrays (left spaced cardioid not shown)
7.2 Training Programme PowerPoint Slides

The following pages are a reproduction of the slides presented to the trainee subjects during the Tutorial phase of the training programmes of the experiments reported in Chapters 4 and 5.

Slide 1

Spatial Audio Listener Training System
“Width”

Slide 2

Objectives

After this course you should have:

- An overview of the spatial attributes of sound reproduction systems
- An understanding of why they are important
- An understanding of one spatial audio attribute description scheme
- The ability to hear the difference between various levels of the “width” of reproduced sound sources
• Increased availability of DVDs and "surround sound" has led to the increase in multichannel sound systems in the home:

• Not only does Surround Sound add two "surround" loudspeakers, it also adds a "centre" channel (and possibly a "subwoofer")
Many different applications involve quality judgements of multichannel audio:
- Loudspeaker System Design
- Microphones and Recording Techniques
- Mixing Techniques
- Sound Processing Algorithms
- Encoding/Transmission System Design
- Others

But how can we evaluate these systems?

Spatial Audio Evaluation?

- Timbral evaluation (Frequency-based)
- Technical evaluation (Noise, Distortion)
- But what of Spatial quality?
- No established method in training listeners to detect and describe changes in the spatial characteristics of sound
- Consider: Loudspeakers are normally evaluated in mono!
- What happens when multiple loudspeaker systems need to be optimised for spatial quality?
Spatial Audio Attributes

- Spatial audio attributes:
  - Concerned with the position and dimensions of auditory sources and the environment within which they are located.
- Examples:
  - Distance
  - Position
  - Depth
  - Width

A Spatial Audio Description Scheme

- There is no universal language for the description of spatial audio attributes
  Why could this be a problem?
  - Different people may use different words to describe the same phenomena
  - Different people may use the same words to describe different phenomena
- This causes understanding problems between people
- A standardised scheme to allow the description of spatial audio scenes would allow listeners to communicate with each other in a meaningful way
Spatial Audio Description Scheme

- The scheme used in this study separates source-related sounds from environment-related sounds.

Imagine yourself in an (enclosed) environment:
Now imagine a sound source, or group of sound sources, in this environment (for example, a guitar or a choir): how would we describe this spatially?

Slide 12

Spatial Attributes (Environment)
Slide 13

Focus of this Study

- In this study, we are interested in source-related spatial audio attributes:
  - Distance
  - Direction
  - Width
  - Depth
- Environment-related spatial audio attributes do not form part of the current study

Slide 14

Visual Analogy Practice

What is changing in the following scenes?
Slide 15

What has changed here?

A

Direction

B

Direction

Slide 16

What has changed here?

A

Distance

B

Distance
Slide 17

What has changed here?

A

Width

B

Width

A

B

Slide 18

What has changed here?

A

Depth

B

Depth
What has changed here?

A

B

Distance and Direction

What has changed here?

A

B

Width

Distance
What has changed here?

Slide 21

What has changed here?

Slide 22
What has changed here?

A

B

Distance, Direction and Width

What has changed here?

A

B

Distance, Direction, Width and Depth
Some Audio Examples

What is changing in the sounds?

Focus of Training

• During this week’s training sessions, we will focus on the WIDTH of single instrument sources
• Over the next few sessions, you will have a chance to gain a mastery of the detection and rating of various widths of two sound sources:
  – Guitar
  – Cornet

(Slide 28 shown is for the Pilot Study. The Transfer Study used a modified Slide 28)
Appendices

Slide 29

Training System

- Through practicing increasingly difficult tasks, you will hone your perception of the width of the sounds
- By the end of the training sessions you will be able to place five different sounds in rank order of width

(Slide 29 shown is for the Pilot Study. The Transfer Study used a modified Slide 29)

Slide 30

..and the point is?

- Knowledge of the spatial description scheme?
  - Provides you with a framework for describing spatial changes in reproduction that others will be able to understand.
- Experience at detecting and rating the individual elements of the scheme?
  - Will improve your ability to detect and rate spatial changes in reproduced audio
  - Improve your confidence and fluency when undertaking spatial audio listening tests
Glossary

Glossary

3/2 Stereo
An arrangement for spatial audio reproduction involving three frontal channels and two surround channels (ITU-R 1992-1994).

5.1 Surround Sound
A term coined by Tomlinson Holman to specify a sound reproduction format of five channels that conform to the 3/2 Stereo format, with an additional, reduced bandwidth channel for 'low frequency effects' intended for reproduction by a subwoofer. See (Holman 1999).

Accuracy
Ability to assign the correct grade to a specific stimulus.

Active Learning
Learning that involves the interactive engagement of the learner.

Behaviourist Learning
A theory of learning based upon the provision of passive learning and repetitive practice.

Consistency
Ability to assign the same or very similar grades or ranks to identical stimuli.

Constructivist Learning
A theory of learning involving the provision of environments or situations within which knowledge or skills are actively assimilated by the learner.

Digital Audio Workstation (DAW)
A computer system for recording, editing, mixing and mastering digital audio.

Drills
Repetitive practice or training sessions.

Experienced listener
A subject who has taken part in listening tests or who has worked in a related field in the audio industry.

Fidelity
The accuracy of a reproduction.

Hedonic Scale / Respose
Related to preference - whether something is appealing or not.

Max/MSP
A graphical programming language for MIDI and audio signal processing.
Glossary

Naïve Listener
A subject with very little or no previous listening experience or training.

Paradigm
A descriptive model.

Passive Learning
Information is presented to the learner without interaction.

Repetitive Practice
Iterating a task a number of times in order to gain proficiency.

Sensitivity
Ability to detect and rate/describe differences between different stimuli.

Subjective
Measures provided by test participants rather than by test equipment.

Trained Listener
A subject who has participated in a listener training programme, or who has practiced on the specific test in concern.

Training Programme
A structured course of instruction

Transfer of Learning/Training
Concerns the ability to apply learning from one situation to another. See Section 2.3.1.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKG</td>
<td>Akustiche und Kino-Geräte</td>
</tr>
<tr>
<td>ALEX</td>
<td>Audio Listening EXperiment software</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>ARCS</td>
<td>Attention, Relevance, Confidence, Satisfaction</td>
</tr>
<tr>
<td>ARD</td>
<td>Absolute Rank Difference</td>
</tr>
<tr>
<td>ASW</td>
<td>Apparent / Auditory Source Width</td>
</tr>
<tr>
<td>BS</td>
<td>Broadcasting Standard</td>
</tr>
<tr>
<td>CCCF</td>
<td>Challenge, Curiosity, Control and Fantasy</td>
</tr>
<tr>
<td>CD</td>
<td>Compact Disc</td>
</tr>
<tr>
<td>CSD</td>
<td>Computer Sound Design</td>
</tr>
<tr>
<td>DAW</td>
<td>Digital Audio Workstation</td>
</tr>
<tr>
<td>DVD</td>
<td>Digital Versatile Disc</td>
</tr>
<tr>
<td>ED</td>
<td>Ensemble Depth</td>
</tr>
<tr>
<td>EQ</td>
<td>EQualiser / EQualisation</td>
</tr>
<tr>
<td>EW</td>
<td>Ensemble Width</td>
</tr>
<tr>
<td>HRTF</td>
<td>Head-Related Transfer Function</td>
</tr>
<tr>
<td>ILD</td>
<td>Interaural Level Difference</td>
</tr>
<tr>
<td>INA</td>
<td>Ideale Nieren-Anordnung (Ideal Cardioid Arrangement)</td>
</tr>
<tr>
<td>IoSR</td>
<td>Institute of Sound Recording</td>
</tr>
<tr>
<td>IRT</td>
<td>Institut für Rundfunktechnik</td>
</tr>
<tr>
<td>ISC</td>
<td>Individual Scene Component</td>
</tr>
<tr>
<td>ITD</td>
<td>Interaural Time Difference</td>
</tr>
<tr>
<td>ITU-R</td>
<td>International Telecommunication Union (Radiocommunication)</td>
</tr>
<tr>
<td>LEV</td>
<td>Listener EnVevelopment</td>
</tr>
<tr>
<td>Max/MSP</td>
<td>Max Music Signal Processing</td>
</tr>
<tr>
<td>MDS</td>
<td>Multi Dimensional Scaling</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MURAL</td>
<td>MUltilevel auditoRy Assessment Language</td>
</tr>
<tr>
<td>MUSHRA</td>
<td>MUltiple Stimuli with Hidden Reference and Anchor</td>
</tr>
<tr>
<td>OCT</td>
<td>Optimized Cardioid Triangle</td>
</tr>
<tr>
<td>OSC</td>
<td>Overall Scene Component</td>
</tr>
<tr>
<td>PC</td>
<td>Personal Computer</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Squared</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SAALTS</td>
<td>Spatial Audio Attribute Listener Training System</td>
</tr>
<tr>
<td>SAAT</td>
<td>Spatial Audio Attribute Toolkit</td>
</tr>
<tr>
<td>SC</td>
<td>Scene Component</td>
</tr>
<tr>
<td>SED</td>
<td>Squared Euclidean Distances</td>
</tr>
<tr>
<td>SGI</td>
<td>Silicon GraphIcs</td>
</tr>
<tr>
<td>SSBP</td>
<td>Simplified Scene-Based Paradigm</td>
</tr>
<tr>
<td>SSED</td>
<td>Sum of Squared Euclidean Distances</td>
</tr>
<tr>
<td>SYWTBAHL</td>
<td>So You Want to be a Harman Listener</td>
</tr>
<tr>
<td>TB7</td>
<td>Teaching Block Room 7 at the University of Surrey</td>
</tr>
<tr>
<td>TBAC</td>
<td>Test of Basic Auditory Capabilities</td>
</tr>
<tr>
<td>WFS</td>
<td>Wavefield Synthesis</td>
</tr>
</tbody>
</table>


References


References


Hiyama, K., S. Komiyama and K. Hamasaki (2002). The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field. AES 113th Convention.


References


References


References


Power FX (1999). The Love & Pop String Thing!, Power FX.


Toole, F. E. (2004). Personal communication with the author.


References


Visio for Mac

Visio mac blog...

- Home
- About

Jan 22

Despite the fact that MS Visio is a part of MS Office suite, the Mac version of MS Office doesn’t contain MS Visio. Latest release of MS Office 2008 for Mac contain only Word, Excel, Powerpoint and Entourage applications.

So, there is no Visio for Mac now.

But don’t lose your heart! There is a way to share your Visio documents on Mac. To work with MS Visio documents on Mac you can use alternative software. You need to save your Visio document as *.vdx file (in Visio XML format) and find any diagramming software which works in Mac OS X and supports import of Visio XML files.

For example, you can use ConceptDraw Professional which is crossplatform (works both in MS Windows and Mac OS X) and supports import from and export to Visio XML. So you have an ability to work with Visio XML files on Mac just as with regular documents.

If you already have a vsd file and you don’t have a Visio installed you can use the free of charge Visio files converter.
Use the standard Open dialog to import your Visio XML file into ConceptDraw then save it as a regular conceptDraw document and work with this document in Mac OS X as well as in MS Windows.

Posted in Visio for mac, Visio mac, visio alternative I No Comments »

Jan 16

MS Visio is a professional diagramming tool which can work both with manually created shapes and with ready shapes. This ready shapes are collected in stencils. Visio has lots of stencils with ready shapes on a variety of thematics.

Each shape is represented both in Metric format of dimensions and in US (inches) format. Many of such ready shapes have unique features added with using of formulas of internal Visio programming language. For example, they can automatically stretch to fit some area or text. Despite the fact that Visio have lots stencils with shapes they are black-and-white in most cases.

So if you want to create a diagram with lots of color shapes, you need to find something else. ConceptDraw is an alternative for Visio both on PC and Mac and in contrast to Visio ConceptDraw has a set of libraries (analog of Visio stencils) with beautiful coloured shapes. This distinction is especially noticeable in ConceptDraw libraries and Visio stencils assigned for Landscape design.

ConceptDraw shapes also supports formulas and custom properties for all shapes. ConceptDraw uses its own internal CDBasic for formulas. So you can import Visio shapes to ConceptDraw and export ConceptDraw shapes to Visio but all “smart” properties given by formulas will be lost.
Visio for Mac
Visio stencils and ConceptDraw libraries
About

Categories

- visio alternative
- Visio for mac
- Visio mac
- Visio stencils

Archives

- January 2008

Blogroll

- Visio mac alternative

Search

\[\text{Search}\]

Meta:

- RSS
- Comments RSS
- Wordpress Themes

designed by: Online Website Builder and: Web Hosting Geeks | available free at: Top Wordpress Themes

3/6/08 13:30
Convert Visio 2003 (PC) documents to PDFs

I occasionally find myself using Visio 2003 on my PC (via Remote Desktop Connection from my Powerbook, of course). Since no one else in the office really uses Visio, I’ve come up with this quick work-around to turn my Visio files into PDFs. The neat thing is that this should work for just about any app that uses a proprietary image format...

1. Open your document in Visio.
2. Save it as a PNG (this will probably work with other image formats, too). Make sure you save it with a decent resolution — I’ve been using 256x256 pixels/inch. That’s usually more than enough, plus it’s an easy number to remember.
3. Move the PNG over to your Mac using your favourite method of file transfer.
4. Open the PNG in Preview, and choose File -> Save As.
5. Give the file a name, change the format to PDF, and click Save.

You can also choose File -> Print -> Print to PDF in Preview, but this will fit your PDF to a piece of paper. Since most of my Visios don’t end up getting printed, that method just has the effect of adding whitespace to the borders of my document.

[6,315 views]
Rate this hint: [ 1 2 3 4 5 ] Average rating: ☠ 1 ☠ 3 ☠ 5

Hint Options
- Mail Hint to a Friend
- Printable Hint Format

From our Sponsor...
PDF Converter for OS X
Docudesk.com/deskUNPDF Mac

User Functions
Username:
Password:
Login
Don't have an account yet? Sign up as a New User • Lost your password?

From our sponsors...

What's New:
HINTS
4 new Hints in the last 1 day
- Use Teleport with VNC for total remot... [+4]
- Back up FileVault directories to a Li... [+2]
- Leech - A powerful yet lightweight do... [+5]
- One way to monitor a child’s use of G... [+3]
- 10.5: View winmail.dat files in Mail....
- 10.5: Enable default permissions on s...
- Set Excel 2008 default zoom level via...
- Change appearance of EyeTV 3.x on-scr...
save the converted doc and delete the original postscript.

The advantage to this method is that you will retain the vector nature of the visio drawing.

This method works for any Windows program that allows printing.

**Convert Visio 2003 (PC) documents to PDFs**

By: diamondsw on Mon, Aug 15 '05 at 10:15AM PDT

Wow, so saving a document in a compatible format is a "hint" now?

---

By: peterjhill on Mon, Aug 15 '05 at 11:58AM PDT

Convert Visio 2003 (PC) documents to PDFs - By: peterjhill on Mon, Aug 15 '05 at 11:58AM PDT

You can also do this on Windows using PDFCreator (open source: http://sourceforge.net/projects/pdfcreator/) which gives Windows a feature similar to OS X's Print to PDF. I don't know if it will add whitespace to the borders of your document, but it should resize more gracefully than a PNG.

---

By: innate on Mon, Aug 15 '05 at 10:51AM PDT

You can also do this on Windows using PDFCreator (open source: http://sourceforge.net/projects/pdfcreator/) which gives Windows a feature similar to OS X's Print to PDF. I don't know if it will add whitespace to the borders of your document, but it should resize more gracefully than a PNG.

---

**Convert Visio 2003 (PC) documents to PDFs**

By: diamondsw on Mon, Aug 15 '05 at 10:15AM PDT

You can also do this on Windows using PDFCreator (open source: http://sourceforge.net/projects/pdfcreator/) which gives Windows a feature similar to OS X's Print to PDF. I don't know if it will add whitespace to the borders of your document, but it should resize more gracefully than a PNG.

---

**Bad resolution choice**

By: merlin on Mon, Aug 15 '05 at 9:36PM PDT

Why would you use 256x256? As most printers work in multiples of 300dpi, you will get less aliasing effects when you print if you use a multiple of 300dpi.

---

**Bitmaps are BAD mmm kay**

By: silicontrip on Tue, Aug 16 '05 at 12:48AM PDT

From memory Visio is a vector based drawing app of sorts?

PDFs support vector based images. So saving to a bitmap file will destroy all the vector based information which would normally be present in the PDF.

It is possible to load PDFs into certain programs and continue to edit them as vector based images. Doing what you describe would prevent this.

Someone mentioned a method to print to a postscript file, then convert the postscript to a PDF. I've used this method for years converting proprietry MS documents into files that can be edited by other operating systems.

Actually, pretty much every machine I've used in the last 15 years had the ability to print to postscript. It's been very useful to know these steps.