Towards A Spatial Ear Trainer

Tobias Neher

March 2004

Department of Music & Sound Recording
School of Arts, University of Surrey

Submitted for the degree of
Doctor of Philosophy

© 2004 Tobias Neher
ABSTRACT
The development of a spatial audio ear trainer has been proposed and investigated. A review of the relevant literature has shown that although numerous researchers have simulated spatial characteristics of sound, the devised simulation algorithms were not (properly) verified with respect to their intended subjective effects, thus making them unsuitable for training purposes. Additionally, whilst various timbral ear-training systems have been set up, there is none for spatial quality. To address these shortcomings four spatial attributes of reproduced sound have been considered: source distance, source width, ensemble width and ensemble depth. For each attribute a processing algorithm was developed that allowed highly controlled changes in the respective percept. Using these algorithms four sets of stimuli were carefully generated with the aim of achieving unidimensional variation in terms of the qualities of interest. In order to allow detailed and reliable validation of the simulations' auditory effects, a sensory evaluation strategy was devised that relies on Multidimensional Scaling techniques and the elicitation of supplementary qualitative data as a means of obtaining a complete sensory profile of a group of stimuli. With the help of this methodology as well as critical listening panels the psychological structures of the sound excerpts were measured. Results showed that the envisaged unidimensionality of the source distance, ensemble width and ensemble depth samples was accomplished. As to the source width simulation, results were not as clear-cut, which was attributed to listener unfamiliarity with this perceptual construct and greater stimulus uncertainty. Nevertheless, the analyses unfolded only one major dimension, which is why all four attribute simulations were deemed suitable for the proposed training purposes.
# CONTENTS

<table>
<thead>
<tr>
<th>ABSTRACT</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENTS</td>
<td>III</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>VII</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>XI</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>XIII</td>
</tr>
<tr>
<td>0 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>0.1 How to evaluate spatial audio systems?</td>
<td>2</td>
</tr>
<tr>
<td>0.2 Training in reproduced sound</td>
<td>3</td>
</tr>
<tr>
<td>0.3 The need for unidimensionality</td>
<td>4</td>
</tr>
<tr>
<td>0.4 Mathematical vs. perceptual unidimensionality</td>
<td>5</td>
</tr>
<tr>
<td>0.5 Aims of the research</td>
<td>8</td>
</tr>
<tr>
<td>0.6 Structure of the thesis</td>
<td>9</td>
</tr>
<tr>
<td>0.7 Original contributions</td>
<td>10</td>
</tr>
<tr>
<td>0.8 Summary</td>
<td>10</td>
</tr>
<tr>
<td>1 THE MULTIDIMENSIONAL NATURE OF SPATIAL QUALITY</td>
<td>12</td>
</tr>
<tr>
<td>1.1 The Berg and Rumsey studies</td>
<td>12</td>
</tr>
<tr>
<td>1.2 The Zacharov and Koivuniemi studies</td>
<td>14</td>
</tr>
<tr>
<td>1.3 Rumsey's scene-based paradigm for spatial audio evaluation</td>
<td>16</td>
</tr>
<tr>
<td>1.4 Summary</td>
<td>18</td>
</tr>
<tr>
<td>2 CORRELATIONS BETWEEN SPATIAL ATTRIBUTES AND PHYSICAL FACTORS</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Source direction</td>
<td>19</td>
</tr>
<tr>
<td>2.2 Source distance</td>
<td>19</td>
</tr>
<tr>
<td>2.2.1 Frequency content</td>
<td>19</td>
</tr>
<tr>
<td>2.2.2 Binaural cues under free-field conditions</td>
<td>21</td>
</tr>
<tr>
<td>2.2.3 Loudness</td>
<td>21</td>
</tr>
<tr>
<td>2.2.4 Direct-to-reverberant sound ratio (D/R)</td>
<td>25</td>
</tr>
<tr>
<td>2.2.5 Finer details of reflected sound</td>
<td>26</td>
</tr>
<tr>
<td>2.2.6 The Doppler effect</td>
<td>35</td>
</tr>
<tr>
<td>2.2.7 Interchannel cross-correlation</td>
<td>35</td>
</tr>
<tr>
<td>2.2.8 Low-frequency decorrelation</td>
<td>37</td>
</tr>
<tr>
<td>2.3 Ensemble depth</td>
<td>39</td>
</tr>
<tr>
<td>2.4 Externalisation</td>
<td>41</td>
</tr>
<tr>
<td>2.5 Source width</td>
<td>42</td>
</tr>
<tr>
<td>2.5.1 Objective measures and applicability to reproduced sound</td>
<td>42</td>
</tr>
<tr>
<td>2.6 Ensemble width</td>
<td>44</td>
</tr>
<tr>
<td>2.7 Environment width</td>
<td>45</td>
</tr>
<tr>
<td>2.8 Listener envelopment (LEV)</td>
<td>45</td>
</tr>
<tr>
<td>2.8.1 Level and spatial distribution of late reflected sound</td>
<td>45</td>
</tr>
<tr>
<td>2.8.2 Temporal distribution of late reflected sound</td>
<td>46</td>
</tr>
<tr>
<td>2.9 Summary</td>
<td>48</td>
</tr>
<tr>
<td>3 METHODS FOR STUDYING THE SUBJECTIVE EFFECTS OF SOUND-PROCESSING ALGORITHMS</td>
<td>50</td>
</tr>
<tr>
<td>3.1 Elicitation of common attribute scales</td>
<td>50</td>
</tr>
</tbody>
</table>
4 EXTANT PERCEPTUAL SOUND PROCESSORS RELATED TO SPATIAL QUALITY....58
4.1 The origin ..............................................................................58
4.2 The spatial reverberator .................................................................59
4.3 The Spatialisateur (Spat) .................................................................60
4.4 Virtual Surround Panning (VSP) .......................................................62
4.5 The SceneBuilder ......................................................................64
4.6 Other approaches for controlling width and distance attributes ......66
    4.6.1 Divergence.............................................................................66
    4.6.2 Stereo shuffling ......................................................................66
    4.6.3 Gerzon's source distance and width controls .........................67
    4.6.4 ZoomFX ...............................................................................68
    4.7 Summary ..................................................................................69

5 UNIDIMENSIONAL SIMULATION AND VALIDATION OF SPATIAL ATTRIBUTES
OF REPRODUCED SOUND ................................................................70
5.1 Which spatial attributes to choose? ..............................................70
5.2 Reproduction layout and experimental set-up ................................71
5.3 General approach to attribute simulations and validations ...........72
5.4 Simulation of source distance .......................................................74
    5.4.1 Creation of source distance stimuli ..........................................74
5.5 Validation of source distance stimuli ...........................................77
    5.5.1 Experimental design .................................................................77
    5.5.2 Results ..................................................................................79
    Analysis of similarity data...............................................................79
    Analysis of verbal data....................................................................82
    INDSCAL analysis ........................................................................84
    5.6 Simulation of source width ..........................................................87
    5.6.1 Creation of source width stimuli .............................................87
    5.7 Validation of source width stimuli I .............................................91
    5.7.1 Results ..................................................................................91
    Analysis of similarity data...............................................................91
    Analysis of verbal data....................................................................93
    INDSCAL analysis ........................................................................94
    5.8 Validation of source width stimuli II ..........................................95
    5.8.1 Results ..................................................................................96
    Analysis of similarity data...............................................................96
    Analysis of verbal data....................................................................98
    INDSCAL analysis ........................................................................100
    5.8.2 Confirmation experiment .........................................................101
    Analysis of classification data .........................................................103
    Analysis of ranking data.................................................................103
    5.9 Simulation of ensemble width .....................................................104
    5.9.1 Pilot study into the perception of ensemble width ....................105
    5.9.2 Spatial attribute processing platform ........................................107
    5.9.3 A new surround panner .............................................................108
    Analytical comparison of new panner and 5-channel PCPP .............109
    Subjective comparison of new panner and 5-channel PCPP .............113
    5.9.4 Creation of ensemble width stimuli ........................................115
5.10 Validation of ensemble width stimuli ......................................................................................................................... 117
5.10.1 Experimental design..................................................................................................................................................... 117
5.10.2 Results ....................................................................................................................................................................... 120
      Analysis of similarity data ............................................................................................................................................... 120
      Analysis of verbal data ....................................................................................................................................................... 121
      INDSCAL analysis .............................................................................................................................................................. 124
      Analysis of graphical data ................................................................................................................................................... 125
      ... but does it sound right? ................................................................................................................................................ 126
5.11 Simulation of ensemble depth I ............................................................................................................................................. 126
5.11.1 Pilot study into the perceptual effects of early reflection pattern characteristics .................................................. 127
      Acoustical parameters and test conditions .......................................................................................................................... 128
      Creation of test materials ..................................................................................................................................................... 130
      Experimental design ............................................................................................................................................................ 133
      Listening panel .................................................................................................................................................................... 134
5.11.2 Results ........................................................................................................................................................................... 134
      Analysis of ABX data .......................................................................................................................................................... 135
      Analysis of Semantic Differential data ............................................................................................................................... 137
      Analysis of directionality judgements .................................................................................................................................. 139
5.11.3 Summary of pilot study ................................................................................................................................................... 141
5.11.4 Creation of ensemble depth stimuli ............................................................................................................................ 142
      Source material .................................................................................................................................................................... 142
      Simulation strategy ............................................................................................................................................................... 143
      Applied processing ............................................................................................................................................................... 143
5.12 Validation of ensemble depth stimuli I ........................................................................................................................... 145
5.12.1 Results ........................................................................................................................................................................... 145
      Analysis of similarity data ....................................................................................................................................................... 145
      Analysis of verbal data .......................................................................................................................................................... 147
      Analysis of graphical data ....................................................................................................................................................... 149
      ... but does it sound right? .................................................................................................................................................... 150
5.13 Simulation of ensemble depth II ........................................................................................................................................ 151
5.13.1 Creation of ensemble depth stimuli ............................................................................................................................... 151
      Source material .................................................................................................................................................................... 151
      Applied processing ............................................................................................................................................................... 152
5.14 Validation of ensemble depth stimuli II ........................................................................................................................... 153
5.14.1 Results ........................................................................................................................................................................... 153
      Analysis of similarity data ....................................................................................................................................................... 153
      Analysis of verbal data .......................................................................................................................................................... 155
      INDSCAL analysis ............................................................................................................................................................... 157
      Analysis of graphical data ....................................................................................................................................................... 158
      ... but does it sound right? .................................................................................................................................................... 159
5.15 Concluding remarks ............................................................................................................................................................ 160
5.16 Summary ............................................................................................................................................................................. 161

6 SUMMARY, CONCLUSIONS AND FURTHER WORK .................................................................................................. 164

6.1 Summary and conclusions ................................................................................................................................................ 164
6.1.1 Chapter 0 ...................................................................................................................................................................... 164
6.1.2 Chapter 1 ...................................................................................................................................................................... 164
6.1.3 Chapter 2 ...................................................................................................................................................................... 165
6.1.4 Chapter 3 ...................................................................................................................................................................... 165
6.1.5 Chapter 4 ...................................................................................................................................................................... 166
6.1.6 Chapter 5 ...................................................................................................................................................................... 166
6.2 Further work ....................................................................................................................................................................... 168
6.2.1 Simulation and validation of other salient spatial attributes .................................................................................... 169
6.2.2 Transferability of attribute simulations ....................................................................................................................... 169
6.2.3 Acoustic analysis of created attribute simulations ....................................................................................................... 170
6.2.4 Development of an intuitive and effective training system ......................................................................................... 170
6.2.5 Evaluation of suitability of system for training purposes ........................................................................................... 172
6.3 Summary ............................................................................................................................................................................. 172
APPENDIX A: GRAPHICAL RESPONSES FROM ENSEMBLE WIDTH VALIDATION

APPENDIX B: ATTRIBUTE GROUPS FROM PILOT STUDY INTO EARLY REFLECTION PATTERN CHARACTERISTICS

APPENDIX C: GRAPHICAL RESPONSES FROM ENSEMBLE DEPTH VALIDATION I

APPENDIX D: GRAPHICAL RESPONSES FROM ENSEMBLE DEPTH VALIDATION II

APPENDIX E: PILOT STUDY INTO TRAINING OF LISTENERS FOR SPATIAL AUDIO EVALUATION

   E.1 Experimental design
   E.2 Results
   E.3 Summary

GLOSSARY

Abbreviations and terms

REFERENCES
LIST OF FIGURES

Figure 0.1: The 3/2-stereo reference arrangement showing the positions of the left (L), right (R), centre (C), left-surround (LS) and right-surround (RS) loudspeakers as recommended in [ITU-R BS.775, 1994]......................................................... 1

Figure 0.2: Graphical illustration of a conceivable relationship between four mathematical dimensions, one of which is a perceptual dimension ('Source distance'), and two of which are physical dimensions ('Direct sound level' and 'Reverberant sound level')................................................................. 6

Figure 0.3: Conceptual model of human perception of multidimensional stimuli (adapted from [Bech, 1999]). If the aim is to create a set of stimuli that differ only in their intensities with respect to one individual sensory impression (or attribute), all associated sub-attributes need to be closely concerted. If agreement is not achieved, a subject presented with that attribute simulation might perceive the various sub-attributes or 'ingredients' (e.g. milk, salt, cherries, eggs etc.). Conversely, if these ingredients are added in the right amounts, they will be perceptually fused by the sensory system, resulting in a unified gestalt perception (e.g. cherry pie). If the latter is the case, the attribute simulation will be considered perceptually unidimensional......................................................... 7

Figure 1.1: Graphical illustration of the concept of source distance as well as different types of width and depth attributes (as defined in [Rumsey, 2002]). The different arrow sizes indicate the transition from micro- to macro-level components of a spatial sound scene................................................................................................. 17

Figure 2.1: Magnitude of air absorption as a function of distance and frequency for an air humidity of 20% and a temperature of 20°C (adapted from [Savioja et al., 1999])......................................................... 20

Figure 2.2: Graphical illustration of the inverse distance law (after [Everest, 2001])................................................................................................................................. 22

Figure 2.3: Graphical illustration of the relationship between (total) sound pressure level and source distance as a function of the diffusivity of the sound field (adapted from [Rumsey & McCormick, 2002])................................................................................................. 22

Figure 2.4: Localisation curve from data of five subjects (solid line) collected under anechoic conditions using natural speech as source material (adapted from [von Békésy, 1949])......................................................... 24

Figure 2.5: Begault's proposed scheme for varying the level of (anechoic) speech to achieve perceptually equal distance intervals (adapted from [Begault, 1987]). All values given in this diagram are relative to Distance 1 (Distance 4 corresponds to Begault's chosen reference level of 65dBC)................................................................................................................................. 25

Figure 2.6: Schematic illustration of Griesinger's ideal reverberation profile (after [Griesinger, 2000])................................................................................................................................. 30

Figure 2.7: Graphical illustration of relationships between timbre, distance and room size perception and the respective impulse response time regions as described in [Pellegrini, 2002]................................................................................................................................. 31

Figure 2.8: Specular reflections (up to second order) calculated using an image-source model algorithm for a rectangular room with a volume of 1200m³ (left graph) and 3780m³ (right graph). The source-receiver arrangement is constant and slightly asymmetric. The direct sound is at 0ms................................................................................................................................. 31

Figure 2.9: Specular reflections (up to second order) calculated using an image-source model algorithm for a rectangular room with a volume of 1200m³. The receiver is positioned at one end of the room and the source is placed in the centre (left graph) or at the other end of the room (right graph). The source-receiver arrangement is slightly asymmetric. The direct sound is at 0ms................................................................................................................................. 31

Figure 2.10: Plan view of the spatial distribution of the early reflections of Figure 2.9 (displayed as image sources). Level is represented by circle size. For each graph the source is located in the centre of the largest circle (the direct sound) and the receiver position is indicated by the black dot................................................................................................................................. 32
Figure 2.11: Graphical illustration of the Doppler effect ................................................................. 35
Figure 2.12: Summarised listener responses from test on impact of LF decorrelation on perceived spatial imagery. The stimuli's cross-correlation coefficients are indicated for each auditory event (adapted from [Martens, 1999]) ................................................................. 38
Figure 2.13: Graphical illustration of the room-related balancing technique for the case of a single spot and monophonic main microphone (adapted from [Theile, 2001]) ........................................................................... 40
Figure 2.14: Graphical illustration of ITD fluctuations as modelled for an omnidirectional source 15m in front and a single reflection coming from a wall 5m to the right of a binaural receiver (reproduced from [Mason & Rumsey, 2001] with permission). The DS signal is composed of three continuous frequencies of 480Hz, 500Hz and 520Hz. Figure 2.14a shows the computed ear signals, while Figure 2.14b displays the ITD fluctuation over time ............................................................................. 43
Figure 3.1: Schematic illustration of the basic steps involved in QDA ............................................. 51
Figure 3.2: Schematic illustration of the basic steps involved in FCP ............................................. 52
Figure 3.3: Schematic illustration of the basic steps involved in the RGT ...................................... 53
Figure 3.4: Schematic illustration of the basic steps involved in CA ............................................. 54
Figure 3.5: Schematic illustration of the basic steps involved in MDS ........................................... 55
Figure 4.1: Simplified block diagram of the spatial reverberator (after [Kendall & Martens, 1984]) ......................................................................................................................... 59
Figure 4.2: The graphical interface used within the Spat. Along the top, the groupings into source- and room-related attributes can be seen ............................................................................. 61
Figure 4.3: Simplified block diagram of a single VSP channel (adapted from [Horbach et al., 2000]) ............................................................................................................................ 63
Figure 4.4: Block diagram of head-related, frequency-dependent panner with stereo width control for a single VSP channel (after [Horbach, 1997]) ...................................................... 64
Figure 4.5: Block diagram of the SceneBuilder (adapted from [Corey, 2002]) ................................. 65
Figure 4.6: Block diagram of stereo shuffling (after [Gerzon, 1986]) ............................................. 67
Figure 4.7: Block diagram of 'pseudo-stereo' image spreading circuit (after [Gerzon, 1992c]) ......... 68
Figure 4.8: Block diagram of the ZoomFX algorithm (after [Sibbald, 1999]) ................................. 68
Figure 5.1: Diagram of general experimental set-up (listening room) ............................................. 72
Figure 5.2: Schematic illustration of the general approach adopted for the attribute simulations and validations ............................................................................................................. 73
Figure 5.3: Generic impulse response of the reverberator used for the simulation of source distance showing the direct sound, the six early reflections and the reverberation tail ......................................................................................................................... 75
Figure 5.4: Block diagram of the processing employed for simulating source distance ..................... 76
Figure 5.5: User-interface employed for the MDS experiments ....................................................... 78
Figure 5.6: Scree plot derived from non-metric MDS analysis (source distance) ............................. 80
Figure 5.7: 1-dimensional stimulus space from non-metric MDS analysis (source distance) .......... 82
Figure 5.8: 2-dimensional subject space from non-metric INDSCAL analysis (source distance) .... 86
Figure 5.9: Generic impulse response of the reverberator used as part of the simulation of source width showing the direct sound, the six early reflections and the reverberation tail ......................................................................................................................... 90
Figure 5.10: Block diagram of the processing employed for simulating source width ...................... 90
Figure 5.11: Scree plot derived from non-metric MDS analysis (source width, listening room) ........ 91
Figure 5.12: 1-dimensional stimulus space from non-metric MDS analysis (source width, listening room) .......................................................................................................................... 93
Figure 5.13: 2-dimensional subject space from non-metric INDSCAL analysis (source width, listening room) .......................................................................................................................... 95
Figure 5.14: Diagram of experimental set-up (Studio 3). The loudspeakers were positioned at 0°, ±30° and ±110° ............................................................................................................. 96
Figure 5.47: Relationships between terminology and attributes (after [Shaw & Gaines, 1995])....161
Figure A.1: Graphical response from Listener 1 (ensemble width).............................................174
Figure A.2: Graphical response from Listener 2 (ensemble width).............................................174
Figure A.3: Graphical response from Listener 3 (ensemble width).............................................174
Figure A.4: Graphical response from Listener 4 (ensemble width).............................................174
Figure A.5: Graphical response from Listener 5 (ensemble width).............................................174
Figure A.6: Graphical response from Listener 6 (ensemble width).............................................174
Figure A.7: Graphical response from Listener 7 (ensemble width).............................................175
Figure A.8: Graphical response from Listener 8 (ensemble width).............................................175
Figure A.9: Graphical response from Listener 9 (ensemble width).............................................175
Figure A.10: Graphical response from Listener 10 (ensemble width).......................................175
Figure A.11: Graphical response from Listener 11 (ensemble width).......................................175
Figure A.12: Graphical response from Listener 12 (ensemble width).......................................175
Figure A.13: Graphical response from Listener 13 (ensemble width).......................................175
Figure A.14: Graphical response from Listener 14 (ensemble width).......................................175
Figure A.1: Graphical response from Listener 2 (ensemble depth, experiment 1)......................179
Figure A.2: Graphical response from Listener 3 (ensemble depth, experiment 1)......................179
Figure A.3: Graphical response from Listener 5 (ensemble depth, experiment 1)......................179
Figure A.4: Graphical response from Listener 7 (ensemble depth, experiment 1)......................179
Figure A.5: Graphical response from Listener 9 (ensemble depth, experiment 1)......................179
Figure A.6: Graphical response from Listener 10 (ensemble depth, experiment 1)......................179
Figure A.7: Graphical response from Listener 11 (ensemble depth, experiment 1)......................180
Figure A.8: Graphical response from Listener 13 (ensemble depth, experiment 1)......................180
Figure D.1: Graphical response from Listener 1 (ensemble depth, experiment 2)......................181
Figure D.2: Graphical response from Listener 2 (ensemble depth, experiment 2)......................181
Figure D.3: Graphical response from Listener 3 (ensemble depth, experiment 2)......................181
Figure D.4: Graphical response from Listener 4 (ensemble depth, experiment 2)......................181
Figure D.5: Graphical response from Listener 5 (ensemble depth, experiment 2)......................181
Figure D.6: Graphical response from Listener 6 (ensemble depth, experiment 2)......................181
Figure D.7: Graphical response from Listener 7 (ensemble depth, experiment 2)......................182
Figure D.8: Graphical response from Listener 8 (ensemble depth, experiment 2)......................182
Figure D.9: Graphical response from Listener 9 (ensemble depth, experiment 2)......................182
Figure D.10: Graphical response from Listener 10 (ensemble depth, experiment 2)....................182
Figure D.11: Graphical response from Listener 11 (ensemble depth, experiment 2)....................182
Figure D.12: Graphical response from Listener 12 (ensemble depth, experiment 2)....................182
Figure D.13: Graphical response from Listener 13 (ensemble depth, experiment 2)....................182
Figure E.1: Screenshots of the familiarisation window, the corresponding viewer window and
the performance window employed to provide the user with feedback during
training exercises.................................................................186
Figure E.2: Wrongness of rank responses – Source distance, trained group..............................187
Figure E.3: Wrongness of rank responses – Source distance, untrained group..........................187
Figure E.4: Response times – Source distance, trained group..................................................188
Figure E.5: Response times – Source distance, untrained group..............................................188
Figure E.6: Wrongness of rank responses – Source width, trained listeners..............................189
Figure E.7: Wrongness of rank responses – Source width, untrained listeners..........................189
Figure E.8: Response times – Source width, trained group.....................................................190
Figure E.9: Response times – Source width, untrained group..................................................190
LIST OF TABLES

Table 1.1: Spatial attributes as derived by Berg and Rumsey. The definitions given in the last column are based on information published in [Berg & Rumsey, 2002].......................... 13

Table 1.2: Spatial attributes as derived by Zacharov and Koivuniemi. The definitions given in the last column are based on information published in [Koivuniemi & Zacharov, 2001] 15

Table 2.1: Reviewed spatial attributes, their physical correlates (relevant to this work) and respective references..........................................................................................49

Table 5.1: RSQ results from non-metric MDS analysis (source distance)............................................................................................................. 81
Table 5.2: Groups of holistic terms and their relative weights (source distance).............................................................. 83
Table 5.3: Groups of analytical terms and their relative weights (source distance).............................................................. 84
Table 5.4: RSQ results from non-metric MDS analysis (source width, listening room).............................................................. 92
Table 5.5: Groups of holistic terms and their relative weights (source width, listening room).............................................................. 94
Table 5.6: Groups of analytical terms and their relative weights (source width, listening room).............................................................. 94
Table 5.7: RSQ results from non-metric MDS analysis (source width, Studio 3).......................................................................................... 97
Table 5.8: Groups of holistic terms and their relative weights (source width, Studio 3).............................................................. 99
Table 5.9: Groups of analytical terms and their relative weights (source width, Studio 3).............................................................. 99
Table 5.10: Accumulated attribute classification judgements from confirmation experiment (source width, Studio 3).....................................................................................103
Table 5.11: Ranking sequences and associated SEDs from confirmation experiment (source width, Studio 3)............................................................................................................. 104
Table 5.12: RSQ results from non-metric MDS analysis (ensemble width).......................................................................................... 121
Table 5.13: Groups of holistic terms and their relative weights (ensemble width).......................................................................................... 122
Table 5.14: Groups of analytical terms and their relative weights (ensemble width).......................................................................................... 123
Table 5.15: Details of ER pattern characteristics for the created test materials. Each test material was investigated for three levels of ED (see text)........................................................................................... 128
Table 5.16: The 24 test items, their associated test materials and the corresponding physical ER characteristics. The one parameter varied in each case is printed in *italics* .................................................................................. 130
Table 5.17: Percentages of correct answers and corresponding *p*-values for each of the 24 test items of the ABX test............................................................................................................. 136
Table 5.18: 'Total difference' coefficients for the 15 test items applied to the Semantic Differential and associated *p*-values ............................................................................................................. 139
Table 5.19: Certainty factors and modes of the directionality judgements of the associated VPA categories for each of the 15 test items submitted to the Semantic Differential analysis............................................................................................................. 140
Table 5.20: RSQ results from non-metric MDS analysis (ensemble depth, experiment 1)............................................................................................................. 146
Table 5.21: Groups of holistic terms and their relative weights (ensemble depth, experiment 1)............................................................................................................. 148
Table 5.22: Groups of analytical terms and their relative weights (ensemble depth, experiment 1)............................................................................................................. 148
Table 5.23: RSQ results from non-metric MDS analysis (ensemble depth, experiment 2)............................................................................................................. 154
Table 5.24: Groups of holistic terms and their relative weights (ensemble depth, experiment 2)............................................................................................................. 156
Table 5.25: Groups of analytical terms and their relative weights (ensemble depth, experiment 2)............................................................................................................. 157
Table B.1: Attribute groups, occurrences and weight factors obtained for '2 flat' test item as part of pilot study into ER pattern characteristics............................................................................................................. 176
Table B.2: Attribute groups, occurrences and weight factors obtained for '2 med' test item as part of pilot study into ER pattern characteristics............................................................................................................. 176
Table B.3: Attribute groups, occurrences and weight factors obtained for '2 deep' test item as part of pilot study into ER pattern characteristics............................................................................................................. 176
Table B.4: Attribute groups, occurrences and weight factors obtained for '3 flat' test item as part of pilot study into ER pattern characteristics............................................................................................................. 176
Table B.5: Attribute groups, occurrences and weight factors obtained for '3 deep' test item as part of pilot study into ER pattern characteristics

Table B.6: Attribute groups, occurrences and weight factors obtained for '4 med' test item as part of pilot study into ER pattern characteristics

Table B.7: Attribute groups, occurrences and weight factors obtained for '4 deep' test item as part of pilot study into ER pattern characteristics

Table B.8: Attribute groups, occurrences and weight factors obtained for '5 flat' test item as part of pilot study into ER pattern characteristics

Table B.9: Attribute groups, occurrences and weight factors obtained for '5 med' test item as part of pilot study into ER pattern characteristics

Table B.10: Attribute groups, occurrences and weight factors obtained for '5 deep' test item as part of pilot study into ER pattern characteristics

Table B.11: Attribute groups, occurrences and weight factors obtained for '6 med' test item as part of pilot study into ER pattern characteristics

Table B.12: Attribute groups, occurrences and weight factors obtained for '6 deep' test item as part of pilot study into ER pattern characteristics

Table B.13: Attribute groups, occurrences and weight factors obtained for '8 flat' test item as part of pilot study into ER pattern characteristics

Table B.14: Attribute groups, occurrences and weight factors obtained for '8 med' test item as part of pilot study into ER pattern characteristics

Table B.15: Attribute groups, occurrences and weight factors obtained for '8 deep' test item as part of pilot study into ER pattern characteristics

Table E.1: Details of source distance training

Table E.2: Details of source width training
ACKNOWLEDGEMENTS

This thesis would not have been possible without the assistance of several individuals and organisations. I would therefore like to thank: Francis Rumsey and Tim Brookes for all their supervision, guidance and advice; Rhonda Wilson for her interest and valuable contributions during various discussions; Meridian Audio Ltd. for sponsorship; my colleagues at the Institute of Sound Recording for asking good questions and giving useful feedback; the AES Educational Foundation for financial support; Peter Craven for provision of his panning algorithm and associated information; Søren Bech for his helpful comments with respect to some problems and ideas; my parents for all their support, encouragement and interest in my education and work; all my listening test subjects for their time and patience; Russell Mason and Ben Supper for their help with proof-reading; and Abigail Martin for checking my grammar and for making sure that I got some daylight and fresh air every now and then.
INTRODUCTION

Over the last decade or so, the audio industry has been witnessing a constant growth in the use of multichannel sound systems. This trend can be primarily traced back to "the 'home cinema' revolution" [Rumsey, 2001], as part of which the DVD-Video format [Holman, 2000] was established as a mass-market medium. Nonetheless, audio-only multichannel productions are becoming more and more common, too. This is mainly due to the DVD-Audio and Super Audio CD formats [Holman, 2000], which were introduced more recently to allow widespread distribution of high-resolution surround sound recordings. The majority of current multichannel sound productions (with or without picture) have in common that their content is/was authored with a particular reproduction layout in mind. Known colloquially as '5.1', this configuration represents "a compromise between the need for optimum spatial enhancement of reproduction and the need for an approach that was practicable and compatible with conventional two-channel reproduction" [AES, 2001]. It utilises five full-bandwidth channels, with an additional bandwidth-limited, low-frequency-enhancement channel being an option. Figure 0.1 shows the reference loudspeaker arrangement as recommended in [ITU-R BS.775, 1994]. This configuration is commonly referred to as the '3/2-stereo' or 'ITU' set-up and, in an attempt to ensure compatibility, has been widely adopted for the reproduction of film sound (both in the cinema and home) and other types of multichannel programme material.

Figure 0.1: The 3/2-stereo reference arrangement showing the positions of the left (L), right (R), centre (C), left-surround (LS) and right-surround (RS) loudspeakers as recommended in [ITU-R BS.775, 1994]
0. Introduction

0.1 How to evaluate spatial audio systems?

As a direct consequence of the recent proliferation of 3/2-stereo (as well as other forms of spatial sound display such as 3-D audio for virtual reality [Blauert, 2002] or Wave Field Synthesis [Berkhout et al., 1993]), there has arisen the need to assess the performance of such systems in terms of their spatial quality. For this purpose, the audio engineer can in principle employ objective (or physical) and subjective (or perceptual) measurements, whereby the former type of evaluation may offer a number of advantages over the latter. Generally speaking, objective measurements can exhibit higher consistency over sequential readings since they are not susceptible to distortions as caused by inter-assessor differences in, for example, familiarity, sensitivity or preference with respect to a certain sensory attribute. Such preconditioning is one reason for inconsistent judgements, which may even be evident in the (repeated) verdicts of an individual subject during a single test session for one particular test condition. Furthermore, there is the issue of erroneous evaluations, "be it due to lack of attention, misinterpretation by the subject or experimenter, vague judgement methods, or mistakes" [Mason, 2002a].

Since physical measurements do not suffer from such human peculiarities they can yield accurately repeatable results, thereby making the evaluation process time- and cost-effective [Grewin, 1995]. However, for the results to be of any practical value it is crucial that they correlate well with subjects' opinions of the perceptual phenomenon, which the physical measurements are meant to have gauged. As for the area of spatial audio assessment, work is under way on the development of suitable objective predictors. In this context, some good progress has recently been made with respect to a few spatial attributes of reproduced sound [Mason, 2002a; Soulodre et al., 2003]. Yet, these measures are far from being ready to substitute subjective assessments, as much more research is required to widen their applicability and to improve their accuracy for a large range of stimulus types (see [Mason, 2002a] and Section 2.8.2). Even if physical measures were eventually able to quantify the numerous facets of spatial auditory perception (see Chapter 1) in a meaningful way, “it is not likely that an objective measurement method can ever totally replace subjective assessments” [Grewin, 1995]. Hence, experimenters have to resort to subjective testing methods to obtain information about the auditory characteristics of spatial sound stimuli.

Sensory analysts make no secret of the fact that using humans as measuring instruments has a number of drawbacks. As mentioned above, humans can be highly variable in their judgements, both on an intra- and inter-subject level. This may cause subjective evaluations to be inefficient and prone to unreliability, e.g. if a certain statistical confidence level is needed to allow extrapolation of findings to the general public. Therefore, to be able to conduct valid and reliable sensory tests, it may be essential to minimise the variability so as to obtain meaningful data on which well-founded decisions can be made. For that purpose, subjects must be put in a frame of mind to understand the characteristics they

1 Actually, this may be impossible (or at least very difficult) to achieve if a given perceptual construct is largely determined by high-level cognitive processes rather than low-level psychophysical relationships.
are asked to measure, which can be achieved by controlled practice and training [Meilgaard et al., 1991]. This is especially important if stimuli are to be evaluated in terms of several specific attributes, as the risk of confusion or different understandings of semantic meanings on behalf of the subjects is even higher in that case.

0.2 Training in reproduced sound

Training is commonly applied in a wide range of disciplines. By simulating the perceptual phenomena of interest, panellists can be exposed to and hence familiarised with the characteristics that they are required to assess at a later stage. As to the field of reproduced sound, there is concrete evidence that training methods can lead to listener improvements regarding specific auditory skills. For instance, trained listeners were shown to perform substantially better than untrained ones in terms of their ability to detect and discriminate between different sound stimuli in a consistent fashion [Olive, 1995]. Consequently, they were more critical and produced data that were statistically more reliable, yielding quantifiable improvements in their performances.

Another study into the effects of training listeners in the assessment of sound-reproducing equipment was conducted by Bech [1992]. Like Olive, Bech was able to show that the training resulted in a higher consistency in the ratings of his subjects. Based on the data analysis he further concluded that from a statistical point of view one trained listener can replace seven untrained ones. This is an important finding since such a big reduction in the number of subjects needed to achieve a certain statistical confidence level can lead to substantial savings in time and hence costs, especially when more comprehensive listening experiments have to be carried out.

Olive [2003] made another crucial discovery. Intending to answer the question of whether training listeners has the effect of making their verdicts unrepresentative of the general public, he conducted a very large study to compare the loudspeaker preference judgements (and performances) of 256 untrained subjects to those of a panel of 12 trained listeners. Since he found that the preference scores of the two groups of respondents were very similar, his study validates the use of expert listeners for product evaluations on the grounds that their preferences can be extrapolated to the intended (untrained) customer population. Also, the trained listeners turned out to be more discriminating and reliable than the untrained ones. In accordance with Bech, Olive therefore deduced that training and experience in controlled listening tests lead to significant gains in subject performance so that fewer listeners are required to achieve a certain statistical power.

Overall then, there are good reasons for putting the necessary resources into setting up a training programme. Thus, it does not come as a surprise that in the area of sound reproduction several programmes have been devised over the years (e.g. [Letowski, 1985; Miskiewicz, 1992; Brixen, 1993; Olive, 1995; Quesnel, 1996; Olive, 2001]). Yet, despite all this activity with regard to the design of ear-training tools, the associated work has focused almost entirely on timbre perception. Apart from
some preliminary steps described in [Koivuniemi & Zacharov, 2001] (see also Section 0.3), a lack of such research is therefore apparent as far as the development of a spatial ear trainer is concerned.\footnote{In this context, it should be noted that at McGill University a system has been developed for the construction of multichannel sound scenes (see Section 4.5), which was also expected to be useful for training listeners in spatial sound assessments [Quesnel et al., 1999]. However, no work has been published in that respect so far.}

0.3 The need for unidimensionality

Regardless of the potential benefits mentioned above, the investment in training makes sense only if (1) it can be implemented in an efficient and cost-effective manner and (2) it can be demonstrated to produce quantifiable improvements in the reliability of the listeners' ratings [Olive, 1995]. It seems logical that in order to satisfy these requirements, a training programme needs to be built on clear and unmistakable demonstrations of the perceptual effects of interest. Otherwise, if ambiguous examples are employed for teaching purposes, subject confusion is likely to occur, which would probably prolong the training process and also impair the training effect. Support for this assertion is available from the literature. For instance, in their textbook on sensory evaluation practices Stone and Sidel [1993] commented on the benefits and problems associated with the use of "references" in training situations. Reference stimuli are valuable when training and retraining subjects in using a descriptive language, because they help listeners to relate to particular sensations that are not easily described or detected. Also, when adding new subjects to a panel they are useful as they allow these individuals to experience what other subjects are talking about, thereby enabling them to contribute to the panel's vocabulary. That is why anchor and training samples are commonly applied in various sensory disciplines (e.g. see [Noble et al., 1987] for a good example in the context of wine assessments).

However, whilst the use of exemplary stimuli seems like an ideal means of focusing subject responses, care is needed when selecting them. In this respect, Stone and Sidel [1993] emphasised that references should not, in themselves, be a major source of variability. They suspected that extensive training time can at least partly be explained by the utilisation of reference stimuli that introduce other attributes unrelated to their purpose. As a result of these sensory interactions, subjects might take considerably longer to learn to discriminate perceptual characteristics and hence to master the associated language. With regard to finding suitable references, it was pointed out that even though a reference sample can be a product that "represents" a particular attribute, "in most training situations, the most helpful references usually are a product's raw materials" or "ingredients". This statement implies that stimuli are required, which can illustrate the individual perceptual dimensions of the field of interest.

Thus, it follows that unambiguous samples are advantageous for training applications in the sensory sciences, as they facilitate the development of a descriptive language and help subjects gain control over it. What is more, such samples could also be of great value to other perception-based work. For instance, when discussing methodological issues in timbre research, Hajda et al. [1997] acknowledged...
that systematically varied timbre stimuli could lead to a better understanding of that perceptual domain. Nonetheless, they also stated that “because of its multidimensional nature, it is difficult to manipulate timbre in a controlled scientific fashion”, which is why “timbre continues to be a poorly understood attribute of music”. To allow experiments to be designed that can help reveal timbral structures, Hajda et al. argued for the perceptual isolation of timbre so as to allow it to be manipulated independently. This necessitates that stimuli be equalised with respect to any additional, confounding psychological changes, whereby the success of the applied equalisations would have to be assessed by means of a formal, perception-based empirical method. Yet, a method suitable for this task does not appear to be readily available.

Evidence for the difficulty in creating stimuli that are free from sensory interactions is also available from the area of spatial audio assessment. Koivuniemi and Zacharov [2001] recognised the need for unequivocal sound examples, as these could be applied as scale anchors, which in turn would facilitate familiarisation and training of subjects in the use of their attribute scales. However, they encountered problems with the creation of such samples, because “it is often difficult to isolate an attribute, and to scale it”. This they traced back to the many psychoacoustic interdependencies that appear to exist (see Chapter 2). Failing to overcome such predicaments, Koivuniemi and Zacharov decided to be less stringent regarding the perceptual effects of their sounds and to produce stimuli that merely “indicated” the attributes in question.

In consideration of Stone and Sidel’s call for interaction-free references (see above), Koivuniemi and Zacharov’s solution is not optimal, especially if subjects are to give meaningful scores to specific spatial characteristics of a stimulus in the presence of other attributes. Ideally, the sound examples have to be clearly related to a particular attribute or, put differently, they should illustrate a unique perceptual facet in an unambiguous fashion. This desirable sensory property can be described succinctly by the term ‘unidimensionality’.

0.4 Mathematical vs. perceptual unidimensionality
The term ‘unidimensionality’ demands closer scrutiny because its precise meaning depends on the context it is used in. In order to avoid confusion, a distinction between perceptual unidimensionality (as applicable to sensory work) and mathematical unidimensionality is believed to be valuable. Consider the geometrical space that is defined by two orthogonal axes labelled ‘Direct sound level’ and ‘Reverberant sound level’ (see Figure 0.2). Movement of a point in a direction parallel to either one of these axes constitutes a mathematically unidimensional variation of the characteristics of that point — a variation which would also have physical meaning. Within the same space, there is a perceptually relevant ‘Source distance’ axis, too. This axis can be defined in terms of ‘Direct sound level’ and ‘Reverberant sound level’ and is therefore orthogonal to neither of them. Figure 0.2 also shows ‘Dimension X’ — an arbitrary axis, again orthogonal to neither ‘Direct sound level’ nor ‘Reverberant sound level’. Movement of a point in a direction parallel to either of these two additional
axes could again be said to constitute a mathematically unidimensional variation, but only movement along the 'Source distance' axis would be perceptually unidimensional.

Figure 0.2: Graphical illustration of a conceivable relationship between four mathematical dimensions, one of which is a perceptual dimension ('Source distance'), and two of which are physical dimensions ('Direct sound level' and 'Reverberant sound level')

It is also important to note that even though on their own 'Direct sound level' and 'Reverberant sound level' may be independent of each other, they lose this independence when dealing with the attribute 'Source distance'. This is because in order to ensure that this spatial facet changes in a perceptually unidimensional fashion, 'Direct sound level' and 'Reverberant sound level' cannot be varied independently but have to be adjusted conjointly to achieve an authentic distance effect.

Thus, the notion of perceptual unidimensionality also implies that one may have to take into account several 'sub-attributes' when attempting to emulate a particular sensory characteristic. This issue was indirectly addressed by Bech [1999] who presented a conceptual model of human perception of multidimensional (auditory) stimuli (see Figure 0.3). He proposed that a given stimulus can be described by a number of physical variables $\Phi$, which are transformed into sensory attributes $\Psi$ by the associated sensory system of a subject exposed to that stimulus. Each of these attributes will have a sensorial strength $S$ that will depend on the magnitudes of the (relevant) physical variables and the characteristics of the sensory system. Based on these sensorial strengths as well as cognitive effects and other mental processes, the subject will arrive at a certain magnitude of impression $I$ for each of the sensory attributes. These individual impressions are then combined – perhaps with different weightings – to form the total sensory impression $I_{tot}$ that could be the basis for a similarity judgement, for example.

To relate Bech's model to the concept of perceptual unidimensionality, if the aim is to produce a number of (multidimensional) reference stimuli that are meant to vary in terms of only one "individual
impression" (e.g. source distance), several physical factors (e.g. the level of the direct and indirect sound) may have to be brought into agreement with each other. Otherwise subjects might perceive these sub-attributes and not the intended holistic effect. Going back to Figure 0.2, improper weighting of ‘Direct sound level’ and ‘Reverberant sound level’ would mean a rotation of the ‘Source distance’ axis, thus making it perceptually multidimensional and therefore less intuitive in a training situation, if not perceptually meaningless. If, on the other hand, the sub-attributes are combined correctly, then the simulation will be successful. Hence, all sub-attributes will be perceptually amalgamated by the sensory system, thereby giving rise to a unified gestalt perception. In this case the simulation would be said to exhibit perceptual unidimensionality. These two possible scenarios are illustrated at the bottom of Figure 0.3, using the attribute of ‘cherry pie’ as an example. Evidently, in order for the (holistic) impression of cherry pie to be created, various ingredients have to be added in a certain way and then carefully cooked.

Figure 0.3: Conceptual model of human perception of multidimensional stimuli (adapted from [Bech, 1999]). If the aim is to create a set of stimuli that differ only in their intensities with respect to one individual sensory impression (or attribute), all associated sub-attributes need to be closely concerted. If agreement is not achieved, a subject presented with that attribute simulation might perceive the various sub-attributes or ‘ingredients’ (e.g. milk, salt, cherries, eggs etc.). Conversely, if these ingredients are added in the right amounts, they will be perceptually fused by the sensory system, resulting in a unified gestalt perception (e.g. cherry pie). If the latter is the case, the attribute simulation will be considered perceptually unidimensional.
To recap then, perceptual unidimensionality, as used within the context of this thesis, implies (1) that an attribute simulation is free from sensory interactions and (2) that it evokes a holistic impression in a subject exposed to it. The reader should be aware, though, that the term ‘unidimensional’ is used broadly in relation to psychological constructs. Classical psychophysics, for instance, is generally concerned with the relationships between physical stimuli and relatively low-level attributes such as loudness. Thus, in this case a single construct (or dimension) will exist at a low perceptual level. Yet, some of the research reported in this thesis deals with higher-level attributes comprising numerous sub-components. The precise meaning of ‘unidimensionality’ therefore depends on the degree of abstraction between physical aspects of the stimulus and the psychological construct.

For reasons of simplicity, for the remainder of this thesis ‘unidimensional’ will be used with reference to the concept of perceptual unidimensionality.

0.5 Aims of the research

In view of the information provided above, it is apparent that an ear-training toolkit would be very useful to the field of spatial audio assessment as it could speed up subjective evaluations and make the resultant data more reliable. Therefore, the development of such a system is proposed. However, in order to ensure an optimal training effect, the system needs to be based on attribute simulations that can convey the intended effects to listeners in a unidimensional way. Such emulations could also further comprehension of the physical cues that give rise to the associated qualitative phenomena, thereby facilitating other research in the field of spatial auditory perception, e.g. the development of physical measures. As was pointed out before, interaction-free (spatial) attribute simulations have yet to be realised, which is why this study was designed to address this deficiency. Specifically, the aim of this research is to answer the following questions:

- Is it possible to vary spatial attributes of reproduced sound in a unidimensional manner?
- How can this be achieved in practice, i.e. what physical parameters should and should not be varied?
- How can unidimensionality in variation be validated?

A logical first step towards achieving this goal is to review previous research into correlations between spatial attributes and sound field characteristics, as knowledge of such psychoacoustic connections can possibly assist with the simulation of certain attributes. Since it may well be that existent spatial sound-processing systems can be used for this purpose, it makes sense to investigate these, too. Moreover, a sensory evaluation technique needs to be found, which is able to provide a detailed sensory profile of an attribute simulation. Based on the findings of this research, spatial

---

3 To prevent misunderstandings, it is emphasised that ‘interaction-free’ is used to denote that an attribute simulation does not contain perceptual changes, which are not part of the intended effect. It does not refer to the absence of interaction of such attribute simulations with each other (see also Section 6.2.4).
attributes can be modelled and then validated so as to assess empirically the ability of the simulations to evoke particular, well-defined percepts in experienced listeners.

It is anticipated that such attribute simulations will provide a good basis for the development of the envisaged spatial ear trainer. Evidently, further work would then be required to arrive at a versatile and effective training system eventually. For instance, the devised simulations might need to be integrated into a single processing environment to facilitate the design of a stand-alone training program. Moreover, such a framework could be equipped with a high-level graphical interface offering a user simultaneous control over the modelled attributes in real-time. It might also be beneficial to implement statistical analysis and feedback modules to enable quantification as well as monitoring of listener performance. These and other ‘further development’ issues will be discussed in more detail in Section 6.2 and Appendix E.

0.6 Structure of the thesis

The structure of the remainder of this thesis is as follows. It starts by introducing the psychological organisation of spatial quality (Chapter 1). Pertinent attribute elicitation studies are summarised, thereby depicting the individual constituents of this perceptual domain and determining associated terminology. This is followed by a comprehensive review of established relationships between discrete components of spatial quality and physical parameters (Chapter 2) so as to provide a basis for the attribute simulations executed as part of this project. Chapter 3 presents different sensory evaluation paradigms and tries to find the most appropriate method for verifying unidimensionality in variation of a set of reference stimuli. Chapter 4 investigates whether a perceptual or a physical attribute simulation approach should be adopted for this project and whether or not any systems already exist that allow manipulating spatial attributes of reproduced sound in a unidimensional manner. Chapter 5 represents the body of this thesis because it contains almost all of the practical work that was conducted. More specifically, it aims to determine if unidimensional simulation of spatial attributes of reproduced sound is possible, and how this can be accomplished for a number of such constructs. The content of this thesis is summarised in Chapter 6 and the main conclusions that can be drawn from the results of this research are reiterated. Also, possible avenues for further work are discussed. The appendices include detailed listener responses collected as part of some of the psychoacoustic studies executed during the course of this project. In addition, a preliminary investigation into training listeners in the evaluation of spatial sound reproduction is outlined that utilised some of the stimuli created by this work. Finally, a glossary is provided to explain the meanings of the abbreviations and some of the special terms used.
0.7 Original contributions

As a result of the research undertaken for this thesis, a number of original contributions have been made to the fields of spatial auditory perception, sensory evaluation and subject training. The more significant of these are summarised below:

- A formal, perception-based validation strategy has been devised as an empirical means of verifying the unidimensionality of sensory attribute simulations (see Chapter 5). This strategy relies on a novel combination of different subjective testing paradigms.
- As part of the development of this strategy, limitations of Multidimensional Scaling techniques have been highlighted, which had not previously been mentioned in the relevant literature (see Chapter 5). Workarounds for these shortcomings have been identified.
- Unprecedentedly, reference stimuli have been created that were shown to illustrate several attributes of spatial sound display to critical listeners in a unidimensional fashion (see Chapter 5).
- The achievement of unidimensionality in variation implies that the employed attribute scales must be unidimensional themselves (see Chapter 1). This could not be inferred from the results of the elicitation studies, which had given rise to these attribute scales.
- A new multichannel panning technique, designed by Peter Craven, was subjectively assessed and found to be superior to 5-channel pairwise constant power panning in terms of two perceptual characteristics (see Chapter 5).
- A novel processing platform has been implemented in software (see Chapter 5), which makes use of this panning technique and which is appropriate for use as the framework for a stand-alone spatial ear-training system.
- With regard to psychoacoustic relationships, the subjective effects of varying the physical parameters of left-right time and azimuth differences, azimuthal distribution and source specificity of early reflection patterns were ascertained for the first time (see Chapter 5).
- New insights were gained with respect to the perceptual salience of finer details of early reflection patterns and the sound field properties of intra- as well as inter-source 'syllabicites' (i.e. discontinuous amplitude envelope characteristics) to ensemble depth hearing (see Chapter 5).
- A pilot study has established the potential benefits of training naïve listeners in the discrimination of selected spatial attributes (see Appendix E).

0.8 Summary

This introductory chapter delineated the background to the research presented in this thesis so as to put the following work into context. In particular, the necessity to assess 3-D sound displays with respect to their spatial performance by means of subjective testing methods was determined. In
addition, training studies related to timbre perception were briefly reviewed and evidence was presented asserting that a systematic training programme can improve specific auditory skills of listeners. In the area of spatial audio the lack of an ear trainer was identified, which is why the development of such a system was proposed, and important requirements regarding its characteristics were established. In conformity with the sensory literature, it was deduced that to ensure an optimum training effect, attribute simulations are needed which are free from sensory interactions. Furthermore, to maximise the simulations’ intuitiveness and hence to enhance the learning process, it was argued that they should evoke a unified gestalt perception in listeners. Based on these two criteria, the concept of ‘perceptual unidimensionality’ was formulated. The objectives of this work were then specified, the main one being the achievement of validated, (perceptually) unidimensional simulations of spatial attributes of reproduced sound for training purposes. This was followed by a short description of the basic approach adopted to achieve this goal as well as an overview of the structure of this thesis. Finally, the original contributions of this research were listed.
THE MULTIDIMENSIONAL NATURE OF SPATIAL QUALITY

Sound quality has been assumed to be a multidimensional phenomenon for a long time. In the field of concert hall acoustics, researchers identified and studied fundamental components such as timbre, loudness and spatial impression (e.g. see [Barron & Marshall, 1981; Beranek, 1996; Schroeder et al., 1974]). Some of that work also addressed more than one spatial phenomenon, thereby suggesting spatial sound perception to be a perceptually complex process itself (see [Zacharov & Koivuniemi, 2001a; 2001b] for a useful overview of relevant research results). As regards the realm of reproduced sound, these findings were adopted and scrutinised (e.g. see [Gabrielsson & Sjögren, 1979]). Due to the increased interest in multichannel audio in the recent past, work has commenced in this particular area as well, aiming to break down spatial quality into its discrete components. This chapter will summarise results of the associated elicitation experiments, which were executed by Berg and Rumsey, whose investigation will be reviewed first, and by Zacharov and Koivuniemi, whose work is to be discussed subsequently. A scene-based paradigm, proposed by Rumsey for spatial audio evaluation purposes, will then be introduced with the intention that it might serve as a conceptual framework for this project. Overall, this synopsis may therefore be seen as some kind of a yardstick for the envisaged ear trainer, since this tool should ideally take into account all dimensions salient to the perception of spatial sound display.

1.1 The Berg and Rumsey studies

Berg and Rumsey [1999a; 1999b; 2000a; 2000b] carried out research with the objective of extracting the individually perceivable constituents of spatially reproduced sound and their relative weights. In an ideal case, these dimensions would be independent of each other and related closely to physically measurable parameters of a sound field, as this should facilitate perceptual optimisation of 3-D audio systems. Regarding the experimental design, test materials were created by recording or electronically processing six sound events in various ways and acoustical surroundings (see [Berg & Rumsey, 1999a] for details). These were played back over a 3/2-stereo sound system, utilising different numbers of (active) reproduction channels per test material. Thus, 18 stimuli were obtained that differed widely in terms of their spatial properties. Making use of an attribute elicitation method called the Repertory Grid Technique (see Section 3.2.2 for details), Berg and Rumsey instructed 18 panellists to listen for differences and similarities between groups of three of these stimuli and to verbalise their perceptions. This yielded more than 300 semantic descriptors based on which listeners had to make direct attribute ratings of the stimuli. In order to group terms with similar meanings together and hence uncover the stimuli's perceptual organisation, the obtained data were subjected to Principal Component Analyses [Berg & Rumsey, 1999b; 2000b] (see also Section 3.2.2) and Cluster Analyses [Berg & Rumsey, 2000a; 2000b] (see also Section 3.3.1). The revealed structures were examined with respect to their meaning and used subsequently for deriving 11 attributes (see [Berg & Rumsey, 2000a] for details).
To validate their findings Berg and Rumsey [2002] conducted an additional experiment, which comprised another elicitation exercise that was very similar to the first one, the main difference being that different listeners and more homogeneous stimuli were employed this time. The elicited constructs were compared to the results from the first study, bringing to light five new attributes. Hence, a revised set of attributes was produced based on which the panel then graded the sound examples. Data analysis showed that the listeners could consistently evaluate the stimuli in terms of the elicited characteristics, thus showing them to be appropriate descriptors for assessing spatially reproduced audio. Table 1.1 presents ten of these attributes (non-spatial ones have been omitted) together with their associated definitions.

<table>
<thead>
<tr>
<th>Spatial attribute</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>How similar to a natural (i.e. not reproduced) listening experience is the sound as a whole?</td>
</tr>
<tr>
<td>Presence</td>
<td>The experience of being in the same acoustical environment as the sound source.</td>
</tr>
<tr>
<td>Ensemble width</td>
<td>The perceived width/breadthness of the ensemble from its left to its right flank. Excludes sounds coming from the environment.</td>
</tr>
<tr>
<td>Source width</td>
<td>The perceived width of an individual source/the angle occupied by this source. Excludes sounds coming from the environment.</td>
</tr>
<tr>
<td>Localisation</td>
<td>How easy is it to perceive a distinct location of a source? How easy is it to pinpoint the direction of a source?</td>
</tr>
<tr>
<td>Source distance</td>
<td>The perceived distance from the listener to the source.</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>The extent to which the source envelops/surrounds the listener. Excludes sounds coming from the environment.</td>
</tr>
<tr>
<td>Room width</td>
<td>The width of angle occupied by a sound source’s reflections in the room. Excludes the direct sound of a source.</td>
</tr>
<tr>
<td>Room size</td>
<td>In cases where a room/hall is perceived, this denotes the relative size of that room.</td>
</tr>
<tr>
<td>Room envelopment</td>
<td>The extent to which a sound source’s reflections envelop/surround the listener. The feeling of being surrounded by reflected sound.</td>
</tr>
</tbody>
</table>

Although not strictly spatial in nature, ‘naturalness’ has been included here to highlight the fact that it formed a major part in Berg and Rumsey’s results. Inevitably, the other attributes contribute to the impression of naturalness, e.g. for a sound source to be perceived as being authentic it has to have a certain width and distance from the listener. Consequently, interdependency between some of the scales can be attested. To examine attribute interrelations in more detail (and hence better understand the dimensionality of the stimulus set as perceived by the listeners), correlation analyses and Factor Analyses (see also Section 3.2.2) were performed on the collected data [Berg & Rumsey, 2002]. Whilst no strong relationships were revealed, the attribute pairs of ‘source width – ensemble width’, ‘source width – localisation’ (negative correlation) and ‘source width – source envelopment’ were found to be moderately correlated.
Going back to the constructs in Table 1.1, another observation can be made. Almost all of them refer to purely descriptive (i.e. objective, value-free) characteristics of sound scene components, e.g. 'source width', 'source envelopment' or 'room size'. An additional Verbal Protocol Analysis (see Section 3.2.1) (performed on the responses obtained from the first elicitation experiment) showed that descriptive terms constituted two thirds of all the verbal data whereas the remaining ones were found to be either of the "emotional-evaluative" or "artificial/natural" type [Berg & Rumsey, 2000a]. By means of a Cluster Analysis Berg and Rumsey [2000b] were able to detect correlations between these three attribute classes, thereby obtaining some information about listener preference. They found that a stimulus containing enveloping (reflected) sound conveyed a sense of the listener being present at the recording space and hence led to a positive experience. Conversely, subjects regarded non-enveloping stimuli as sounding unnatural. Similarly, phantom mono (as opposed to multichannel) reproduction resulted in a lack of room and width perception, which was not preferred by Berg and Rumsey's listeners. While none of these results is really surprising, a more unexpected finding was that the localisation attribute turned out not to be strongly correlated with naturalness and positive sensations.

1.2 The Zacharov and Koivuniemi studies

Recently, Zacharov and Koivuniemi [2001a; 2001b; 2001c] performed an investigation that was also concerned with trying to determine how subjective opinion of spatial sound display is formulated. As to their applied strategy for generating suitable stimuli, Zacharov and Koivuniemi chose three microphone techniques that were different from the ones used by Berg and Rumsey (see [Zacharov & Koivuniemi, 2001a] for details). Using these, 13 miscellaneous sound events and acoustical environments were recorded and rendered with the help of eight reproduction formats (many of which had not been used in Berg and Rumsey's study). This enabled a broad range of spatial sound characteristics to be captured and reproduced. As a result, as many perceptual constructs as possible were excited during the elicitation stage, which comprised a single stimulus paradigm (for absolute attribute elicitation), followed by a multiple stimulus comparison test (for differential attribute elicitation) [Koivuniemi & Zacharov, 2001]. Each of 12 subjects was required to listen to the stimuli as often as (s)he needed to verbalise all detectable auditory features so as to ensure that every possible aspect of spatial sound experience was covered. A total of 1400 descriptors were elicited and redundant ones removed by filtering out identical and (grammatically) highly similar terms. In accordance with the principles of descriptive analysis (see Section 3.1), group discussions were then conducted in order to further evaluate the remaining 532 terms and to arrive at a manageable set of consensual attributes that could satisfactorily describe all stimulus characteristics. This very lengthy and detailed process of developing a descriptive language resulted in eight spatial attributes, which are displayed in Table 1.2 together with their corresponding definitions.

4 Four timbral attributes that were also derived are not shown.
1. The multidimensional nature of spatial quality

Comparing Table 1.2 to Table 1.1, it can be seen that Zacharov and Koivuniemi obtained results similar to the ones reported by Berg and Rumsey. Again, descriptive terms dominate the picture. What is more, many attributes seem to have a relative in Berg and Rumsey’s list of elicited constructs (e.g. ‘sense of direction’ and ‘localisation’ or ‘sense of space’ and ‘presence’). This is remarkable since the sound events, recording techniques, reproduction formats and attribute elicitation methodologies were all substantially different. As a consequence, the findings of both studies are strengthened. The large overlap between the two sets of results further implies that most of the elicited spatial attributes should be externally valid and hence can be used for other spatial audio applications such as the design of an ear-training tool.

It is also worth pointing out that Zacharov and Koivuniemi [2001b] found their developed scales to be employed consistently by the subjects, thus providing them with good confidence in both their definition and usage. Yet, as was the case with Berg and Rumsey’s study, some attribute scales appear to be logically related with regard to their semantic meanings. Nonetheless, Zacharov and Koivuniemi [2001c] were able to show that many of their scales are “very poorly correlated”. Not only is this a significant finding for the audio engineer who seeks to improve sound displays in terms of their spatial performance, it is also reassuring for this work to see that the acquired scales are statistically highly independent of each other as this should pave the way for attribute simulations that are free of sensory interactions.

Like Berg and Rumsey, Zacharov and Koivuniemi also reported some results with respect to their elicited attributes’ relatedness to listener preference. Before engaging in the descriptive language development process, preference judgements of the stimuli had been collected from the panel. In addition, all listeners had graded each stimulus on the developed scales. Using a statistical technique
1. The multidimensional nature of spatial quality

called Partial Least Squares Regression (e.g. see [Martens & Martens, 2001]) Zacharov and Koivuniemi [2001b] were able to establish relationships between these preference and direct attribute ratings. To be more specific, ‘penetration’, ‘distance’ and ‘depth * distance’ (the * denoting interaction) turned out to be the least desirable spatial characteristics. However, it was also emphasised that their preference mapping model was not fully validated in terms of its prediction ability [Zacharov & Koivuniemi, 2001b] and that a superior model may be possible with a larger set of preference data [Zacharov & Koivuniemi, 2001c]. Therefore, it may be concluded that until these results have been replicated with more data sets one should refrain from making inferences about the perceptual importance of Zacharov and Koivuniemi’s attributes – especially because such findings are “likely to be highly context and subject dependent” [Rumsey, 2002]. This statement also holds true for Berg and Rumsey’s listener preference findings (see Section 1.1).

Finally, it has to be stressed that if scales are found not to be completely independent this does not preclude their unidimensionality5. The moderate correlation obtained by Berg and Rumsey and by Zacharov and Koivuniemi for some of their scales might have been due to the stimuli that were used during the elicitation stage. It is very likely that the sound examples did not exhibit unidimensional variation in terms of the elicited spatial characteristics, thus prohibiting listeners to rate these changes independently. As part of this project, however, some of these scales were used as the basis for simulating the corresponding attributes (see Chapter 5). Since the simulations were found to be unidimensional, it follows that this must also be the case for the scales themselves. This constitutes a novel contribution to the field of spatial sound perception.

1.3 Rumsey’s scene-based paradigm for spatial audio evaluation

As the above summary of the two elicitation studies indicated, the associated findings are potentially very helpful for spatial audio engineers who until recently have not been able to rely on such palpable insights with regard to their domain’s perceptual structure. As a matter of fact, earlier work has frequently suffered from an incomplete vocabulary of spatial attributes that have often been ill-defined, thus hindering advancements in this field [Rumsey, 2002]. Endeavouring to circumvent problems due to non-standardised terminology and hence to lay the foundation for experimental work producing comparable results, Rumsey proposed a scene-based paradigm for the description and assessment of spatial quality based on his own and other people’s research. Its (primarily descriptive) attributes were derived by categorising elements of a reproduced sound scene according to their function within that scene and at levels appropriate to the task. That is why they can be grouped into micro- and macro-attributes, whereby the former describe spatial features of individual elements whilst the latter refer to perceived properties of the sound scene as a whole. In Figure 1.1 several spatial attributes, as defined by the scene-based paradigm, are illustrated in graphical form. For

5 And neither would complete independence guarantee the scales to be unidimensional for that matter. Two attributes that are multidimensional in nature (e.g. ‘frontal image’ and ‘spatial impression’) might still enable listeners to grade stimuli independently along such scales [Rumsey, 1998].
reasons of clarity, the arrows indicating the various scene components have different sizes to
demonstrate the transition from the micro-level (i.e. individual sources, small arrows) to the macro­
level (i.e. the entire environment, large arrows) with a third level – the ensemble-level (i.e. groups of
sources, medium-sized arrows) – in-between. The reader should note the distinction between the
notions of ‘source depth’, ‘ensemble depth’, and ‘environment depth’ and how these compare to the
concept of ‘source distance’. Likewise, there are important differences between ‘source width’,
‘ensemble width’ and ‘environment width’. Such precise definitions are expected to be beneficial to
subjective assessments of spatial sound reproduction insofar as they should enable listeners to
generate more consistent responses. Support for this assertion is available from Berg and Rumsey
(2003) who observed that when data from spatial audio tests exhibited low consistency across subjects
this was mainly attributable to confusing semantics being used. In particular, both unclear and too
“wide” (i.e. multidimensional) descriptors/definitions led to confusion amongst listeners, whereby the
less precise the chosen language was the more uncertainty it produced.

Figure 1.1: Graphical illustration of the concept of source distance as well as
different types of width and depth attributes (as defined in [Rumsey, 2002]). The
different arrow sizes indicate the transition from micro- to macro-level components
of a spatial sound scene.

At the time of writing, the scene-based paradigm represents the only systematic approach to profiling
spatial quality that was specifically aimed at overcoming deficiencies of previous work. Therefore, it
makes sense to adopt its concepts and definitions for this project. Yet, the paradigm must not be
regarded as the ultimate spatial quality lexicon. Rather, it should be seen (and used) as a means to an
end, namely to help further research in this area. As such, it has to adapt to new insights. Experience
from other sensory disciplines suggests that ascertaining the number and identity of a sensory
domain’s perceptual dimensions is usually a long-winded process, which is also true as far as the
development of a suitable, internationally approved vocabulary is concerned. For example, referring
to timbre perception, Berg and Rumsey [1999a] pointed out that:

"The dimensionality of 'timbre space' has been hotly debated ... for some thirty years
or more, without satisfactory conclusions, yet there are numerous researchers around the
world using a variety of attribute scales for subjective experiments on sound timbre,
each with differing degrees of usefulness and applicability".

With respect to the relatively young field of spatial audio, future discoveries are anticipated to lead to
a reworking of currently accepted viewpoints. Indeed, work carried out as part of this project shows
that the 'ensemble width' definition of Rumsey's paradigm is not sufficiently detailed for a
completely unmistakable classification of the associated spatial effect (see Section 5.9.1).

1.4 Summary
This chapter summarised the results of attribute elicitation experiments carried out in the context of
spatial sound display, thereby setting a benchmark for the envisaged ear trainer. Two independently
conducted studies confirmed the existence of several spatial dimensions that are mostly descriptive in
nature. More precisely, they describe discrete sound scene components such as the distance, depth or
width of single or groups of sources. Since a hierarchical system of spatial attributes is available that
(1) was built on these and other pertinent findings and (2) was designed to remedy inadequacies
evident in previous research, it has been embraced as a conceptual framework for this study. As for
the spatial attributes' relations to subjective preference (and hence their perceptual importance), more
research has to be carried out to allow valid inferences to be made. This is because a scale's adequacy
as a predictor for listener preference will strongly depend on both the context it is used in and the
subject it is used by.
2. Correlations between spatial attributes and physical factors

2 CORRELATIONS BETWEEN SPATIAL ATTRIBUTES AND PHYSICAL FACTORS

Having introduced the various components of spatial quality, this chapter will review established relationships between these attributes and their physical correlates. It is expected that these will be applicable to or at least offer a good starting point for the unidimensional attribute simulations to be tackled in Chapter 5. Although an effort has been made to structure this chapter in terms of these dimensions, separation has not always been possible because many interdependencies appear to exist. Still, the following systematisation of pertinent findings should help to make clear distinctions ultimately possible.

2.1 Source direction

Without a doubt the direction of a sound event is the most researched spatial attribute, the first reported scientific studies dating back about 100 years [Rayleigh, 1907]. As a result, a vast amount of literature is available that documents the underlying lateralisation mechanisms in great detail (e.g. [Blauert, 1997; Moore, 2003]). Since a discussion of associated findings would be beyond the scope of this thesis, no summary has been attempted here. In this respect, it should also be noted that the angle of incidence of a source is only of secondary interest to this work, which focuses on some comparatively unexplored attributes of spatial hearing.

2.2 Source distance

Even though not in the same league as source direction, distance perception has been the subject of a considerable number of investigations. While a consensus seems to have been reached with regard to the most salient distance cues, some disagreements appear to exist concerning the importance of less well understood psychoacoustic relationships for judging the proximity of a sound source. Generally speaking, monaural and binaural cues are available to the human auditory system for distance evaluation [Blauert, 1997]. Essential monaural attributes of ear-input signals are time and level differences between individual spectral components of each ear-input signal. The elementary interaural signal attributes are time and level differences between corresponding spectral components of the two ear-input signals. For a long time, the scientific community took the stance that apparent source range is mainly estimated based on information given by monaural signal attributes. In recent years, however, this view has been challenged by claims that non-medial reflections are crucial for hearing a distance effect. An outline of the various monaural and binaural cues is given below.

2.2.1 Frequency content

Changes related to the frequency content of a direct sound (DS) belong to the category of monaural cues. For the case of a sound source positioned in the median plane and radiating a constant broadband signal the available cue depends on the distance of the sound event. If the source is located
far away, i.e. 15m or more, high-frequency (HF) loss due to air absorption can be distinctly audible above 10kHz [Blauert, 1997]. Analytical expressions for calculating the magnitude of the air absorption effect as a function of frequency, temperature, humidity and pressure have been published in [ISO 9613-1, 1993]. Using these standardised equations, Savioja et al. [1999] estimated the magnitude responses for six different source-receiver distances, the results of which are shown in Figure 2.1. It follows that this phenomenon can be modelled as a low-pass filter whose cut-off frequency decreases and slope increases in parallel with decreasing source proximity.

![Figure 2.1: Magnitude of air absorption as a function of distance and frequency for an air humidity of 20% and a temperature of 20°C (adapted from Savioja et al., 1999)](image)

As far as proximal sources are concerned, it has been suggested that an increase in low-frequency (LF) sound energy can take place [von Békésy, 1960]. A wavefront radiated by a distant source is usually planar by the time it reaches a listener, but for a nearby source it will exhibit curvature. According to Begault [1994], this results in an added emphasis to lower versus higher frequencies. In this regard, Blauert [1997] acknowledged that for sound sources closer than 3m spectral distortions of the signal may occur, but he also pointed out that their influence is small for distances greater than 25cm. Moreover, he stressed that the specific spectral attributes evaluated by the auditory system for close sources still need to be determined.

From the above, it appears that only the simulation of HF roll-off is worthwhile for this work. For reasons outlined in Section 5.2, an ITU-compatible reproduction set-up is to be used, which is why the spectral changes applicable to close sources do not need to be taken into account⁶. In any case, inclusion of such frequency-related changes in a distance simulation would only make sense if the resultant audible effects were known.

---

⁶ The preferred minimum loudspeaker distance for an ITU layout is 2m [ITU-R BS.1116, 1997]. Since no special (e.g. transaural) processing is to be applied for this work, the maximally possible source proximity will be limited to the radius of the reproduction set-up.
2. Correlations between spatial attributes and physical factors

2.2.2 Binaural cues under free-field conditions

Although interaural signal properties are primarily correlated with lateral displacements of auditory events, they may also supply distance cues. Research related to the development of virtual auditory displays showed that a change in the distance of a source within 1m of the head of a listener causes significant changes in interaural level differences (ILDs) if the source is displaced to one side of a listener [Shinn-Cunningham, 2000]. For sources positioned along the median plane ILDs are essentially zero, regardless of source distance. This also applies for lateral sources more than ~1m away from the head. In practice, the modelling of these binaural source-proximity cues is complicated by the fact that their strength varies with source direction. However, Shinn-Cunningham pointed out that ILD cues do not seem to contribute to distance perception when reverberation is included in a simulation and that even large ILDs incorporated into anechoic headphone simulations do not lead to robust distance percepts. Hence, it may be concluded that (even under free-field conditions) binaural distance cues provided by the DS are limited in their effect and therefore can be considered to be of minor importance – a view also endorsed by Begault [1987].

As was already mentioned, the chosen reproduction format (see Section 5.2) prevents simulation of such nearby sources. Besides, due to the strong dependence of distance perception on the presence of reverberation (see below) inclusion of indirect sound in a distance simulation is virtually mandatory, which is why this type of ILD can be safely disregarded. Yet, since lateral reflections also give rise to binaural differences, it is of interest to find out if these interaural cues are perceptually more salient to distance hearing. This aspect will be discussed in Section 2.2.5.

2.2.3 Loudness

As evident from numerous psychoacoustic studies, loudness perception is related to sound pressure level (SPL). Under free-field conditions, all the sound energy is radiated away from the source and none of it is reflected. Thus, the inverse distance (or 1/r) law entirely dictates the SPL for any source location. This implies a drop of 6dB per doubling of distance (see Figure 2.2), which will be less in non-anechoic environments, the exact value depending on the diffusivity of the sound field [Blauert, 1997] (see also Figure 2.3).
The fact that variations in loudness also give rise to changes in tone colour is one example of the complexity of the subject of distance perception. Perceived loudness rises with increasing signal level at the ears whilst at the same time the tone colour becomes 'darker'. This can be explained by means of the well-known equal-loudness contours, i.e. as the level is raised the LF components of a broadband signal gain more perceptual weight relative to the mid- and high-frequency components. Nevertheless, Blauert [1997] stated that at best the relationship between loudness and tone colour serves as an auxiliary parameter for distance perception.
Chomyszyn [1992] intended to clarify whether distance judgements of sound sources in a reflective environment are still precise after their loudness has been matched and therefore eliminated as a cue. An experiment was conducted whereby an anechoically recorded sound was played back in a reverberant environment at several distances and recorded binaurally using an artificial head. The recordings were reproduced over headphones and subjects were instructed to judge which stimulus of a given pair was perceived to be closer (Stage 1). For each pair the subjects also had to adjust the loudness of the closer sound to match that of the more distant one. The obtained attenuation factors were used to create equally loud stimuli and subjects were asked to estimate which sound of each pair appeared to be closer when listening to them a second time (Stage 2). The two sets of data were then compared in order to determine the influence of removing the loudness clues. The results of Stage 1 showed a high percentage of correct answers (>85%), i.e. the distances of the sounds were clearly discernible by the subjects. During the second stage the overall performance decreased by less than 5%. It appeared that for sounds close to each other the differences in reverberation were not large enough to allow for accurate distance estimation when the loudness changes were missing. Apart from these closely spaced pairs, the performance was usually very good. Hence, it was concluded that distance judgements can still be made even when loudness cues have been removed and only the parameters specific to the reverberant environment such as the direct-to-reverberant sound ratio (D/R) are present.

Regarding Chomyszyn's work, it has to be noted that judging the apparent distance of a sound source can be problematic when listening over headphones. Binaural reproduction systems are infamous for their susceptibility to 'in-the-head locatedness' (see Section 2.4). Unfortunately, Chomyszyn provided very few details as to the binaural recording and playback conditions. Even though he did not mention a lack of externalisation on behalf of his subjects, his results are not conclusive and therefore need to be treated with care. At any rate, a psychoacoustic simulation is likely to benefit from including as many cues as possible in order to increase its stability with respect to adverse technical or listening conditions, "with a side-effect of low listening fatigue since the ears and brain are having to do less work to decide what is going on" [Gerzon, 1992a].

While the $1/r$ law proclaims a 6dB increase in DS level for a halving in physical source distance, a different level increment may be required for evoking the corresponding psychological change. Von Békésy [1949] published a localisation curve based on distance judgements from five (blindfolded) listeners that were collected under anechoic conditions using natural speech as the stimulus. As can be seen from Figure 2.4, continuously increasing the distance of the sound source caused listeners to progressively underestimate its location, leading von Békésy to surmise that auditory space is of limited extent.
In [Begault, 1987] a study was presented that investigated the preferred level change for the sensation of half distance. In particular, the amplitude of a sound was varied according to the inverse distance law, i.e. a 6dB increment, and a perceptually based scale, i.e. a 9dB increment (which roughly causes a doubling in loudness). The latter had been chosen based on findings of an earlier study that had suggested a 9dB difference to be generally preferable over 3dB or 12dB increments for perceiving a halving in apparent source distance. A piano tone (which had been closely recorded in a non-anechoic environment) was used as the source material and played back monaurally over headphones at a chosen reference level of 65dBC. Next, the same tone was reproduced at a level that was either 6dB or 9dB higher. Six subjects were asked to indicate if the louder sound seemed half as distant as the quieter one. Results showed that (in the absence of other cues) a 9dB change was more successful at producing a halving of perceived source distance compared to the $1/r$ law, thus reinforcing von Békésy’s results.

As Begault admitted himself, the significance of his results is limited due to the use of earphones and the exclusion of all other cues crucial for distance hearing. Despite the fact that he asked his listeners beforehand if they could externalise the sound (and rejected one subject as a result of this screening process), assessing the apparent proximity of a diotically reproduced, ‘dry’ source is problematic. Due to the complete lack of cues that might enable externalisation (see Section 2.4) it seems that listeners had to rely on their imagination more than anything else when making their judgements, so the validity of Begault’s study is doubtful.

Regardless of the large standard deviations evident in the data and hence the inconclusive findings, it is interesting to note that Begault’s aforementioned preliminary investigation had shown the actual reference level to be the most important factor in determining whether a 6dB or a 9dB difference
produces the desired distance-halving effect. In general, the more the level of the reference sound was raised (above 65dBC), the more listeners preferred a 6dB increase in level to achieve a halving of perceived distance, whereas preference for 9dB became greater when the reference level was lowered again. Based on this finding Begault proposed a scheme for adjusting the level of (anechoic) speech sounds to achieve equal intervals of distance. This is illustrated graphically in Figure 2.5.

Incidentally, a similar trend of non-linear scaling of the DS level also emerged during the simulations of source distance (see Section 5.4.1) and ensemble depth (see Section 5.11.4), for which musical source materials were used.

2.2.4 Direct-to-reverberant sound ratio (D/R)

Nielsen [1993] investigated the influence of several physical factors on the perception of distance, the D/R being one of them. He distinguished between absolute and relative cues, whereby the former can provide information about the distance of a given auditory event even when heard for the first time under certain conditions. Relative cues, on the other hand, can only be used to identify changes in distance. In this context, it was stated that the D/R depends on the acoustical properties of an enclosure. Since such (reflected sound) information is often obtained subconsciously, a listener can usually tell whether a source is proximate or far away even when (s)he has not been exposed to the sound event before, which is why the D/R always provides absolute cues to distance perception.

As part of his work, Nielsen carried out listening tests in three different rooms: an anechoic chamber, an IEC 268-13 standardised listening room and a classroom. Several loudspeakers were positioned at distances ranging from 1m to 5m. Listeners were visually isolated from both the room and test set-up by means of an acoustically invisible curtain. Regarding the test material, an anechoic recording of a human voice was chosen. The influence of the reproduction level was tested by playing back the stimulus at 58Phon, 68Phon and 78Phon, measured at the listening position. 32 subjects were asked to indicate the perceived location of each auditory event graphically by drawing their responses on a computer screen. The obtained results showed that the differences in room acoustics had a significant impact on perceived distance. For the two non-anechoic environments a clear relationship between physical and perceived source distance could be detected, i.e. listeners estimated the source to be at about the same distance as each loudspeaker position, regardless of the playback level. This was not the case for the anechoic chamber where the physical source distance was found to have no influence.
on its perceptual counterpart. Hence, it was concluded that reflections are crucial for distance hearing, whereby "the main factor changing [in a reverberant room] as a function of distance is the ratio between direct and reverberant sounds".

Although Nielsen’s results imply that the D/R has a strong impact on distance estimation, other research has shown reflected sound energy to also define a boundary for distance perception. That is, there seems to be an upper limit to how much reverberation can be mixed with a direct sound before it reaches an “auditory horizon” [Begault, 1994]. What is more, it may be asked whether the salient distance cues are not provided by some finer details of the reverberant energy. Surely, the human hearing system does not perceptually integrate the reflected energy as a whole for evaluation purposes? These or similar questions probably led recent research to question the long prevailing view that humans perceive source distance almost solely on the basis of the D/R. A discussion of research into finer details of reflected sound is presented below.

### 2.2.5 Finer details of reflected sound

To date few psychoacoustic studies have been reported that have endeavoured to identify directly finer details of reverberant sound, which may be influential to perceived source distance. In light of the considerable amount of research that has addressed the impact of reflections onto apparent source width perception (e.g. see [Beranek, 1996]), this is rather surprising. As a matter of fact, a lot that is taken for granted regarding the salience of finer features of reflected sound to perceived distance (and depth) stems from theoretical contemplation or informal experimentation. Notwithstanding, an attempt has been made to summarise the relevant literature. Since the role of early reflections (ERs) to distance hearing seems to be a hotly debated topic at the moment, it is hoped that the following overview will assist with the design of any future studies that can meet scientific standards.

**Relative distribution of early and late sound energy**

Research into distance perception conducted by Michelsen and Rubak [1997] focused on two factors: the temporal distribution of reverberant energy and the finer structure of ERs. It was pointed out that the properties influencing the perception of distance are the level and spectral content of the DS and various properties of the reflected sound. However, as the influence of the DS level diminishes in reverberant environments, reflected sound has to provide the distance cues primarily. Hence, Michelsen and Rubak designed an investigation to address this issue. By keeping the D/R constant and changing the temporal distribution of the reverberant energy, the clarity index $C_{80}$ was manipulated. As to the finer structure of the ERs, they intended to clarify whether room-specific

---

7 The clarity index, expressed in decibels, is the ratio of the early (0ms to 80ms) to the late (80ms to 3s) sound energy [Beranek, 1996]:

$$C_{80} = 10 \log \left\{ \frac{\int_0^{80} p^2(t) \, dt}{\int_{80}^{3000} p^2(t) \, dt} \right\}, \, dB$$

where $p$ = sound pressure and $t$ = time.
structures of the first 10ms of ERs can lead to an improvement in distance evaluation. The reflections were adjusted both in delay and amplitude and directional properties were simulated by means of level differences between the (2-channel stereo) loudspeaker signals. Several different impulse responses were synthesised and the general format of the DS, ERs (t ≤ 10ms) and diffuse reverberation (t > 10ms) was adopted. In order to create differences between the stimuli, at least one of the following changes was applied to each impulse response:

- The discrete ERs were substituted with diffuse reverberation;
- The influence of the overall sound level was eliminated by scaling a particular impulse response for the two chosen distances to have equal total energy;
- $C_{80}$ was in/decreased by applying a weighting function to the reflections and diffuse reverberation.

For the listening test a graphical response method was chosen, i.e. subjects had to indicate the location of each stimulus on an outline of the surroundings. Analysis of the results suggested that a better distinction of distance took place when simulated ERs were present than when they were substituted by diffuse reverberation. This finding led to the (fairly general) conclusion that ERs support the estimation of source proximity. Also, a strong relationship between $C_{80}$ and distance hearing was evident from the data. In general, it was found that an increase in $C_{80}$ caused the auditory events to be perceived closer whereas a decrease resulted in the opposite impression.

On the whole then, this investigation showed that distance perception does not just rely on the well-known D/R, but that it is also dependent on the temporal distribution of reverberant energy. Besides, the finer structure of ERs seems to provide auxiliary clues.

As to Michelsen and Rubak’s created test stimuli, their chosen time limit of 10ms may be criticised as being inappropriate for verifying the perceptual salience of properties of early reflected sound energy to distance hearing. Under natural listening conditions, hardly any reflections will usually arrive at the receiver within this time window unless the source is far away (see next section). Furthermore, other researchers have suggested (1) that the first 50ms relative to the DS are crucial to distance judgements and (2) that ERs provide the main cues to distance estimation (see below). Clearly, the second point somewhat opposes Michelsen and Rubak’s findings. However, as will be seen in Section 5.4.1 and Section 5.5.2, results obtained as part of this work support the outcomes of their study.

Temporal and spatial distribution of early reflections
Begault [1987] conducted a pilot experiment to scrutinise the effect of varying the angle of incidence of a single reflection onto perceived source distance. Two loudspeakers were set up in an anechoic chamber to reproduce the direct and reflected sound. The DS was fixed at 0° azimuth and elevation, while the reflection was varied in terms of its horizontal angle of incidence (30°, 90° and 150°),
2. Correlations between spatial attributes and physical factors

intensity (-6dB, 0dB and +6dB) and time delay (11ms, 21ms and 41ms) relative to the DS. Reflection-free male speech and a piano tone (which had been closely recorded in a non-anechoic environment) were employed as the programme materials. All resultant sound fields were captured using a binaural head system and presented dichotically to three listeners who compared 108 randomised pairs of both the speech and piano stimuli. For each dyad they had to indicate which stimulus sounded closer. Significant results were obtained when the reflected sound had a level of +6dB and a delay time of 21ms or 41ms. In these cases all responses indicated that for the 150° test condition the source appeared closer than when the reflection arrived from 30° or 90°, irrespective of which programme material was used.

Begault’s result suggests that this binaural parameter serves as an auxiliary, albeit weak cue to distance perception. However, questioning of its ecological validity is legitimate. Begault acknowledged that no single reflection can in actuality be louder than the DS. Nonetheless, he claimed that the auditory system groups reflected energy in time intervals small enough for perceptual fusion of their individual levels to occur, which is why a synthesised reflection can be thought of as being representative of the energy of many individual, higher-order reflections. Yet, even if such a scenario is practically conceivable, the result needs to be treated with care due to the exploratory nature of this (pilot) study and the lack of reverberation, which can put externalisation of a (binaural) sound event at stake (see Section 2.4).

Gerzon [1992a] claimed that the most important information for distance estimation is obtained from ERs. This assumption was based on a hypothesis by Peter Craven (presented in [Gerzon, 1992a]) who proposed that the auditory system detects the time delay and amplitude gain of each reflection relative to the DS to deduce apparent source distance. By applying this method to a large number of reflections, a reasonable estimate of distance can be obtained. Since the Craven hypothesis relies on the auditory system to be able to discriminate the ERs individually, it also anticipates that the distance effect will break down when two reflections arrive at substantially the same time. This implies that artificially generated distance cues avoiding overlaps of synthesised reflections may actually work better than simulated natural room cues. However, Craven’s assertion that the perception of distance will become more reliable as the number of ERs is increased (provided that no temporal overlaps occur) contradicts another assumption, namely that additional ERs reduce the human hearing system’s ability to deduce information about sound sources. In this respect, Gerzon claimed that practical work had shown that on the whole the sense of distance was poor if very few ERs were reproduced.

Still, Gerzon also admitted that there are uncertainties about many details of the Craven hypothesis. For instance, it does not seem to be clear how closely spaced in time the ERs can arrive at the ears before they start to overlap psychoacoustically. Likewise, the role of the spatial distribution of ERs for distance judgements is not understood. In general, Gerzon believed that ERs arriving from different directions are desirable since they help reduce subjective colouration and increase the “sense of
spaciousness”. In addition, the ears tend to be better at discriminating different signals if they come from different directions (due to spatial unmasking). Yet, he acknowledged that there is little evidence that this contributes to perceived source proximity. In fact, the good sense of distance conveyed by omnidirectional monophonic recordings (or even telephone conversations) may indicate that the direction of ERs only supports the perception of source distance to a marginal degree. The influence of the consistency of the late reverberation with ERs on range hearing also seems unclear. In this context, Gerzon referred to research by Kendall and Martens [1984] who found that the first 33ms of an acoustical space’s impulse response give rise to a strong distance effect, even if the reverberation tail is entirely omitted (see also Section 4.2). Nonetheless, he disagreed with their chosen time limit of 33ms. In his opinion, it is the reflections arriving within the first 50ms after the DS that supply the predominant cues for distance estimation [Gerzon, 1992c]. Also, Gerzon stated that Kendall and Martens’ finding does not prove that additional cues might not make distance perception more robust, even though he considered the cues defined by the Craven hypothesis to be the most important ones [Gerzon, 1992a]. That is why he recommended maximising the number of available cues in order to create a reliable distance illusion.

While the demand for maximising simulated distance clues seems very sensible, this author does not accept the correctness of the Craven hypothesis. In Section 5.11.1 a psychoacoustic study will be presented that quantified the subjective effects of manipulating several finer features of ER patterns. Incidentally, time-symmetrical sets of reflections were compared to asymmetrical ones and no breakdown in perceived source distance was found. Also, as will be shown in Section 5.4.1 and Section 5.5.2, it is possible to produce changes in source proximity with the help of only a small number of ERs. Regarding the influence of late-arriving reverberant energy, several investigations, including the ones by Nielsen or Chomyszyn, have shown this parameter to have a strong bearing on distance hearing as well.

Griesinger [2000] also addressed the perceptual effects associated with ERs. He argued that an impression of source distance is created by reflections arriving at the listener between 20ms and 50ms after the end of a DS. These reflections lead to fluctuations in the ILD and interaural time difference (ITD) (see also Section 2.5.1), which enable the human hearing system to detect the distance properties of the auditory event. In contrast, fluctuations occurring 150ms or more after the end of a DS are perceived as reverberation and envelopment (see also Section 2.8). Thus, Griesinger concluded that the acoustical properties of a particular room or hall can be divided into two separate perceptions, i.e. source distance and envelopment. Therefore, to create a recording with sources placed at different distances (and hence an overall sense of ensemble depth (see Section 2.3)), decorrelated reflections in the 20ms to 50ms time window should be added. Based on these claims Griesinger suggested the “ideal reverberation profile”, which is shown in Figure 2.6. As far as the spatial distribution of the reflected energy is concerned, Griesinger pointed out that ERs from anywhere on the median plane are inaudible (but he did not specify why this should be the case). Further, reflections in the 20ms to 50ms
time range coming from the same direction as the source are either inaudible or they cause colouration without adding any sense of distance. That is why Griesinger stressed the need for early lateral reflections in order to be able to evoke the subjective effect of source distance.

![Figure 2.6: Schematic illustration of Griesinger's ideal reverberation profile (after [Griesinger, 2000])](image)

In this author's opinion, Griesinger's assertions are the result of an oversimplification of the spatial hearing mechanism. Basically, Griesinger argues that only two spatial attributes (i.e. source distance and envelopment) exist, which numerous other research results clearly refute (e.g. see [Berg & Rumsey, 2001; Koivuniemi & Zacharov, 2001]). What is more, in his earlier work (e.g. [Griesinger, 1997]) he explicitly stated that reflected energy arriving within the first 50ms after the DS gives rise to the perception of source width rather than source distance. In this context, experimental work to be presented in Chapter 5 showed that reflections arriving within this time window influence both width and distance perception. Hence, it is likely that specific features of ER patterns are responsible for the occurrence of each of these two qualitative phenomena. This assertion is substantiated by the availability of ample evidence that source width perception depends on distinct properties of the source material [Mason & Rumsey, 2001], which is why it cannot be generalised that an upper time limit of 50ms is applicable in each case (see also Section 2.5.1).

Pellegrini [2002] proposed a system giving independent control over the apparent distance (and direction) of a sound source as well as the apparent room size. It utilises an image-source model (which is slightly asymmetric to prevent unfavourable interferences and thus colouration) to compute four reflections for simulating source distance and six reflections for the control of room size impression. Regarding their temporal distribution, Pellegrini pointed out that reflections with delay times of 0ms to 20ms influence sound timbre rather than to enhance spatial attributes (unless they arrive from 30° to 65° relative to a frontal sound source) and therefore should be minimised. In contrast, discrete reflections arriving at the receiver 20ms to 50ms later than the DS provide salient cues for distance hearing. Referring to [Griesinger, 2000], Pellegrini further stated that perceived room size predominantly depends on reflected sound energy within the 50ms to 200ms time window.

---

8 This system seems to provide the basis for Studer's 'Virtual Surround Panning', which will be discussed in Section 4.4.
Figure 2.7 displays the connections between the various perceptual effects and respective time windows of an impulse response as described by Pellegrini.

Figure 2.7: Graphical illustration of relationships between timbre, distance and room size perception and the respective impulse response time regions as described in [Pellegrini, 2002]

Assumedly, the choice of the upper time limit was based on the observation that the spatial distribution of a sound field in a large room takes up to 200ms to become diffuse (whereas in a small room it will be diffuse after only 100ms). Likewise, it was noted that:

- The bigger the room, the later the (room size) reflections will arrive and the more spread out in time they will be and vice versa;
- The farther away the source, the closer in time the (distance) reflections will be with respect to each other and the DS;
- The farther away the source, the more frontal the angles of incidence of the (distance) reflections will become and vice versa.

These relationships are illustrated in Figure 2.8, Figure 2.9 and Figure 2.10, respectively.

Figure 2.8: Specular reflections (up to second order) calculated using an image-source model algorithm for a rectangular room with a volume of 1200m³ (left graph) and 3780m³ (right graph). The source-receiver arrangement is constant and slightly asymmetric. The direct sound is at 0ms.
2. Correlations between spatial attributes and physical factors

Figure 2.9: Specular reflections (up to second order) calculated using an image-source model algorithm for a rectangular room with a volume of 1200m³. The receiver is positioned at one end of the room and the source is placed in the centre (left graph) or at the other end of the room (right graph). The source-receiver arrangement is slightly asymmetric. The direct sound is at 0ms.

Figure 2.10: Plan view of the spatial distribution of the early reflections of Figure 2.9 (displayed as image sources). Level is represented by circle size. For each graph the source is located in the centre of the largest circle (the direct sound) and the receiver position is indicated by the black dot.

Depending on the system's distance and room size settings, these temporal and spatial modifications are applied to the synthesised reflection patterns. In terms of level processing, both DS and ERs are adjusted in accordance with changes in source closeness, thereby manipulating the D/R. An increase in room size causes a diffusion filter to spread out the energy and hence fill the gaps between adjacent reflections. The levels of the room size reflections are chosen so as to give an exponential decay that matches a specific reverberation time (RT), which depends on the chosen volume of the simulated environment and which is controlled by means of diffuse reverberant energy. Further, Pellegrini postulated that the distribution of (reflected) sound energy over time and the extent to which the listener's ear signals are uncorrelated are important cues to room size perception, but it is unclear how these cues are exploited in his processing.

To validate his devised system Pellegrini conducted a listening experiment, putting to test the hypothesis that it allows source distance to be controlled independently of room size. Eight stimuli were created illustrating six degrees of source range (ranging from the smallest to the largest achievable distance effect) and four degrees of room volume (133m³ to 741m³). For some reason, all distance-simulating reflections were processed to have maximal angles of incidence of about ±30° relative to the frontal DS (no details were provided as to the spatial distribution of the six other discrete reflections). It is also interesting to note that each stimulus comprised at least two room size reflections with delays of less than 50ms, the smallest simulated room having all six reflections.
occurring below this temporal (distance) threshold. The sounds were presented over a binaural system
with head-tracking, but no individualised head-related transfer functions (HRTFs). Ten listeners had
to estimate the absolute distance (pre-specified to be between 0m to 10m) of a source three times for
each stimulus. A (possibly non-anechoic) acoustic guitar recording was used as the DS. Results
showed that the panel perceived the stimuli in the correct order with respect to the intended distance
effect, irrespective of the variations in room size settings. Additionally, presentation of stimuli that
had identical distance settings but differed in terms of simulated room size produced non-significant
changes in apparent source range.

In this author's opinion, Pellegrini's experiment helped little to validate the proposed room size
algorithm (and perhaps was not intended to do so). At best, it can show that the room size processing
did not diminish the listeners' ability to make (reasonably accurate) distance judgements. Since the
subjects were told to assess source distance changes only, it cannot be inferred from the results that
the room size algorithm evoked the envisaged sensation or, in fact, that either of the two attribute
simulations was free from any additional perceptual effects. Also, assuming that some form of verbal
feedback had shown perceived room size to in/decrease, the algorithm would have still been
considered inappropriate for this work, because of the multidimensional nature of the room size
attribute; after all, a room has a certain width, depth and height. The above summary of Pellegrini's
work also indicates that more knowledge is needed regarding sound field characteristics that
distinguish source distance from room size hearing. There are clear discrepancies between declared
psychoacoustic relationships and applied processing and it seems that (at least some of) the room size
reflections helped control perceived source range.

It is also worth noting that the examination of physical models to identify (finer) cues relevant to
source distance and room size perception formed a major part in Pellegrini's work. However, while
such real-world observations may be valid in their own right, this does not necessarily mean that their
perceptual relevance is warranted, too. Fundamental psychoacoustic research is required to confirm
that this is really the case. It is true that a listening test was conducted to verify the system, but the
multitude of parameters changed to manipulate source proximity prohibits any inferences to be made
as to the perceptual salience of varying the delay times and angles of incidence of the distance
reflections, for example. Controversies such as whether or not early lateral reflections are important to
perceived distance/depth will not be resolved through sole reference to physical facts.

Finally, the above is an example of inadequate experimental design that is not uncommon for studies
aiming to validate a certain signal-processing technique. While such a scenario is preferable to not
conducting any subjective testing at all (as is the case for some of the other research described above),
it is still far from ideal in helping to lay open connections between perceptual and physical
characteristics. From a psychoacoustician's point of view, rigour in the modelling of perceptual
effects should be matched by rigour in the psychological verification of the developed algorithms.
Tracking early reflections

Corey [2002] carried out an experiment to test the supposition that reflections which track the location of a sound source allow subjects to better judge source distance than static ones. The tracking reflections were calculated using a fourth-order, horizontal-only image-source model of a small rectangular room. A custom-built software application was used to render the resulting 40 reflections over a conventional 3/2-stereo layout. Conversely, the static reflections were generated by means of a commercially available 5-channel reverberation (and panning) unit for a room model of the same size. Only the DS and the ERs were included in the synthesised sound fields and the level of the DS was kept constant. Two source locations, directly ahead of the listener, were selected: one near the front wall (A) and the other (B) close to the listening position, which was at the centre of the simulated room. Three types of monophonic, anechoic programme material (female speech, bongos and electric guitar) were used. Four subjects were presented with two sound examples (A and B) containing either static or dynamic reflections and were asked if they could detect a change in distance. If so, they had to choose which of the two sources appeared to be farther away. An acoustically transparent curtain was hung directly in front of the loudspeaker array to avoid auditory-visual interaction. For the stimuli containing tracking reflections 90.7% of the responses indicated that listeners could hear a change in source proximity. Conversely, for the static condition only 5.5% of the answers indicated an audible distance change. Also, with respect to the tracking reflections, 58.4% of the responses implied that listeners perceived A as being more distant. This was claimed to be a statistically significant result, but it was not specified if/how many repetitions were included in the experimental design or how the data were analysed.

The fact that the static test condition did not cause a change in perceived distance does not come as a surprise because it appears that A and B were physically identical in this case. As to the tracking test condition, the two source locations presumably differed in terms of the spatial and temporal distribution of the reverberant sound (see the above discussion of Pellegrini’s work). Consequently, it is expected that the audible differences were small, which is why it would have been helpful to employ a subjective testing methodology allowing reliable and unbiased verification with respect to their detectability. A suitable candidate for this task would have been the well-known ABX paradigm, which was utilised for an investigation into finer features of ER patterns executed as part of this work (see Section 5.11.1). Based on the outcome of the ABX test, a 2-alternative forced-choice test (e.g. see [Levitt, 1971]) could then have helped determine the practical significance of these differences. In this context, Corey’s result of 58.4% does not give the impression that these cues supply perceptually robust source range information. It is imaginable that tracking reflections might give a clearer indication of source location, since they produce changes in interaural comb-filtering patterns that are also present under natural listening conditions. However, the conclusion that tracking reflections are a critical component of a distance simulation should be regarded as a tentative one until their perceptual weight has been assessed in the presence of other prominent distance clues.
2. Correlations between spatial attributes and physical factors

2.2.6 The Doppler effect

Similar to changes in the temporal and spatial structure of ER patterns, the Doppler effect (or shift) occurs whenever there is a dynamic change in source range. Named after the Austrian physicist C. J. Doppler (1803-1853) who first described its cause, the phenomenon implies that as a sound-emitting source approaches a receiver the peaks and troughs of the radiated wavefronts are compressed (i.e. closer together in space), whereas if the source retreats they are rarefacted (i.e. farther apart in space). This results in changes of the effective wavelength (and therefore frequency) at a listener's ears that in turn manifest themselves in the form of pitch variations. Figure 2.11 illustrates this concept graphically.

In [Strauss, 1998] a pilot study was briefly summarised that investigated the perceptual influence of Doppler shifts in virtual auditory displays. A scenario of a car passing the listener at high speed was auralised for binaural reproduction and presented to 21 subjects. Several simulation parameters – one of them being the Doppler effect – were varied and the listeners had to judge the plausibility of the presentations using a paired comparison method. Although the magnitude of the other parameters varied within the test was much larger than the perceptual change caused by in- or excluding the Doppler effect in the simulation, a clear majority of subjects found the stimuli with Doppler shifts to be more convincing than those without them. This suggests that the Doppler effect can be considered a salient cue for perceiving a front-back movement of a source – a point also made by Begault [1994].

2.2.7 Interchannel cross-correlation

Kurozumi and Ohgushi [1983] investigated the relationship between the cross-correlation of 2-channel acoustic signals and sound image quality. For the experiments seven kinds of white noise
with cross-correlation coefficients\(^9\) ranging from +1 to −1 were created. Each stimulus was low-pass (<1kHz), band-pass (1-3kHz) and high-pass (>3kHz) filtered, giving 21 stimuli in total. These were reproduced in an anechoic chamber by means of a conventional 2-channel stereo set-up. For all possible pairs of stimuli within a given filtered set subjects were asked to judge the degree of similarity with respect to sound image quality on a 5-point category scale. The resultant data were subjected to a Multidimensional Scaling (MDS) (see Section 3.3.2) analysis, which revealed that sound image quality comprised “approximately two independent psychological factors”. In order to identify the underlying dimensions, subjects were instructed to compare all pairs in terms of ‘width’, ‘distance’ and ‘elevation’. The obtained results showed that width depended strongly on the absolute value of the cross-correlation coefficient of the noise samples, i.e. 0 produced the widest and |±1| the narrowest image. Distance, on the other hand, depended greatly upon the cross-correlation coefficient itself, i.e. −1 produced the closest and +1 the most distant image. As to the judgements of elevation, Kurozumi and Ohgushi obtained a wide range of responses from their subjects and thus were unable to identify a clear relationship between this perceptual factor and the cross-correlation coefficient. With regard to the effect of the frequency content of the stimuli, it was stated that the perception of sound image width was even perceptible in the HF range, which was not the case for distance perception. Hence, the absolute effect of the cross-correlation coefficient was greater for frequencies below 1kHz compared to frequencies above 3kHz.

The obtained results are questionable because of a flawed experimental layout, i.e. Kurozumi and Ohgushi did not meet certain requirements that are essential for a stable MDS analysis (see Section 5.5.2). In particular, at least nine stimuli would have had to be compared in terms of their overall similarity for each filtered set to allow unravelling of two meaningful perceptual dimensions. Also, the use of a 5-point category scale definitely led to strongly quantised responses, which can explain the neat stimulus configurations evident in the presented MDS ‘stimulus spaces’ (e.g. see Figure 5.7). Finally, asking subjects to choose one of three answers predetermined by the experimenters did not enable an unbiased identification of the perceptual attributes. While such an approach would have been suitable to confirm findings of an earlier elicitation stage, it was inappropriate for this study, which was essentially exploratory in nature. On the whole then, it seems that this study was bound to give the answers that Kurozumi and Ohgushi asked for.

\(^9\)The cross-correlation coefficient is mathematically defined as follows:

\[
\text{Cross-correlation coefficient } (\tau) = \left[ \frac{\int_{t_1}^{t_2} x(t) y(t + \tau) dt}{\left( \int_{t_1}^{t_2} x^2(t) dt \right)^{1/2} \left( \int_{t_1}^{t_2} y^2(t) dt \right)^{1/2}} \right]
\]

where \(x\) and \(y\) are the two signals whose correlation is to be measured, \(t\) is time, \(t_1\) and \(t_2\) define the period over which the correlation is measured, and \(\tau\) is an offset between the two signals under measurement [Mason, 2002a]. It varies between +1 and −1, whereby a high value indicates that two signals are almost identical and a low value the opposite.
Building on the work by Kurozumi and Ohgushi, Kendall [1995] intended to clarify which acoustic factors affect image\textsuperscript{10} distance and which image width. In this respect, he claimed that width is associated with the randomisation of interchannel phase relationships, whereas distance is related to a shift from a correlation measure of $+1$ to $-1$. However, no objective or subjective measurements were included to support this statement. Kendall also emphasised that the technique of controlling image distance (and width) by continuously changing the correlation measure between two signals is strongly dependent on the listening position. This ‘sweet spot’ dependency is similar to the one typical of transaural reproduction systems and hence a clear disadvantage for critical listening exercises, since relatively small head movements can introduce unwanted perceptible changes that might have a considerable effect on subjects’ judgements.

Regardless of the scientific validity and reliability of the findings outlined in this section, it is apparent that the manipulation of the cross-correlation coefficient of two sound signals does not allow independent variation of either of the two affected perceptual phenomena. Hence, this method seems to be unusable for this work. In fact, as will be shown in the following section, this type of processing appears to influence the perception of a third attribute, too.

2.2.8 Low-frequency decorrelation

Martens [1999] also expanded upon Kurozumi and Ohgushi’s findings related to the cross-correlation coefficient and the perception of image distance and width. In particular, his research was concerned with decorrelated LF reproduction and its impact on auditory spatial imagery. A comparison was made between two reproduction modes whereby the first one (Mode 1) employed two woofers, which were both fed with the sum of the left- and right-channel signals below their crossover frequency. Hence, as the two signals were identical this approach excluded the decorrelation contained in the lowest frequency band ($<250\text{Hz}$). The second reproduction set-up (Mode 2) also utilised two woofers, but in this case the left- and right-channel signals were kept separate to enable faithful reproduction of LF decorrelation. With regard to the physical test layout, two studio monitors and woofers were set up at $\pm30^\circ$ in an anechoic chamber. Two experiments were carried out. As part of the first one, listeners had to make forced-choice judgements as to which of two given stimuli dominated with respect to one of the two spatial dimensions under investigation (image width and distance). The second experiment was based on pairwise comparisons of a set of ten stimuli, which differed in terms of their cross-correlation coefficients (five different values) and/or the reproduction mode (the two modes described above). In this case subjects had to judge the overall similarity between each pair of stimuli on a 5-point scale, and the obtained data were analysed using (non-metric) MDS. Listeners were also asked to

\textsuperscript{10} The reader should note that throughout this thesis the term ‘image’ will be used as a prefix for a spatial attribute if a particular statement cannot be generally made with respect to a certain perceptual change. To illustrate, a processing method might affect width perception. Yet, depending on the characteristics of the processed programme material this might either be source width in the case of an anechoic recording of a single instrument and/or environment width if the recording contains acoustic information about the recording space. Therefore, ‘image’ will be applied if a categorical distinction is considered impossible.
make drawings on a standardised form to indicate the spatial extent, shape and location of the auditory spatial image associated with a certain stimulus.

Results of the first experiment confirmed Kurozumi and Ohgushi’s findings, i.e. out of the chosen cross-correlation values of -0.9, -0.7 and -0.3 that had been introduced into the synthesised stimuli the last produced the widest spatial image. Also, the sound images created by Mode 2 were generally perceived to be wider than those of Mode 1. In addition, the reproduction of LF decorrelation (Mode 2) caused the occurrence of two other spatial dimensions: image depth and distance. For instance, the lowest cross-correlation coefficient (-0.9) resulted in a narrow, very close sound image surrounding the listener, while a value of -0.7 created a wider auditory event in-between the listener and the loudspeakers. In the case of Mode 2, an increase in the cross-correlation coefficient produced a sound image with more depth whilst at the same time moving it towards the plane of the loudspeakers, hence making it more distant. For the sound images produced by Mode 1 both source depth and distance seemed to be constant, i.e. the image remained in the plane of the loudspeakers regardless of the cross-correlation coefficient. Thus, these results suggest that the width dimension is more dependent on the cross-correlation coefficient, whereas depth and distance are influenced by reproduced LF decorrelation. A summary of the panel’s drawings is depicted in Figure 2.12. Yet, Martens stressed that not all listeners perceived the stimuli in the same way. Indeed, it seems to be common to observe a large range of individual differences in tests on spatial images associated with variations in the cross-correlation between two sound signals. Analysing the data of the second experiment confirmed the results of the first one, except that the results also indicated a perceived increase in distance for Mode 1 as the cross-correlation coefficient was shifted from -0.9 to -0.1.

Figure 2.12: Summarised listener responses from test on impact of LF decorrelation on perceived spatial imagery. The stimuli’s cross-correlation coefficients are indicated for each auditory event (adapted from [Martens, 1999]).
To summarise, the experiments showed that the use of two woofers for delivering decorrelated LF signals enables extended control over three perceptual attributes of the resulting spatial image, but most particularly image width and distance. These changes in spatial imagery are absent when LF decorrelation is not reproduced faithfully.

Martens' results provide further evidence for the unsuitability of manipulating the cross-correlation coefficient in order to achieve unidimensional variation of one of the associated spatial attributes. Although Martens managed to shed some light on the relationships between physical factors that are responsible for the occurrence of image width and those that result in a distance perception, his experiments uncovered another subjective quality (image depth) that seems to be caused by this type of processing, too. To verify the existence of this third attribute it would have been useful (1) to present the measures of fit obtained from the MDS analyses to be able to deduce whether a third meaningful dimension might exist and (2) to carry out an MDS experiment allowing for the extraction of three dimensions from the subjective data complemented by some verbal reporting to aid the interpretation of the obtained MDS solution (see also Section 5.5). What is more, Martens could confirm previous findings related to varying the cross-correlation coefficient of two signals, which had shown that such methods give rise to highly individual listener perceptions. Evidently, this prohibits the creation of unequivocal training stimuli.

2.3 Ensemble depth

In Section 2.2.8 the relationship between the spatial attribute of image depth and LF decorrelation was delineated. In this section some findings with regard to the simulation of ensemble depth are discussed. However, it has to be emphasised that depth-related work is often difficult to discern from research dedicated to the distance dimension, since researchers have only just started to differentiate deliberately between these two spatial concepts. As a consequence, many of the results reported in the literature are ambiguous due to the two terms being used interchangeably – sometimes even in conjunction with 'sense of perspective', which seems to be employed to describe a similar perceptual dimension. The distinction between depth and distance appears to be important as the results from elicitation experiments like the ones mentioned in Chapter 1 indicate. In summarising the following work, this author has made the assumption that the spatial attribute of ensemble depth was the subject of the research and not the depth of the overall sound scene.

Wohr et al. [1990] discussed the issue of how to simulate depth. In this regard, it was mentioned that to create depth artificially, not only the DS but also suitable reflections need to be reproduced. If such spatial information is missing from the acoustic signals that are recorded these components should be generated artificially, so that the auditory system can recognise both the position and “nature” of a sound when it is played back.
2. Correlations between spatial attributes and physical factors

Woehr et al. also addressed the implications of this claim with respect to suitable microphone and mixing techniques. Spot microphones are often used in conjunction with a suitable main microphone, e.g. during the recording of extended sound sources like an orchestra. When these signals are mixed together care has to be taken so as not to destroy the reflection pattern of the recording room, which is contained in the main microphone signal. It is common practice to lessen the 'space-disturbing' effect of spot microphone signals by adding reverberation and/or compensating for the delay of the main microphone signal. Yet, these techniques are not satisfactory because delay compensation leads to notching effects. That is why Woehr et al. proposed the "room-related balancing technique", which avoids such notching errors and at the same time achieves a desirable increase in level by adding sound energy from artificially generated reflections. According to this method, the spot microphone signals should be delayed so as to fall into the region of the ERs where they are less damaging to the overall effect. As a result, the favourable imaging characteristics of a main microphone with respect to spatial impression are scarcely altered by this type of processing. Figure 2.13 illustrates this concept graphically for the (simplified) case of one spot and main microphone.

In practice, Woehr et al. found that depth perception could be significantly improved by using the room-related balancing technique rather than simple intensity balancing or delayed intensity balancing and that the loss of depth could be avoided satisfactorily by adding just two reflections. Theile [2001] expanded on this point by stating that reflection patterns containing approximately 10 to 20 ERs are advantageous to avoid colouration effects that arise due to comb-filtering. More precisely, the added reflections should be mainly reproduced from lateral directions and arrive within the time window of 15ms to 50ms in order to minimise the audibility of spectral changes and to achieve "a realistic perception of distance and spatial depth".

There are many similarities between this and Griesinger [2000] and Pellegrini’s [2002] work on distance perception. In each case, the importance of ERs arriving from non-medial directions and
having time delays of up to 50ms relative to the DS is stressed for achieving a front-back perspective in a sound scene. This is probably because the spatial attributes of source distance and ensemble depth complement each other, as several sources that are perceived to be at different distances from the listener give rise to ensemble depth.

In consideration of the various findings related to the distance and depth attributes summarised above, one might conjecture that the finer structure of ERs is not crucial for evaluating the distance of single sources, but that ER patterns contain essential information for perceiving ensemble depth. However, it has to be repeated that many of the relevant publications did not include any subjective data that could back up such a hypothesis. In fact, in Chapter 5 results will be presented suggesting that this physical distinction may not be responsible for separating the two constructs psychologically.

2.4 Externalisation

Blauert [1997] discussed a phenomenon known as intracranial or in-the-head locatedness (IHL). IHL is a well-documented phenomenon of (deficient) binaural sound reproduction and is affected by various factors such as the absence of reverberation. As decorrelation is an essential component of diffuse sound fields and hence reverberation, it appears that it influences the perceptual dimension of IHL or rather its antipode externalisation. Indeed, Kendall [1995] proposed that adding decorrelated reverberation to a stimulus can help externalise correlated signals.

The optimisation of binaural sound systems is still the subject of considerable research as headphone-based spatial audio displays can (in principle) offer several advantages over loudspeaker-based alternatives. Other researchers have looked at the benefits of including head-tracking and individualised HRTFs in the rendering algorithms. While it seems that the former is a guarantor for reduced front/back reversals, neither of the two appears to be of any value as far as externalisation is concerned [Begault et al., 2000].

Blauert [1997] stressed that certain loudspeaker-based replay systems are susceptible to IHL, too. For instance, if several loudspeakers are arranged symmetrically on either side of the median plane and radiate identical signals or if two loudspeakers radiate identical signals that are out of phase with each other, IHL is likely to occur. Griesinger [1998a; 1998b] went further by claiming that this sonic property is unique to small rooms. According to him, LF instruments in popular music (reproduced in listening rooms with the help of loudspeakers) are almost always perceived as coming from inside the head. This is not the case for music having a substantially random phase relationship between two (stereo) channels. Also, for performance spaces where LF decorrelation is present the sound is likely to be perceived as external and enveloping. Hence, the perception of externalisation somewhat depends on the properties of the listening environment, whereby rooms with a high degree of symmetry and carefully controlled RT at low frequencies are especially susceptible to IHL.
The above findings are only indirectly relevant to this work. Due to the spatial distortions typically encountered with (non-optimal) headphone reproduction (i.e. IHL and front/back reversals), binaural rendering is deemed inappropriate for conveying well-defined spatial dimensions to a group of listeners in an unambiguous manner. Thus, only the problem of IHL experienced with loudspeaker playback is of concern. Nevertheless, even though all practical work was conducted in critical listening environments that are acoustically well treated and therefore have a low RT across the audible frequency band, subjective measurements can confirm that this qualitative phenomenon was not perceived during any of the various listening tests reported in Chapter 5.

2.5 Source width

Awareness of the importance of apparent or auditory source width (ASW) perception emanated in the area of concert hall acoustics where the fundamental psychoacoustic principles behind this phenomenon were established. In particular, perceived source width was shown to depend on the amount of laterally reflected sound energy arriving at the listening position within 80ms of the DS, which in turn affects the degree of interaural cross-correlation (e.g. see [Beranek, 1996] for an overview of the subject). Besides, the reflections' frequency content was also found to have an impact on ASW hearing (e.g. see [Blauert, 1987][1]).

The notion of ASW implies that a sound source appears to occupy more space than its physical dimensions suggest. However, as indicated by reproduced sound research, it is often difficult for subjects to gauge the left and right boundaries of such a source, because the ERs appear to make it sound “fuzzy” and difficult to localise [Rumsey, 2002]. That is why listeners sometimes find terms like ‘source focus’ or ‘diffuseness’ more applicable. Below, work related to sound reproduction is discussed in more detail.

2.5.1 Objective measures and applicability to reproduced sound

Intending to capitalise on the psychoacoustic (concert hall) findings, researchers started to develop objective measures for predicting ASW perception. In this context, Theile [1987] noted the tendency that these measures either depend on the interaural degree of coherence or the ratio of lateral to frontal or total sound energy arriving at the listener. However, he also emphasised that while the basic physical parameters related to “spaciousness”[12] are known and used to evaluate and optimise room acoustical quality, their suitability for assessing reproduced sound is not clear yet. That is why the applicability of these measures to the field of sound reproduction has been looked at more closely recently. For instance, Mason [2002a] carried out an investigation into objective measures that relate to auditory spatial perception. Following a comprehensive summary of the associated literature, he

---

[1] Blauert used the term ‘auditory spaciousness’ to describe the “characteristic spatial spreading of auditory events, so that they fill a larger amount of space than is defined by the visual contours of an ensemble of sound sources” (i.e. the extents of the auditory events are more spacious than under free-field conditions). This seems to refer to the perception of ASW.

[2] Similar to Blauert, Theile used the term ‘spaciousness’ to describe sources that appear to “fill larger amounts of spaces than in a free field under comparable conditions”.

deduced that there are limitations to these measurements because they are not as successful when applied to the assessment of sound-reproducing equipment. In particular, he argued that none of the measures accurately matches a particular subjective attribute. Hence, Mason concluded that the physical parameters, which are responsible for certain subjective effects, need to be investigated further. This led him to examine the perceptual effects of ITD fluctuations, which are changes over time of the relative phase between the two audio signals measured at the ears. Such fluctuations arise due to the interaction of the DS with at least one reflection. For illustration purposes, Figure 2.14a shows a plot of the modelled ear signals of a binaural receiver with an omnidirectional source (emitting three continuous frequencies of 480Hz, 500Hz and 520Hz) 15m in front and a single reflection from a wall 5m to the right of the receiver. As can be seen, the interaural time (and amplitude) difference changes over time, thereby causing an alternate leading and lagging in the phase of the two audio signals (see Figure 2.14b).

Figure 2.14: Graphical illustration of ITD fluctuations as modelled for an omnidirectional source 15m in front and a single reflection coming from a wall 5m to the right of a binaural receiver (reproduced from [Mason & Rumsey, 2001] with permission). The DS signal is composed of three continuous frequencies of 480Hz, 500Hz and 520Hz. Figure 2.14a shows the computed ear signals, while Figure 2.14b displays the ITD fluctuation over time.

ITD fluctuations are dependent on a multitude of factors related to both the source signal (e.g. amplitude and spectral content) and the reflection pattern (e.g. delays, directions and frequency content) [Mason, 2002a]. Besides, changes in the source signal can cause the properties of the ITD fluctuations to change significantly over a short period of time. Also, for a pair of symmetrical reflections no ITD fluctuations are created due to the interactions being identical at each ear. This may have important consequences when creating artificial reflection patterns to vary certain perceptual phenomena unidimensionally. In this respect, Mason found that, depending on the source signal, the ITD fluctuations in the sound source segment that comprises the DS have different audible characteristics compared to the purely reverberant sound source segment (see below). If the relative interaural phase of an audio signal fluctuates slowly, the subjective effect will be a change in the perceived position of a sound source. If the fluctuations occur at a frequency above a few Hertz, the
2. Correlations between spatial attributes and physical factors

auditory hearing system is not capable of tracking the movement due to the perceptual effect of "localisation lag" [Blauert, 1972]. In this case the source is perceived to be stationary in the presence of a 'surround'. To investigate the perceptual effects due to the localisation lag in greater detail, Mason conducted a number of subjective experiments. From these be found that an increase in the magnitude of the ITD fluctuations caused an increase in the perceived width of a certain attribute of a scene, depending on the nature of the signal. For sound segments with an audible DS, increasing the magnitude of the fluctuations appeared to result in a perceived source widening. For purely reverberant sound segments, increasing the magnitude of the fluctuations seemed to cause a perceived environment widening.

From the above it seems that researchers have used different terms to describe a supposedly similar or identical phenomenon. However, it is not clear whether all the terms that have been employed signify the same perception, because the meanings of these verbal descriptors have not been compared very closely. Besides, there are indications that complete separation of psychoacoustic relationships has yet to be accomplished. For instance, recalling that distance/depth perception is claimed to depend on non-medial reflections with time delays of less than 50ms, perceptual 'overlaps' between the width and distance/depth attributes are to be expected. Likewise, Mason’s results imply that it may not be possible to exert precise control over the width of a source through the synthesis of early laterally reflected energy, as environment width is likely to be changed, too. The controlled insertion of ITD fluctuations into selected source material for the creation of reference stimuli does not appear to be a suitable approach either. Such a method will result in a highly unnatural sound quality if musical programme material is used [Mason, 2002b]. This is because frequency modulation techniques, which would have to be applied, are known to cause unpleasant audible side-effects (e.g. see [Klipsch, 1968]). It is true that by using suitable microphone arrangements, which produce different degrees of interchannel cross-correlation (e.g. coincident vs. spaced microphone set-ups), sound images having different widths could be recorded. However, such a strategy would affect numerous other aspects of the overall sound quality as well – changes in timbre and localisation accuracy being two obvious candidates. Hence, finding an appropriate methodology for simulating source width in a unidimensional manner appears to be a difficult task.

2.6 Ensemble width

Regarding the spatial attribute of ensemble width, this author is not aware of any documented work that has dealt (directly) with this aspect of spatial quality. Under natural listening conditions, interaural differences in the acoustic signal radiated by each component source of a sound scene enable the human auditory system to determine the lateral positioning of each of these sources (e.g. see [Blauert, 1997]). As far as reproduced sound is concerned, ensemble width perception is the result of source-specific interchannel amplitude and/or time differences, which can be controlled using panoramic potentiometers (panpots), Mid-Side (M-S) processing or different microphone techniques, for example [Rumsey, 2002]. As such, ensemble width has not been of interest to psychoacousticians
so far. Nonetheless, in Section 5.9.1 a study will be presented showing it to be perceptually more complex than anticipated, thus justifying its further investigation.

2.7 Environment width
In Section 2.5.1 it was already mentioned that ITD fluctuations affect the perception of environment width, provided that they take place during purely reverberant sound segments. Therefore, these results will not be reiterated here. For an in-depth treatment of the subject the reader is referred to [Mason, 2002a].

2.8 Listener envelopment (LEV)
Like ASW, the concept of listener envelopment (LEV) stems from findings made by concert hall acousticians who were concerned with unravelling the individual components of spatial quality. It is the other constituent commonly thought to be relevant to concert hall preference. Related research by Bradley and Soulodre [1995] showed that the sense of LEV is due to reflected energy arriving at least 80ms after the DS. Further, LEV was found to depend on the level of the reverberant sound energy. Yet, it was emphasised that reflections have to arrive from lateral directions in order to produce a strong sense of LEV. This implies that for a maximal perception of LEV the correlation of the late parts of a listener's ear signals needs to be minimised. Interestingly, Bradley et al. [2000] also found that an increase in early sound energy tends to decrease perceived LEV. Therefore, it was concluded that early and late sound energy influence the perception of LEV concurrently, which is why it is theoretically even possible to have too many ERs.

Very recently, LEV hearing was scrutinised in the context of multichannel sound reproduction, leading to some new insights with respect to salient sound field characteristics. The associated work is presented below.

2.8.1 Level and spatial distribution of late reflected sound
Soulodre et al. [2003] conducted two experiments to investigate the perception of LEV in relation to multichannel audio systems. In the first experiment three parameters were varied systematically: RT, $C_{60}$ and the angular distribution of the late sound. Holding the DS and four ERs ($t < 80ms$) of each synthesised sound field constant, three degrees of $C_{60}$ (7dB, 4dB, −2dB) were obtained by altering the level of the late energy only. RT was set to one of three values (0.5s, 1.2s or 1.9s) and the angular distribution parameter was controlled by having the reverberant energy come from either one (0°), three (0°, ±30°) or all five (0°, ±30°, ±110°) loudspeakers. For the DS a short extract of an anechoic orchestral recording was selected, which (presumably) was fed to C in each case. The ERs were routed to C, L and R. A modified version of the MUSHRA methodology [ITU-R BS.1534, 2001] was employed as the test protocol, allowing subjects to instantly compare and rate several sound fields in terms of perceived LEV relative to a reference sound. Analysis of the results showed that (for an ITU-
compliant listening environment) perceived LEV can be controlled by either varying the relative level or the angular distribution of the late energy and that it is also affected, albeit to a lesser extent, by altering the RT of a sound field.

The second experiment was almost identical to the first one except that this time RT was kept constant (1.9s), whilst three different playback levels (74dBA, 77dBA and 80dBA) were included instead. Results confirmed the previous finding that it is possible to vary perceived LEV by either changing the relative level or the angular distribution of the late energy. In addition, it was found that the overall playback level influences the perception of LEV as well, whereby a higher level causes perceived LEV to increase and vice versa.

2.8.2 Temporal distribution of late reflected sound
Acknowledging that the choice of 80ms as the transition point between early and late energy may bias the outcome of such studies, Soulodre et al. [2003] executed another formal subjective test to scrutinise directly the lower temporal limit salient to LEV perception. In order to eliminate the fixed temporal distribution previously strictly adhered to by LEV researchers, discrete reflections having delays of more than 80ms were included this time and their number, levels, delays and directions varied between sound fields. Similarly, the level as well as the temporal and spatial distribution of the diffuse reverberant energy was broadly altered among the test items. For some sound fields the diffuse energy started immediately after the DS, whereas for others the onset of the reverberation was delayed by 20ms to 120ms. RT was also varied across stimuli and the generated sound fields were reproduced over a 3/2-stereo loudspeaker set-up again. Subjects rated a total of 12 test items in terms of their degree of LEV relative to a reference using the multi-stimulus comparison method employed previously. Data analysis showed that the sound fields were uniformly distributed in terms of perceived LEV and that subjects could consistently discriminate many levels of LEV. In addition, impulse responses were measured for each sound field using a sideways-facing bidirectional microphone that was set up at the listening position. To be able to examine the effect of the early-late transition point, the obtained measurements were processed to compute \( LG^w_90 \). This is a slightly modified version of the \( LG^w_90 \) measure\(^\text{13}\), which allows quantifying the level of the late sound arriving from lateral angles and which had previously been found to be the most suitable predictor of LEV. \( LG^w_x \) was calculated for \( x = 5-200\text{ms} \) and correlated against the corresponding mean LEV score of each sound field. Incidentally, the highest correlation was obtained for a lower integration limit of 105ms, suggesting that 80ms is indeed not the best choice for predicting perceived LEV. In order to

\(^{13}\) The \( LG^w_90 \) measure (as used in this reproduced sound context) is mathematically defined as follows:

\[
LG^w_90 = 10 \log \left( \int_{80}^{x} p^2_r(t) dt \right) \text{ dB}
\]

where \( p_r(t) \) is the instantaneous lateral sound pressure measured by the bidirectional microphone [Soulodre et al., 2003].
verify this finding, the sound fields and data from the first two experiments (see Section 2.8.1) were analysed in the same way. This time it was found that although a value of 105ms gave a higher correlation when averaged over all three experiments, it did not provide the highest correlation for each individual experiment. Hence, Soulodre et al. concluded that the optimal value may be frequency-dependent and set out to derive perceptually motivated integration limits taking into account the forward masking properties of the human auditory system. A comprehensive search for the optimum early-late energy transition point as a function of frequency resulted in values ranging from 160ms to 45ms for the octave bands with centre frequencies of 125Hz to 8kHz, thus corroborating that there is more forward masking at low compared to high frequencies. Based on these findings, a new objective measure was proposed that turned out to be the best predictor of LEV for all three subjective experiments.

The results published by Soulodre and his colleagues are potentially very useful for various applications in spatial audio, but it remains to be seen whether equally high correlations between perceptual and physical LEV measurements are achievable for other types of source material, too. In this respect, it would be helpful not to play back an orchestral recording monophonically, as this would increase the ecological validity of future findings. Also, similar to the criticism expressed with regard to Pellegrini’s validation experiment (see Section 2.2.5), it is not guaranteed that the identified physical characteristics affect a single spatial dimension that can be meaningfully described by the term LEV and hence that the proposed metric actually gauges a unidimensional percept. Soulodre et al. instructed their listeners that, while many subjective parameters of the sound fields may be varying, they were to rate only perceived LEV for each sound field. However, allowing subjects to express their perception of a single (pre-specified) attribute only is problematic, because it means that the existence of other qualitative changes happening simultaneously cannot be ruled out. These in turn could have interacted with the subjective effect of interest and hence might have influenced the listeners’ LEV judgements\footnote{While a subject may be asked to assess a single attribute of a stimulus, it is very likely that the resultant judgement will be affected by other perceptual changes if no special care has been taken to exclude them from the item under test. This is because different sensory messages interact, leading to a confounding of the response and hence to a potential misinterpretation of the results [Stone & Sidel, 1993].}. Additional verbal feedback could have alleviated such uncertainties and thus helped to strengthen Soulodre et al.’s findings.

The perception of envelopment in reproduced sound was also addressed by Griesinger [2000]. He stated that for an optimal sensation of LEV decorrelated, reverberant energy with a delay of at least 150ms relative to the DS should be supplied. Hence, his chosen lower integration limit for LEV perception is in broad agreement with the recent discoveries made by Soulodre and his colleagues.
In the context of source width perception, it was already noted that due to the lack of internationally agreed terminology research findings are not directly comparable as a potential for misinterpretation exists. In the case of LEV the situation is even more complicated, because the underlying concept is more abstract in nature compared to the relatively tangible notion of the width of a source. Of course, it is possible that there are significant differences in the meanings of the verbal descriptors used by different people. To illustrate, Barron and Marshall [1981] described “spatial impression” in terms of “the difference between feeling ‘inside’ the music and looking ‘at’ it, as through a window”. However, this description may also apply to LEV, which Griesinger [1999] defined as “the perception of being enveloped by the music”. Similarly, Beranek [1996] described LEV as “the subjective impression of being enveloped by reverberant sound in a hall”. In contrast, in [Bradley & Soulodre, 1995] a definition of LEV was given depicting it as “the fullness of sound images around a listener”, which is semantically different again. While all four descriptions imply that the listener is surrounded by acoustic information, only Beranek specifically stated that it is reverberation, which envelops the listener. In the other cases it could also be the DS radiated by a single very wide source or a group of sources distributed around a listener that leads to an enveloping impression.

To avoid this and similar problems related to semantic differences and misconceptions, Rumsey [2002] proposed the aforementioned hierarchical system of spatial attributes (see Section 1.3). With respect to the term LEV, he suggested to differentiate between source, ensemble and environmental envelopment. Applying these concepts to the research presented in this section, the spatial attribute of LEV would be described more accurately as environmental envelopment.

### 2.9 Summary

This chapter reviewed established relationships between spatial dimensions and their physical correlates so as to pave the way for the unidimensional attribute simulations to be undertaken in Chapter 5. To date, numerous attributes have been examined and for each of them at least one perceptually salient acoustical parameter has been identified. Bearing in mind that an ITU-conformant loudspeaker set-up is to be used for this project, not all of these have to, or indeed can, be taken into account. The relevant relationships are summarised in Table 2.1. It should be noted that neither is this list of attributes complete, nor is it claimed that the shown physical parameters are the only ones that influence the occurrence of a particular spatial effect. These shortcomings are due to the fact that many of the underlying psychoacoustic connections are not completely understood or possibly have not even been discovered yet. Elucidating this situation is not straightforward because reflected sound constitutes the basis for most of these dimensions, which is why the chance for perceptual interactions is high. Substantial amounts of work will have to be carried out before it will be known which physical parameters of sound control which attributes. Notwithstanding, the research covered in this chapter represents the current standard of knowledge in this field. Thus, the pertinent findings are expected to prove useful with respect to varying spatial attributes in a unidimensional manner.
Table 2.1: Reviewed spatial attributes, their physical correlates (relevant to this work) and respective references

<table>
<thead>
<tr>
<th>Spatial attribute</th>
<th>Physical parameter</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-frequency content</td>
<td>[Blauert, 1997]</td>
<td></td>
</tr>
<tr>
<td>Loudness</td>
<td>[Blauert, 1997]</td>
<td></td>
</tr>
<tr>
<td>D/R</td>
<td>[Nielsen, 1993]</td>
<td></td>
</tr>
<tr>
<td>Relative distribution</td>
<td>[Michelsen &amp; Rubak, 1997]</td>
<td></td>
</tr>
<tr>
<td>Early (t &lt; 50ms)</td>
<td>[Gerzon, 1992a]</td>
<td>[Griesinger, 2000]</td>
</tr>
<tr>
<td>Spatial and temporal</td>
<td>[Pellegrini, 2002]</td>
<td></td>
</tr>
<tr>
<td>Cross-correlation</td>
<td>[Martens, 1999]</td>
<td></td>
</tr>
<tr>
<td>Doppler effect</td>
<td>[Strauss, 1996]</td>
<td></td>
</tr>
<tr>
<td>Cross-correlation</td>
<td>[Martens, 1999]</td>
<td></td>
</tr>
<tr>
<td>Early (t &lt; 50ms)</td>
<td>[Wöhr et al., 1990]</td>
<td>[Theile, 2001]</td>
</tr>
<tr>
<td>Decorrelated reverberant energy</td>
<td>[Kendall, 1995]</td>
<td>[Begault et al., 2000]</td>
</tr>
<tr>
<td>Absolute value of</td>
<td>[Martens, 1999]</td>
<td></td>
</tr>
<tr>
<td>ITD fluctuations during</td>
<td>[Mason, 2002a]</td>
<td></td>
</tr>
<tr>
<td>DS (as caused by asymmetrical, lateral reflections)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source-specific</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>ITD fluctuations during</td>
<td>[Mason, 2002a]</td>
<td></td>
</tr>
<tr>
<td>reverberant sound</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Direct sound of single</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Direct sound of sources distributed around listener</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Lateral reflections</td>
<td>[Souloire et al., 2003]</td>
<td></td>
</tr>
<tr>
<td>(125Hz octave band)</td>
<td>(as caused by asymmetrical, lateral reflections)</td>
<td></td>
</tr>
<tr>
<td>(8kHz octave band)</td>
<td>(as caused by asymmetrical, lateral reflections)</td>
<td></td>
</tr>
<tr>
<td>(as caused by</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>ITD fluctuations during</td>
<td>[Mason, 2002a]</td>
<td></td>
</tr>
<tr>
<td>reverberant sound</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Direct sound of single</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Direct sound of</td>
<td>[Rumsey, 2002]</td>
<td></td>
</tr>
<tr>
<td>Lateral reflections</td>
<td>[Souloire et al., 2003]</td>
<td></td>
</tr>
<tr>
<td>(125Hz octave band)</td>
<td>(as caused by asymmetrical, lateral reflections)</td>
<td></td>
</tr>
<tr>
<td>(8kHz octave band)</td>
<td>(as caused by asymmetrical, lateral reflections)</td>
<td></td>
</tr>
</tbody>
</table>

15 To be precise, individual source or component distances are also required for the perception of ensemble depth.
3 METHODS FOR STUDYING THE SUBJECTIVE EFFECTS OF SOUND-PROCESSING ALGORITHMS

This chapter will present sensory evaluation approaches that allow identification of the perceivable attributes of a set of (reproduced sound) stimuli. The need to verify the unidimensionality of the attribute simulations to be created means that a technique is required that permits unbiased elicitation of all perceivable constructs. Subjects should therefore be able to use their own language to describe what they perceive without being influenced in their assessment or verbalisation. Established procedures which may qualify for this purpose can be roughly divided into three categories [Rumsey, 1998]:

1. Those that aim to arrive at a common set of attributes for grading by all panel members;
2. Those that are based on free categorisation or individualised scales;
3. Those that use multivariate analysis techniques carried out on some form of non-verbal similarity judgements.

For each of these categories in turn, the most commonly used techniques will be presented and their associated strengths and weaknesses assessed with respect to the particular aim of confirming the unidimensional variation of the simulations to be generated.

3.1 Elicitation of common attribute scales

Techniques that derive attribute scales for a set of stimuli based on the verbalised perceptions of a group of subjects usually come under the heading of descriptive analysis (DA). According to Bech [1999], the basic steps that most forms of DA entail are: selection of subjects, development of a descriptive language, subject training, reporting evaluations in quantitative form, and analysis and presentation of the obtained data. The best known example of this type of sensory evaluation philosophy is the Quantitative Descriptive Analysis® (QDA) method, which is described below.

3.1.1 Quantitative Descriptive Analysis (QDA)

QDA was designed to overcome problems intrinsic to other DA approaches, e.g. the sole reliance on qualitative information or the use of attribute scales pre-defined by the experimenter (so-called 'provided constructs') [Stone & Sidel, 1993]. Typically, QDA is applied as a means of setting up an evaluation panel to be employed for industrial applications such as product development or quality control. It requires subjects, who have been screened and chosen for their discriminatory abilities, to come up with a list of attributes based upon their collective assessment of a set of stimuli. After a (usually extensive) process of discussion, further auditioning and attribute modification, carried out under the guidance of a panel leader, a common language and a set of rating scales are arrived at. The development of such scales and associated definitions forms an essential part of QDA. The panel is
then trained in the use of the scales (which ideally have been anchored with the help of exemplary stimuli) and in the discrimination of the perceptual factors that they refer to. As a result, it is hoped that subjects can grade company products reliably and consistently in terms of these sensory properties. The resultant direct attribute ratings can be analysed using traditional statistical methods such as ANOVA [Field, 2000]. Figure 3.1 outlines the basic steps involved in QDA.

Figure 3.1: Schematic illustration of the basic steps involved in QDA

Regarding the suitability of QDA to elicit the detectable attributes of a stimulus set in an unbiased manner, some concerns have to be raised. Strictly speaking, QDA leads to a cross between provided and elicited constructs, as subjects are influenced and possibly biased by each other during the stimulus presentation and discussion stages, but at least they have a say in the choice of scales and their definitions [Berg & Rumsey, 1999a]. Also, the subjects will not completely agree on all their perceptions any more than will there be complete agreement on all the attributes [Stone & Sidel, 1993]. The individuality of each panellist (in terms of sensitivity, motivation and personality) is likely to further add to the complexity of the process of devising a shared terminology.

3.2 Elicitation of individualised attribute scales

In contrast to QDA, the techniques presented next do not arrive at a structured definition of attributes through discussion. Instead, they allow for individualised attribute scales. That is why methods such as Free-Choice Profiling® (FCP) and the Repertory Grid Technique (RGT) are claimed to have advantages of lack of bias [Berg & Rumsey, 1999a]. Below, these two methods will be briefly outlined, pointing out those aspects that have a bearing on the design of a strategy suitable for verifying the perceptual organisation of a group of stimuli.

3.2.1 Free-Choice Profiling (FCP)

Having been developed as a solution to the problem of panellists using different wordings for a given attribute, FCP allows each subject to invent and use as many terms as (s)he needs to describe the sensory aspects of a set of samples. Therefore, the need for agreement between assessors, which is often very difficult if not impossible to achieve [Williams & Langron, 1984], is eliminated. The collected semantic data are submitted to a Procrustes Analysis [Dijksterhuis, 1996] – a class of statistical techniques combining aspects of Principal Component Analysis (see Section 3.2.2) with methods of rotation, translation, centring and reflection to produce a consensus configuration or
perceptual map of stimuli from individual data [Bech, 1999]. Figure 3.2 outlines the basic steps involved in FCP.

![Figure 3.2: Schematic illustration of the basic steps involved in FCP](image)

One advantage of FCP is that it represents a time-effective alternative to the protracted QDA by not requiring any screening or training of the panellists. Also, since each subject is exposed to the stimuli of interest separately and panel conformity is not an issue, the obtained attributes are more likely to reflect the subjects’ true perceptions. Nonetheless, while the absence of a response format can help minimise the experimenter’s influence, statistical analysis will be more complicated as subjects may produce widely differing descriptors. Consequently, questions regarding the ability of the sensory analyst to interpret the resulting terms, combined from all panellists, need to be addressed. Since the experimenter must decide what the terms actually mean, the results may be coloured more by the perspective of the analyst than by the participants’ original verdicts [Meilgaard et al., 1991].

Fortunately, however, approaches exist that can help with systematising such qualitative data. Verbal Protocol Analysis (VPA), for example, is a methodology that allows classification of verbal descriptors of certain properties into different groups [Ericsson & Simon, 1993; Samoylenko et al., 1996]. This is often achieved by first determining a number of coding categories and then parsing the verbal reports in such a way that each resultant segment constitutes one aspect of the theoretical model under consideration. Depending on the task, appropriate cues for segmentation might be pauses, intonation or contours contained in the verbalisations. A slightly different segmentation rule, which is based on separating ideas, places emphasis on the actual content. Next, the obtained segments are encoded by finding the category that expresses the same information as the original verbalisations. Finally, the frequency of occurrence of each segment in the derived categories is computed. As such, a comprehensive original protocol can usually be represented more succinctly by a set of process frequencies, which in turn may be converted into numerical magnitudes that can be analysed by standard statistical methods.

### 3.2.2 Repertory Grid Technique (RGT)

The RGT is similar to FCP in that it allows subjects to develop their own individual descriptive language for a stimulus set. Being a means for structuring experiences [Kjeldsen, 1998], it encourages personal reflection upon the qualities of the items under test and definition of a personal set of constructs that can differentiate between them [Berg & Rumsey, 1999a]. Unlike FCP, the RGT is equipped with a specific elicitation protocol. In particular, the experimenter presents triads of stimuli
to one subject at a time who is asked to describe in what way two of them are mutually alike but different from the third stimulus. This process is repeated until the subject stops creating new answers. In this way a list of terms is generated that are assumed to fully describe the group of stimuli. A grid is then constructed upon which the subject rates the different stimuli according to his/her own previously elicited constructs. Optionally, the ratings can be analysed both for the individual subjects and the panel as a whole to see whether it is possible to condense the obtained descriptors to a common set of attributes. In this context, multivariate analysis methods such as Principal Component Analysis (e.g. [Martens & Giragama, 2002; Zacharov & Koivuniemi, 2001c]), Factor Analysis (e.g. [Solomon, 1958; Gabrielsson & Sjögren, 1979]) or Cluster Analysis (see Section 3.3.1) can be employed. Hence, a form of data reduction is carried out to extract a smaller number of sensory attributes (called ‘components’, ‘factors’ or ‘clusters’, depending on the technique that is used). These can be labelled by examining their respective weightings on the different (elicited) terms and how the analysis grouped the original information. The resultant summated scales are then used to collect data containing a complete product profile. Figure 3.3 outlines the basic steps involved in the RGT.

Figure 3.3: Schematic illustration of the basic steps involved in the RGT

The fact that subjects have to comply with a pre-specified response format means that analysis of RGT data should be easy compared to FCP. Yet, by restricting subjects in such a way there is a risk that they may be unable to express their sensations accurately, although this problem should be less of an issue than with QDA. Bech [1999] pointed out that the RGT has been criticised for (still) being influenced by the experimenter and that this criticism resulted in FCP. However, as part of an empirical study that directly compared RGT to FCP it was found that the influence of the experimenter in the RGT is limited, but that FCP would be the better choice as an experimental technique as no time is spent on developing terms [McEwan et al., 1989].

3.3 Non-verbal scaling techniques

Non-verbal scaling techniques are statistically based approaches that enable disentanglement of the multidimensional structure of data sets stemming from sensory analysis applications. The main difference to DA methods is that these techniques work on some form of similarity judgements rather than verbal descriptors and direct attribute ratings. The two procedures that fall into this category are Cluster Analysis (CA) and MDS. Again, a short overview of each of these two techniques will be given, followed by a delineation of their respective values and handicaps.
3. Methods for studying the subjective effects of sound-processing algorithms

3.3.1 Cluster Analysis (CA)

CA is an analytical technique for obtaining meaningful subgroups of stimuli. It has been described as a "tool of discovery" [Anderberg, 1973], because one of its most useful purposes is to generate ideas about stimulus structure. The number of 'clusters' suggested by such a preliminary analysis is often treated as a hypothesis to be tested on a new data set. Specifically, the objective often is to classify a set of stimuli into a smaller number of mutually exclusive groups based on their similarities.

CA usually involves three steps [Hair et al., 1998]. Firstly, some sort of similarity (or correlation) measurements among the stimuli are collected in order to determine how many groups exist in the sample. Next, the stimuli are partitioned into clusters. The final step is to profile these groups of stimuli to determine their attribute composition. Figure 3.4 outlines the basic steps involved in CA.

Figure 3.4: Schematic illustration of the basic steps involved in CA

CA can be a very useful data-reduction tool, but there are several caveats. For instance, an immediate difficulty in applying CA is that there is no straightforward way of determining the actual number of clusters found. Although some rules of thumb have been derived that can assist with this decision, "the literature stresses that CA is more or less an iterative process where the analyst's conception of the process which generated the data is important" [Berg & Rumsey, 2000a]. Hence, the experimenter him/herself may play an instrumental role in the making of key decisions such as ascertaining the dimensionality of a solution. Furthermore, empirical evidence is available which shows that different clustering algorithms applied to the same set of data will often produce structures that are substantially different [Chatfield & Collins, 1980], so solutions are not (necessarily) unique. Similarly, the ability of CA methods to detect non-existent clusters is well established. This led Hair et al. [1998] to emphasise that because of its rather subjective nature CA is more an art than a science. Much caution is necessary when trying to interpret the results of a CA, which (on its own) can only be used to 'get a feel' for a data set.

3.3.2 Multidimensional Scaling (MDS)

Like CA, MDS is essentially non-verbal in nature and hence does not impose (or require the elicitation of) complex descriptive terms upon (from) subjects. As a result, there is little chance of bias or distortion owing to differences in understanding of semantic meanings [Borg & Groenen, 1997]. When using this technique, the primary objective is normally to search for and define the fundamental perceptual constructs or dimensions assumed to underlie a set of stimuli. This is accomplished...
through examination of the mathematical relationships inherent in similarity (or sometimes preference) judgements made by subjects.

In essence, MDS is concerned with finding a configuration of points that reflects the stimuli’s psychological organisation. More precisely, the objective is to transform judgements of similarity into (Euclidean) distances to be represented in multidimensional space. All stimuli are then positioned in the way that best corresponds to these distances. The resulting perceptual map shows the relative locations of the stimuli, which can provide an intuitive graphical representation if the solution does not contain more than three salient dimensions. Figure 3.5 outlines the basic steps involved in MDS.

![Figure 3.5: Schematic illustration of the basic steps involved in MDS](image)

The main disadvantage of MDS (and CA) lies in the difficulty with interpreting these dimensions (or clusters). The results of the analysis are purely numerical in nature and cannot provide an insight into their (psychological) meaning. Additional information is needed to make sense of the pattern revealed, i.e. which attributes predict the position of each stimulus. In practice, the labels given to the dimensions are usually based on the findings of ancillary descriptive evaluation. MDS also suffers from the problem of how to determine the ‘correct’ dimensionality, though to a lesser degree than CA as clearer numerical guidelines are available (e.g. see [Hair et al., 1998]). This is because MDS has been applied in innumerable studies aiming to identify the perceptual relationships among a group of stimuli (see also Section 3.4). Hence, a wealth of experience and information with respect to the use (and interpretation) of MDS (solutions) is at hand.

### 3.4 Discussion

As evident from above, each of the three presented attribute identification paradigms has distinct advantages and disadvantages, which have to be taken into account when designing a strategy suitable for confirming the unidimensionality of a stimulus set. From a statistician’s viewpoint, it is beneficial to have common scales, because the responses from multiple subjects can be readily analysed together and inferences can be drawn regarding the preferences of the general population. However, due to its consensus approach to language development and the experimenter’s influential role in this process, a DA technique is deemed unsuitable for achieving impartial stimulus verification. Techniques in the second group are less biased as they specifically avoid subject training and allow for individualised responses with regard to stimulus qualities, but data analysis is not as straightforward. The third
category is even less 'contaminated' by external influences, because semantic difficulties are eliminated entirely, but interpretation of the obtained stimulus structure is problematic.

Since there does not appear to be a single technique that supports valid and reliable sensory verification, a hybrid method has to be devised. The difficulty humans have in articulating their sensations and the inherent ambiguity of semantics are both well documented (e.g. see [Mason et al., 2001]). By using the similarity judgements of non-verbal scaling, the experimenter does not have to rely on a highly subjective and potentially incomplete list of descriptors. Also, adjective data can be extremely noisy and often contain less structure when compared with similarity spaces [Schiffman et al., 1981]. This is because similarity ratings are easier to make for subjects, as problems with semantics (such as misunderstandings of attributes) are virtually non-existent. Thus, non-verbal scaling appears to be useful for establishing an 'objective' basis for unravelling the perceptual properties of a group of stimuli. Verbal descriptors, on the other hand, seem best employed to help understand a multidimensional space derived from similarity data – not to derive one.

In view of the indisputable deficits of CA and its highly exploratory nature, MDS clearly is the better choice of non-verbal scaling techniques for a confirmatory application like this one. For the collection of supplementary semantic data a 'free verbalisation' approach (such as the one taken by FCP) seems to be most valuable, because the subjects' verbalised perceptions are expected to best correspond to their actual sensory experiences. Parenthetically, this combination of MDS for determining dimensionality and additional verbal data for naming the revealed constructs has also been the preferred technique for other perceptual researchers who studied sound quality in general (e.g. [Gabrielsson et al., 1974]) or specifically timbre perception (e.g. [Martens et al., 2000; Mattila, 2003]). It is true that the problem of systematising the semantic responses will have to be dealt with, but due to the highly controlled perceptual properties of the stimuli to be analysed the data will not be very diverse.

Finally, it is important to note that the methods reviewed in Section 3.1 and Section 3.2 are usually employed to produce a sensory profile of complex, multidimensional stimuli. Consequently, they rely on the collection of long lists of descriptors in order to ensure that no salient characteristics are left out, as having too few attributes would mean that subjects might not be able to differentiate stimulus differences. Since the stimuli to be verified are intended to be very homogeneous, it makes sense to initially use a technique like MDS that can help decide whether the stimuli's organisation can be portrayed in a perceptual space of low dimensionality or not. In consideration of the intended application, the suggested strategy is thus more efficient than the other procedures outlined above. Its efficiency is further enhanced by the absence of a rigid attribute elicitation and training protocol.
3. Methods for studying the subjective effects of sound-processing algorithms

On the whole then, if used in conjunction, MDS solutions and free verbalisations provide the most appropriate way to investigate empirically the degree of success achieved in unidimensional attribute simulations.

3.5 Summary

This chapter presented different methodologies that enable identification of the perceived attributes of a set of stimuli. In particular, three categories of sensory evaluation paradigms were presented and their respective benefits and drawbacks highlighted. It was concluded that in order to accomplish valid and reliable validation, the most appropriate option for this research is to use a hybrid procedure that utilises MDS to disclose the (low) dimensionality of a group of (notionally homogeneous) samples as well as subject-dependent attribute elicitation as a means of assigning meaningful labels to those dimensions. A welcome side-effect of this approach is that it is more efficient than the possible alternatives.
4. Extant perceptual sound processors related to spatial quality

4 EXTANT PERCEPTUAL SOUND PROCESSORS RELATED TO SPATIAL QUALITY

When dealing with the simulation of psychoacoustic phenomena, audio engineers have two options at their disposal. They can either adopt a physically or a perceptually based modelling strategy. The objective of the former is to emulate the transformation of sound in an enclosure as precisely as possible. In this case, simulation of all acoustical relationships that contribute to the formation of the sensation of interest is required. This usually means that a very elaborate and hence computationally expensive geometrical model has to be constructed, which then serves as the frame of reference for any subsequent calculations. Conversely, a modelling strategy based on human perception can make use of perceptually valid simplifications warranted by findings from associated psychophysical studies. By deciding on what is important for emulating a particular effect and what is not (e.g. what will be masked), the audio engineer can make substantial savings in terms of signal-processing costs. Further, opting for the latter approach means that a user-interface can easily be designed, which will enable direct manipulation of the stimulus properties that give rise to specific sensations in a listener. Such an interface is advantageous in that it saves the user from having to deal with any low-level implementational details. Control can be provided with the help of a few buttons and sliders, whereby moving each of these produces the same perceived effect as regulating several (relevant) physical parameters concurrently. Hence, the total number of variables that need to be adjusted for synthesising or manipulating an auditory scene can be kept to a minimum. Although it should also be possible to devise a user-interface with such favourable characteristics for physically based simulations, ensuring its psychological veridicality would take much longer.

Thus, in order to be able to design a training toolkit that provides intuitive demonstrations of certain auditory effects and that is also user-friendly, a perceptual approach is more efficient. That is why existent spatial sound processors featuring a perceptually motivated processing framework and/or interface will be examined in the following sections, to determine their applicability to unidimensional attribute simulations. For reasons of clarity, processing engines offering control over several spatial attributes are to be presented as a whole, rather than discussing their discrete ‘building blocks’ separately. This will be followed by a review of other available attribute synthesis methods, which do not belong to any spatial sound processing ‘package’.

4.1 The origin

Chowning [1971] probably introduced the principle of a perceptually oriented user-interface to the audio community. His moving sound source simulator gave the user independent control over the lateral displacement and perceived range of a sound source for the first time. The implementation was based on a 4-channel reverberation scheme and relied on varying the D/R to create the illusion of changes in the closeness of a sound image. In addition, the Doppler effect was included in the
simulation to supplement the credibility of the movement of a source. Other researchers followed him along the same lines, increasing the sophistication and complexity of their algorithms as well as adding other spatial attributes. All the systems that this author is aware of are outlined below.

4.2 The spatial reverberator
Kendall and his colleagues devised a system – much more complex than Chowning’s – which they called ‘spatial reverberator’ [Kendall & Martens, 1984; Kendall et al., 1989]. A signal-processing network was developed so as to be able to imitate closely the characteristics of reflected sound predicted with the help of an image-source model. The network accurately simulates all first- and second-order reflections, which are passed through a special “directionalizer” sub-system. This sub-system superimposes idealised pinna cues onto each reflection to allow them to be reproduced from the correct directions in the modelled room. A simplified block diagram of the signal-processing network is given in Figure 4.1.

Moreover, an interface based on perceptual qualities related to reflected sound was designed to provide users with direct control over auditory space percepts. In this context, the following four subjective impressions (believed to be important to the manipulation of spatial imagery in music production and reproduction) were incorporated: ‘distance’, ‘definition’, ‘spaciousness’ and ‘spatial texture’. Descriptions provided for each of them specified that ‘definition’ includes source-related attributes such as ‘spatial extent’ and ‘focus’. In contrast, ‘spaciousness’ comprises room-related attributes such as ‘liveness’, ‘size’ and ‘shape’ of a space. This suggests that both these impressions constitute categories of perceptual constructs rather than discrete attributes of a spatial sound scene. ‘Spatial texture’ was (somewhat vaguely) described as “the perception of the interaction of the sound with its environment” and “the quality, which differs between two spatial images having the same distance, definition, and spaciousness”. The latter description is reminiscent of the way timbre has
often been defined$^{16}$, i.e. it goes only as far as specifying what spatial texture is not instead of pinning down its perceptual effects directly. Such a definition is problematic, because it cannot elucidate what kind of spatial change Kendall et al. were referring to, thus prohibiting it from comparison with verbal descriptors used by other researchers that potentially describe a similar or even identical phenomenon.

In terms of psychoacoustic relationships, Kendall et al. stated that all four subjective impressions are determined by particular spatial and temporal characteristics of reflected sound. With respect to 'distance', it was reported that changes could be perceived even when only reflections with time delays of 33ms or less relative to the DS were included in the simulation. Yet, they admitted that the resultant spatial image was 'dry' and gave little impression of the modelled room upon which the simulation had been based. Regarding the other three subjective impressions, no details were given as to how they relate to sound field characteristics or how their manipulation was accomplished. Therefore, rebuilding the system (which is commercially not available) is not an option. In any case, no effort was made to verify the perceptual outcomes of the devised simulations. Such formal verification is considered to be of paramount importance to this work if attribute simulations are to be as unambiguous as possible.

Hence, it can be deduced that because of all the aforementioned shortcomings the spatial reverberator is inappropriate for this work.

### 4.3 The Spatialisateur (Spat)

Arguably the most comprehensive spatial sound processor to date, the 'Spatialisateur' (Spat) was developed at the Institut de Recherche et Coordination Acoustique/Musique (IRCAM). Being aimed primarily at musicians and sound engineers, the Spat was equipped with a user-interface giving control over parameters "directly related to audible sensations" [Jot & Rault, 1998]. The system's design was based on psychoacoustic research conducted at IRCAM to determine "the perceptive factors governing human perception of room acoustics" [Jullien et al., 1992]. During various studies selected objective acoustical criteria were varied independently to construct different artificial sound fields, which in turn were played back to a listening panel in a pairwise fashion. Subjects were then asked to grade each sound field pair's degree of similarity and the resultant data were analysed using Individual Differences Scaling (INDSCAL) to extract the number of perceptually salient dimensions contained in the judgements$^{17}$. This resulted in 52 dimensions (but no other details regarding the INDSCAL solutions were given). In a further step, correlations between these dimensions and various room acoustical criteria were computed so as to gain an insight into the psychological effects experienced by the listeners. It was found that the 52 dimensions could be described by 11

$^{16}$ The American National Standards Institute [1973] defined timbre as "that attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" (other definitions also include the duration of the two sounds). Thus, timbre is defined by the absence of relationships with other perceptual variables.

$^{17}$ INDSCAL is a derivative of MDS and will be discussed in Section 5.5.2.
4. Extant perceptual sound processors related to spatial quality

independent perceptual factors, which were (somehow) verbalised subsequently and for each of which a corresponding objective criterion (with the best overall correlation) was proposed [Jullien et al., 1993].

When devising a generic impulse response to be used as the basis for simulating such perceptual effects, the Spat’s developers found that a separation into DS, ERs and late reverberation was insufficient. Proposing that two different kinds of reflections exist in-between the DS and the onset of diffuse reverberation, they subdivided the impulse response into four parts [Jullien et al., 1993]:

1. Direct sound: \( t < 20 \text{ms} \)
2. Directional early reflections: \( 20 \text{ms} < t < 40 \text{ms} \)
3. Diffuse early reflections: \( 40 \text{ms} < t < 100 \text{ms} \)
4. Diffuse late reverberation: \( t > 100 \text{ms} \)

In addition to these temporal limits, each part is specified further by the energy contained in three frequency bands (low, mid, high) [Jot & Rault, 1998]. Parameterising the impulse response in this way enables emulation of various source-related (i.e. ‘source presence’, ‘source warmth’ and ‘source brilliance’) and room-related (i.e. ‘room presence’, ‘running reverberance’, ‘envelopment’, ‘late reverberance’, ‘heaviness’ and ‘liveness’) perceptual parameters. These have been implemented in the Spat and their respective groupings can be seen in the graphical user-interface shown in Figure 4.2. However, it is unclear how these nine parameters relate to the 11 (differently labelled) perceptual factors that were identified using INDSCAL and correlation analyses (see above).

Figure 4.2: The graphical interface used within the Spat. Along the top, the groupings into source- and room-related attributes can be seen.

In order to investigate whether the Spat is suitable for unidimensional attribute simulations, an informal listening test was executed by this author. With regard to the Spat’s 3/2-stereo compatible configuration, two general observations could readily be made: (1) none of IRCAM’s perceptual
4. Extant perceptual sound processors related to spatial quality

parameters (as controlled by each slider) is truly independent and (2) non-spatial changes occur as well. To give an example, increasing 'envelopment' clearly causes a source broadening, but this is accompanied by a separately perceivable HF boost. 'Running reverberance', on the other hand, appears to produce some form of environmental envelopment, i.e. an initially frontal room impression changes to a surrounding one. Yet, the sound source also gets wider because 'envelopment' increases simultaneously (see below).

When scrutinising the perceptual interface and its way of enabling manipulation of the various parameters, one finds that most of the sliders themselves are not independent either, i.e. they act as master controls for up to two other sliders. For instance, altering 'source presence' also affects 'running reverberance' as well as 'envelopment'. Jot and Rault stated that this particular feature was implemented so as to allow for convincing changes in source proximity by simply varying the 'source presence' slider [Jot & Rault, 1998]. Nonetheless, it was found that by changing the settings manually a more natural sounding and hence credible simulation of source distance could be achieved, especially if 'source brilliance' was varied, too. Similarly, the slider 'running reverberance' also controls 'envelopment', which is why it does not seem to be possible to increase the spatial attribute of environmental envelopment without affecting source width perception at the same time.

Overall then, the Spat appears to be deficient as far as the unidimensional simulation of spatial attributes is concerned. This is because the perceptual parameters as controlled by the user-interface are far from orthogonal and often cause non-spatial changes, too. In fact, it seems that most of them are related only indirectly to the spatial attributes of reproduced sound as defined in Section 1.3. For example, 'heaviness' and 'liveness' relate to acoustical qualities of a space, but they are not inherently spatial in nature. In this context, Rumsey [2002] pointed out that "there is rarely a direct mapping from these "virtual acoustics" parameters to ... spatial attributes". This led him to distinguish between 'attributes of spaces' (as used by IRCAM) and 'spatial attributes' (as employed for this work), whereby the latter embrace constructs relating to the directionality, height, depth and width of reproduced sources, groups of sources and acoustical environments. On the contrary, IRCAM's perceptual parameters appear to apply to the perception of natural acoustical spaces and are thus deemed unusable for this work. It should also be remembered that the Spat does not offer manipulation of discrete attributes by means of single faders or buttons – a desirable feature for a spatial ear trainer that is to be as intuitive and easy to use as possible.

4.4 Virtual Surround Panning (VSP)

Studer developed a collection of signal-processing techniques for their D950S digital mixing console, which they termed 'Virtual Surround Panning' (VSP). The aim was to provide tools that could be used to manipulate and enhance spatial sound scenes based on the 3/2-stereo surround format. Extending traditional capabilities of mixing desks, VSP comprises head-related, frequency-dependent panning algorithms that are claimed to enable stable positioning of sound images even outside the
optimal listening position [Horbach, 1998]. The design of these panpots was based on the psychoacoustic principles of Theile’s "association model". Theile [1991] proposed that a natural stereophonic sound image with convincing sensations of depth and space can be achieved only if interaural differences typically produced by real sources are introduced into the loudspeaker signals that are used to create phantom images of such sound events.

Another feature of VSP is the automatic generation of ERs that allow positioning of a sound image behind the frontal loudspeakers [Horbach et al., 2000]. The simulation of these reflections is dynamic, i.e. the reflection patterns are modified according to the chosen source location. For each channel 16 discrete reflections are generated, whereby their finer structure is controlled using four perceptual parameters. 'Ambience' and 'distance' affect the level ratio between DS and discrete reflections; 'distance' and 'room size' influence the temporal distribution of the reflections. The fourth parameter, 'absorption', controls the time-dependent level attenuation of the reflections. All of these parameters can be set within each channel to allow for source-specific distance adjustments. In Figure 4.3 a simplified block diagram of a single channel is shown.

Further, Horbach et al. pointed out that care was taken to optimise perceived distance whilst at the same time minimising colouration effects. In addition, a global diffuse reverberation tail was devised, the design of which was based on the Spat, hence utilising the perceptual parameters derived at IRCAM. Horbach [1997] also provided some information as to the development of a "stereo width" control based on a crosstalk cancellation circuit that was implemented in conjunction with the frequency-dependent panning algorithm. Figure 4.4 shows a block diagram of this circuit. Apparently, this technique enables the width of the sound image to be extended, resulting in "a very convincing beyond-the-speaker localization around the sweet spot".

Figure 4.3: Simplified block diagram of a single VSP channel (adapted from [Horbach et al., 2000])

Figure 4.4: Block diagram of the "stereo width" control circuit
Apart from Pellegrini's distance study discussed in Section 2.2.5, Studer did not report any form of perceptual verification of VSP. While it seems plausible that the combined use of these techniques can improve the spatial quality of a recording and that it offers greater flexibility than when using conventional amplitude panning and external reverberators, for example, the design of VSP somewhat departs from the goals of this work. To illustrate, 'absorption' clearly is not a spatial attribute and neither are most of IRCAM's perceptual parameters (see Section 4.3). Furthermore, several controls have to be adjusted in order to manipulate source closeness. These do not appear to be independent of each other, e.g. both 'distance' and 'room size' affect the temporal structure of the ERs employed to simulate source proximity. Thus, VSP resembles the Spat, which was already found to be unsuitable for this work.

### 4.5 The SceneBuilder

Work performed at McGill University was concerned with the development of an intuitive, perceptually based mixing tool that would allow users to synthesise sound scenes through manipulation of perceptual attributes of auditory spaces [Quesnel et al., 1999]. Termed the 'SceneBuilder', the system combines simplified physical and perceptual models [Corey, 2002]. It makes use of a digital mixer, a multichannel panner and reverberator, and two 2-channel reverberation units as well as custom-made digital signal processing (DSP) components (see Figure 4.5 for a block diagram of the whole system). The DSP components were devised for the synthesis of tracking reflections and the simulation of axial room modes encountered in real enclosures. According to Corey et al. [2001], these phenomena need to be taken into account when modelling acoustic spaces because the location of a source and receiver within a real room have a direct impact on many spatial characteristics such as the apparent level of room reverberation, the level of room modes, the apparent size of the sound source, and the frequency spectrum of both the source and the room.

Some details regarding the implementation of tracking reflections were already given in Section 2.2.5. To recapitulate, a fourth-order image-source model is used for calculating the gains, delay times and directions of 40 reflections according to source location and room dimensions, as referenced to a central listening position. The reflections are generated for a 2-D model of a rectangular room.
Pairwise constant power panning is used for their spatial encoding. To approximate wall and air absorption effects low-pass filters are employed. In addition, global ER gain adjustments are made as a function of source position, i.e. all reflections are attenuated by several dB as the source moves from a wall to the centre of the room [Corey et al., 2001] – presumably to complement the 'boundary effect' algorithm (see below).

To emulate the broadening of a source perceivable under natural listening conditions when it approaches a room boundary and that is accompanied by a LF build-up [Corey, 2002], a width control was devised. As part of this ‘boundary effect’ algorithm four “fuzzy sources” are generated by sending parallel feeds of the direct source (or sound) signal to the two external 2-channel reverberation devices (see Figure 4.5). These are configured to provide short impulse responses exhibiting a dense series of echoes that help create a diffuse sound image. The fuzzy sources surround the direct source and maintain the same relative distance from it, irrespective of source location. Dynamic level adjustments are made to ensure that when the direct source is located in the centre of the room all fuzzy sources are attenuated completely, but that two of them are increased in level as the direct source is moved towards a wall. At the same time, low- and high-shelving filters are applied to alter the spectral balance of both the direct and fuzzy sources in order to simulate air absorption and to help widen the sound image.

![Figure 4.5: Block diagram of the SceneBuilder (adapted from [Corey, 2002])](image)

In its favour, the SceneBuilder was subjected to a series of (fairly basic) psychoacoustic experiments to check if the modelling of various location-dependent sound field properties can perceptually enhance the simulation of apparent source position. The tests included verifying if listeners perceive the boundary effect emulation in the intended way, evaluating whether the addition of the dynamic properties of the system improves the simulation of source position as displayed on a visual interface and checking if the SceneBuilder supports small distance changes (see [Corey, 2002] for details). Nevertheless, the validation attempt falls short with respect to one aspect crucial to this work. Since
panellists were not told to listen for other changes in sound quality as well when judging, say, source distance, it was not demonstrated that the proposed system is free from any additional, unintended perceptual effects that could potentially confuse naïve listeners.

For the generation of exemplary stimuli, the SceneBuilder could in principle provide a useful processing environment. However, as to its suitability as a training system modifications to the user-interface would be required, which (apparently) does not provide direct control over discrete spatial attributes. Put simply, there is no single slider labelled ‘source width’, for example. Rather, the SceneBuilder contains elements of conventional room acoustics software, i.e. it requires the input of data for configuring various parameters related to the properties of the simulated model (e.g. room size), thereby making it less attractive for training exercises.

4.6 Other approaches for controlling width and distance attributes

Other approaches for the control of width and distance perception exist that were not developed as part of a stand-alone spatial processing application. A description of various techniques is given below. Anticipatorily, it can be said that their perceptual effects have yet to be scientifically validated.

4.6.1 Divergence

A widely known technique for spatially spreading monophonic signals is the so-called ‘divergence’ feature, which can be found in mixing consoles such as the Sony OXF-R3 or DMX-R100 [Robjohns, 2000]. This technique normally utilises simple pairwise intensity panning across a number of speakers (usually L, C and R) and thus represents an extension to the conventional L-R panpot. Divergence controls the L/C/R panning parameters, thereby regulating the proportion of sound mixed into each channel. Whilst it is far from being psychoacoustically optimised for manipulating perceived source width, it scores in terms of simplicity.

4.6.2 Stereo shuffling

A more sophisticated approach to width control is stereo shuffling, which is based on Alan Blumlein’s M-S concept [Blumlein, 1931]. Stereo shuffling was revived by Gerzon [1986] and can involve frequency-dependent width processing, i.e. it allows equalising the sum (or M) and difference (or S) signal of 2-channel stereophonic source material in different ways before recovering the original left and right channels (see Figure 4.6). This can be beneficial to widen certain stereophonic programme material, e.g. recordings made with two coincident cardioid microphones, which produce a monophonic output at low frequencies. Nevertheless, Gerzon stressed that stereo shuffling is not uniformly effective, i.e. there can be unwanted side-effects with source material that does not rely entirely on amplitude differences to construct the sound image. For instance, with time-based stereo, frequency-dependent cancellation effects occur when summing/subtracting the left and the right
channels. That is why the sound image will be perceived to be wider at some but narrower at other frequencies, resulting in a "possibly confused and degraded stereo image".

Due to the fact that stereo shuffling changes the degree of difference between the left and right channels, it can affect the attributes of source, ensemble and environment width all at once, the exact magnitude of each spatial change depending on the characteristics of the programme material that is processed. For a multiple-source stereo input signal, for example, it will broaden/narrow the individual sources as well as the ensemble as a whole. If reflected sound is present, variations in environment width are likely to be perceivable, too. Thus, this technique may be problematic as far as unidimensional attribute simulations are concerned.

4.6.3 Gerzon's source distance and width controls
As already mentioned in Section 2.2.5, Gerzon [1992a] argued that early lateral reflections with delays of up to 50ms relative to the DS provide the most important cues for distance perception. In this respect, he also claimed that simulating ERs based on room models does not give as reliable and controllable results as creating artificial ones generally unlike those found in a real room. Likewise, emulation of the cues produced by actual rooms can involve rather large amounts of signal processing and the resultant ERs may produce unpleasant comb-filter effects. Conversely, artificial reflection patterns can be designed to give an optimum distance illusion for all positions with a low degree of colouration. Based on these assertions, Gerzon designed an ER simulator that would imitate such cues without the need for complete room modelling. The quality of the perceived distance effect was claimed to depend on the number of simulated reflections, their spacing in time from one another, as well as their apparent stereo position, but none of these assertions was verified psychoacoustically. The simulator allows for gain and time delay adjustments in the direct and indirect signal paths. Amplitude panning is employed to enable the DS and reflections to be rendered with different angles of incidence, as directional diversity helps reduce subjective colouration and results in an improved spatial quality.

For the control of "source size" Gerzon [1992c] proposed various 'pseudo-stereo' processing methods that rely on a monophonic input signal being sent to two all-pass filters set up in parallel. These cause frequency-dependent phase shifts between the channels, leading to a perceived spreading of the sound if the interchannel phase difference changes sufficiently rapidly with frequency. A block diagram of one such circuit is shown in Figure 4.7. As pointed out by Gerzon, this image spreading makes
subtleties of acoustical resonances or reverberation contained in a monophonic sound signal more audible because of the psychoacoustic phenomenon of directional unmasking.

Regrettably, Gerzon did not go to the length of assessing his proposed algorithms by means of listening tests, nor did he investigate the perceptual salience of finer details of ERs he declared to be important to distance hearing. With respect to the second point, a formal subjective study carried out as part of this work suggests that for the successful simulation of source proximity changes a large number of ERs is not required (see Section 5.4.1 and Section 5.5.2). Neither does the finer structure of the reflections appear to be crucial for achieving a realistic distance illusion (see Section 5.4.1 and Section 5.5.2).

### 4.6.4 ZoomFX

To improve the authenticity of transaural sound rendering, British company Sensaura developed a source width algorithm, which they named ‘ZoomFX’ [Sibbald, 1999]. Their aim was to let large objects that are close to a listener fill the auditory space in the same way as they would fill the visual space, thereby making the synthesised sound scene more realistic. In essence, an array of secondary sources is created from a single original source in such a way that the sum of these secondary signals is equal to the original one. In addition, the secondary sources are decorrelated with respect to the primary sound and each other in order to enable the brain to perceive them individually. For this purpose, a processing method called Dynamic Decorrelation™ is employed. By distributing the sources over a prescribed area or volume that corresponds to the physical size of the sound-emitting object to be synthesised, the brain is tricked into interpreting these similar but distinct signals as a (composite) ‘volume source’. A block diagram illustrating this process is shown in Figure 4.8.
4. Extant perceptual sound processors related to spatial quality

ZoomFX was designed with virtual reality applications and the like in mind, which is why its design requirements depart widely from the ones of this work. Splitting musical programme material into discrete, decorrelated components and rendering the resultant signals from different positions is very likely to sound unnatural, which is why this approach was not pursued any further.

4.7 Summary

This chapter examined existent perceptually motivated spatial sound processors in order to determine their suitability for unidimensional attribute simulations. It was found that even though various systems have been devised, none of them has been (thoroughly) validated with regard to (all) the perceptual changes it produces. Yet, such detailed subjective testing is deemed essential for this work, because in order to achieve an optimum training effect, it has to be ensured that the attributes of interest can be demonstrated to listeners in a unidimensional manner. What is more, many of the systems do not model spatial attributes of reproduced sound as applicable to this work, but rather feature other psychological dimensions, e.g. room acoustical percepts. Finally, several of the reviewed processors fail to offer control over their respective perceptual changes 'by the push of a button'. Often, input of (textual) data and/or the adjustment of several faders are required to achieve the intended effect. While this aspect does not impede the creation of suitable reference stimuli, it jeopardises the accomplishment of an intuitive and user-friendly training tool.

In conclusion then, none of the extant perceptual sound processors related to spatial quality can be directly employed for this work. However, it is conceivable that elements of some of these systems could be used as starting points for the development of suitable attribute simulation methods. This will be discussed in more detail in Chapter 5.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

5 UNIDIMENSIONAL SIMULATION AND VALIDATION OF SPATIAL ATTRIBUTES OF REPRODUCED SOUND

Having delineated the background to this research, this chapter will present almost all of the practical work that was executed\(^{18}\). As stated in Chapter 0, this project aims to achieve unidimensional variation of spatial attributes of reproduced sound so as to facilitate listener training. To begin with, a set of attributes will therefore be chosen that shall be investigated. In addition, a reproduction system suitable for spatial ear-training applications is to be selected. This will be followed by an outline of the general experimental set-up used for the various psychoacoustic tests that were carried out. Next, an overview is to be given with regard to the basic approach that was adopted for the modelling and verification of the percepts under consideration. The bulk of this chapter, however, will be concerned with how each attribute was simulated and subsequently validated with regard to the intended subjective effect. For some of these attributes pilot studies had to be conducted and a suitable processing infrastructure implemented in order to allow their unidimensional variation. This work shall also be reported.

5.1 Which spatial attributes to choose?

As the outline of the multidimensional structure of spatial quality in Chapter 1 showed, a multitude of spatial attributes exist. In an ideal case, a spatial ear-training toolkit should be able to provide demonstrations for each of these dimensions. However, due to the imposed time limits not all of them can be addressed by this work. Thus, the question of which attributes should be covered for this thesis arises. In Section 1.4 it was pointed out that more research needs to be performed before valid inferences can be made with regard to the relationships between the various spatial dimensions and listener preference. That is why a decision is made based on considerations of practicality rather than perceptual importance. To be able to determine whether spatial attributes of reproduced sound can be simulated in a unidimensional manner (and ultimately whether naïve listeners can be trained in spatial sound perception), it seems sensible to start with constructs related to the geometrical dimensions of sources (see Table 1.1 and Table 1.2). Most psychoacoustic research conducted to date has dealt with such (descriptive) dimensions of spatial quality, e.g. source distance or source width (see Chapter 2). Since it is likely that the associated results are applicable to or can at least offer a good starting point for this work, these attributes should be easier to simulate than (semantically) more abstract percepts. Besides, it is expected that subjects will be more familiar with geometrical concepts compared to less tangible ones like presence or room envelopment, which are probably not as meaningful to them. As a consequence, the validation (and training) part should be less complicated, too. On these grounds, the constructs of source distance (SD), source width (SW), ensemble width (EW) and ensemble depth (ED) are made the subjects of the unidimensional attribute simulations to be undertaken for this thesis.

\(^{18}\) In Appendix E a pilot study into the training of listeners for spatial audio evaluation purposes will be outlined that was based on some of the attribute simulations described in this chapter.
5.2 Reproduction layout and experimental set-up

In addition to having to decide on which attributes to simulate, a reproduction format has to be picked that is able to convey these spatial characteristics to listeners in an unambiguous manner. As discussed in [Rumsey, 2001], various forms of spatial sound rendering are available, the majority of which are loudspeaker – rather than headphone – based. In Section 2.4 it was argued already that binaural reproduction is deemed inappropriate for spatial ear-training applications due to its susceptibility to spatial distortions. In-the-head locatedness, for example, inevitably imperils the unequivocal simulation of source distance and ensemble depth. Loudspeaker stereophony is generally immune from these problems, but, depending on the system that is used, may have other limitations as far as the reproduction of certain spatial dimensions is concerned (e.g. the inability of most loudspeaker-based systems to reproduce height information or to create ‘near head’ phantom sources). Although often criticised for its inadequacy as a musical spatialisation system, 3/2-stereo has some irrefutable advantages. Due to it being standardised 3/2-stereo is more widespread than other configurations, which should facilitate the transferability of a spatial ear trainer featuring this system at the reproduction end. It is true that much more sophisticated loudspeaker-based rendering approaches are available (e.g. Wave Field Synthesis [Berkhout et al., 1993], which, unlike the ITU set-up, does not treat different sectors of the horizontal plane unequally). Yet, due to their substantial requirements in terms of hardware and processing power they are very expensive to install and hence not commonly used. Rumsey [2001] pointed out that since 3/2-stereo does not directly support the concept of 360° horizontal image localisation, it may be better to treat the format as one with a conventional 3-channel frontal image and two surround channels for providing ambience, effects or ‘room impression’. Nonetheless, he also stated that “with such an approach it is still possible to create very convincing spatial illusions”. Therefore, for reasons of practicality and wide applicability the 3/2-stereo layout is chosen as the basis for the spatial ear trainer envisaged by this work.

In the remainder of this chapter various listening tests will be described, which were very similar in terms of their experimental set-ups. That is why the general physical configuration will be briefly outlined here, with any inter-experiment differences being reported in the appropriate sections. The psychoacoustic experiments carried out for this thesis took place in a listening room meeting the design specifications of [ITU-R BS.1116, 1997]. In accordance with [ITU-R BS.775, 1994], five Genelec 1032A loudspeakers were set up at 0°, ±30°, and ±110° and a distance of 2.1m from the optimal listening position. The loudspeakers were level-aligned to within 0.2dBA of each other using a pink noise test signal and a Brüel & Kjær 2123 real-time spectrum analyser with the measurement microphone being located at the listening position. Listening test software, used to automate the experiments and to save subjects’ responses to hard disk, was run on a computer, which was connected to a Yamaha 02R mixer via an optical interface for digital-to-analogue conversion of the stimuli. The computer monitor was positioned directly in front of the listening position, thereby allowing subjects to control the speed of each listening test as well as to switch between stimuli at their leisure. To eliminate the influence of any visual cues on respondents’ judgements, the listening
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

room was darkened and an acoustically transparent curtain was hung from the ceiling to conceal the position of the loudspeakers. A diagram of the experimental set-up is shown in Figure 5.1.

Figure 5.1: Diagram of general experimental set-up (listening room)

5.3 General approach to attribute simulations and validations

In Section 0.3 the benefits of unidimensional attribute simulations were outlined. To recap, such simulations will make it easier for subjects to learn a descriptive language and hence should lead to a reduction in training programme duration. Besides, they are expected to be of great value in helping to lay open connections between the perceptual and physical domains. That is why putting the necessary resources into the production of suitable reference materials seems worthwhile.

Chapter 4 reviewed extant spatial sound manipulators featuring a perceptually motivated processing framework and/or interface of some sort to give the user control over spatial characteristics of a sound image. Most of these systems allow the chosen attributes to be altered in real-time, thus offering an effective means of familiarising a listener with the associated qualitative effects. Although the implementation of a real-time controllable system equipped with a high-level interface is also envisaged by this work, such a tool is not a prerequisite for training listeners in spatial sound perception. As the summary of issues related to training procedures in the sensory sciences in Section 0.3 indicated, subjects are commonly acquainted with perceptual characteristics by means of references that differ in their intensities with respect to the attributes of interest. This is also the approach pursued for this thesis, mainly because of time limitations and the fact that sets of stimuli are required for the validation experiments (see Section 5.5.2).
The general course of action taken for the creation of unidimensionally varying stimuli entailed a form
of ‘analysis by synthesis’. Through an iterative process involving informal listening and a clear
knowledge of theoretical factors that have been found to relate to the construct in question, this author
would attempt to find a processing method allowing a convincing illustration for each chosen attribute
scale to be rendered. This would be followed by formal validation of the simulation’s success or
otherwise using a panel of critical listeners. The reader may recall that in Chapter 3 MDS was chosen
as the main sensory analysis tool, because it is a relatively neat way of verifying the dimensionality of
a group of (notionally homogeneous) samples. In addition, it was decided to elicit subject-dependent
verbal descriptors so as to be able to assign meaningful labels to the MDS dimensions, which could
then be analysed using a technique like VPA. Depending on the outcome of the validation stage, the
stimuli would then either have to be modified or could be declared appropriate for training purposes.
The latter scenario would also allow the physical characteristics of the modelled attributes to be
determined based on acoustic analyses of the created signals. Figure 5.2 outlines the basic steps of the
adopted approach.

It has to be pointed out that in generating such exemplary stimuli the aim was to achieve unmistakable
representations of four well-defined perceptual effects rather than physical verisimilitude. Therefore,
when trying to develop suitable algorithms, any means was felt justified as long as the resultant
samples would exhibit the intended auditory features. That is why the exaggeration of cues was
considered legitimate, too. The need for a certain range of stimuli for the validation phase (see Section
5.5.2) and the fact that reproduced sound can create images that do not necessarily occur in natural
listening make the departure from physical reality a valid means to an end.

In order to be able to produce sound excerpts altering along their respective perceptual scales in a
unidimensional way, all variables had to be limited to solely those of interest. Creating reference
samples based on live recordings was deemed inappropriate for this task, because personal experience
has shown that varying the microphone technique and/or placement with respect to a sound source
will give rise to a multitude of perceptual changes (both spatial and non-spatial). What is more, almost
all of the spatial attributes reviewed in Chapter 2 have been associated with reverberant sound. That is
why it seemed much more likely to achieve unidimensional variation with the help of highly
controllable synthesis techniques offering precise control over the temporal and spatial distribution of
reflections, for example. For these reasons, the decision was made to synthesise each set of stimuli
using anechoically recorded (or at least acoustically ‘dry’) monophonic source material as the DS
signals. Previous research has emphasised the advantages of simple yet representative programme material when conducting subjective tests on spatial sound perception [Rumsey, 1998; Ford et al., 2001]. Hence, source signals commonly encountered in natural hearing were employed for the various attribute simulations to be depicted below. To cater for the percepts’ dependency on reflected sound energy artificially generated reverberation was added to the DS signals, its physical properties being carefully adjusted to accommodate particular attribute requirements. Some aspects of the processing set-ups were changed between simulations (e.g. the number of reproduction channels used) – the various reasons for doing this being theoretical, practical and evolutionary in nature. Attribute-specific details regarding the stimulus creation process will be given in the following sections.

5.4 Simulation of source distance
In Section 2.2 the various established source range cues were discussed, suggesting that distance perception is most strongly influenced by changes in loudness, the D/R, the finer structure of ERs and the frequency spectrum. To be able to devise a processing method suitable for the creation of a set of exemplary stimuli, the relative merit of each of these factors was informally evaluated. The outcome of this work is presented below.

5.4.1 Creation of source distance stimuli
The SD stimuli were created in Studio 3 of the University of Surrey’s Department of Music & Sound Recording. Studio 3 is a multichannel surround sound control room, which conforms to many of the design criteria of [ITU-R BS.1116, 1997] including dimension ratios, noise floor, RT and loudspeaker arrangement. Nonetheless, it departs from the recommendation with respect to the required attenuation of ERs (arriving at the listening position within 15ms of the DS) due to the presence of a Sony OXF-R3 digital mixing console, a 19” computer monitor and an equipment rack. The 3/2-stereo reproduction set-up comprises five full-range, active loudspeakers (ATC SCM100As), each being located at a distance of 2.3m from the optimum listening position. A diagram illustrating the layout of Studio 3 is given in Figure 5.14.

In order to avoid problems caused by asking subjects to judge too complex stimuli, it was decided to synthesise a stationary source (positioned along the front-back axis directly ahead of the listener) for each stimulus. The (anechoic) DS signals consisted of an acoustic guitar or a cornet recording (i.e. a more transient and a more steady-state source), which were taken from the Archimedes CD [Hansen & Munch, 1991]. Both programme items were edited to a length of ~9s. Care was taken to ensure that the musical integrity of the extracts was preserved so as not to cause annoyance of the subjects during the subsequent listening test.

For the synthesis of reflected sound energy a Lexicon 480L reverberation processor was used. This reverberator is equipped with a 4-channel digital output and offers control over up to six discrete reflections, which are individually real-time adjustable in terms of their delays and amplitudes relative
to the input signal. A room preset was selected having a RT of ~1s at mid frequencies. Based on the findings of other researchers outlined in Section 2.2.5, a generic impulse response was designed consisting of the following three time regions:

1. Direct sound: \( t = 0 \text{ms} \)
2. Early reflections: \( 15 \text{ms} < t < 40 \text{ms} \)
3. Reverberation tail: \( t > 40 \text{ms} \)

The choice of 40ms (rather than 50ms) as the upper time limit for the ERs was meant to prevent them from being spaced too widely along the time axis. Otherwise they might become audible as discrete echoes in the case of transient source material. To avoid conflict with the ERs the onset of the diffuse reverberation was delayed appropriately. Figure 5.3 illustrates the generic impulse response of the reverberator graphically.

![Figure 5.3: Generic impulse response of the reverberator used for the simulation of source distance showing the direct sound, the six early reflections and the reverberation tail](image)

The stimuli were mixed for reproduction over L, C and R only, the DS being routed to the centre speaker. Since ERs arriving from lateral directions have been claimed to be important to distance hearing (see Section 2.2.5), it was decided to send three reflections to L and R, respectively. All reflections had the same level (see below) and were processed using a 6dB/octave low-pass filter with a cut-off frequency of ~4.5kHz. Once a suitable ER pattern had been found this was not altered any further. Admittedly, this is a rather crude approximation to the conditions encountered in a natural acoustic, and it also contradicts some of the research findings outlined in Section 2.2.5. Yet, it was found that the reflection pattern provided sufficient flexibility for creating a number of sounds with different degrees of source proximity, which bear a strong qualitative resemblance to what happens in the real world. The two channels of decorrelated reverberation were also mixed into L and R, respectively. LS and RS were not used because informal listening tests had revealed that they are not required for simulating source distance. Thus, in order not to run the risk of exciting any unwanted spatial dimensions by providing reflected energy from the sides and rear of the listening position, they were simply omitted.
The perceptual effect of changes in the range of a source was produced by varying the relative gains between the three chosen time regions with the help of the mixing console. To simulate a decrease in source closeness the DS level was reduced in a monotonic manner, resulting in a maximal attenuation of approximately -12dB for the most distant stimulus. At the same time, the ERs (which for the closest stimulus were at a level of about -19dB relative to the DS) were raised slightly before being reduced in amplitude again, the most distant stimulus having an ER level of approximately -25dB with respect to the DS of the closest source. Whilst it is acknowledged that such an intermediate level increase does not occur under natural acoustic conditions, it was found appropriate in the case of this simulation\(^\text{19}\). As to the level of the late reverberation, it can be assumed that in a small space there will be little variation at the receiver position for a sound source radiating a signal at constant intensity but varying distances [Chowning, 1971]. Yet, to intensify the distance effect the level of the diffuse reverberation was raised slightly (i.e. by a few dB) for those sound examples to be used at the 'far' end of the distance scale. In addition, a low-pass filter with a cut-off frequency of 10kHz and a slope of 12dB/octave was inserted into the signal path of the DS for the most distant stimuli, thereby giving rise to a slight HF attenuation. This was meant to imitate the effect of air absorption. For reasons outlined in Section 2.2.1, LF boosts measurable for (real) close sound sources were not accommodated in the simulation. An attempt was made to achieve linear spacing of the sounds to facilitate the detection of differences during the validation (and training) stage of this study. To achieve perceptually equal step sizes it was found necessary to change the DS level in a non-linear manner. More precisely, the level reductions needed to be bigger for the more distant stimuli, which substantiates the findings of other researchers, e.g. the ‘distance compression’ phenomenon identified by von Bekesy (see Section 2.2.3). Parenthetically, it may be that the intermediate increase in ER level (see above) helped compensate for insufficient reductions in the DS level of these stimuli. In other words, perhaps a monotonic decrease in the level of the ERs combined with larger level changes in the DS signal of the intermediary stimuli could have also resulted in a credible distance simulation.

A block diagram of the algorithm devised for manipulating source distance is shown in Figure 5.4.

\(^{19}\) This could be due to the physical layout of the loudspeakers, emphasising the fact that reproduced sound may not be the same as natural acoustics (depending on the method of rendering).
Based on the methodology delineated above, nine different SD stimuli were created as required for the validation phase (see Section 5.5.2). The entire processing took place in the digital domain and the sound excerpts were recorded to hard disk as 3-channel Audio Interchange File Format (.aiff) files with a sampling rate of 44.1kHz and a resolution of 16 bit.

5.5 Validation of source distance stimuli

Once a set of test stimuli had been generated, a validation experiment was conducted to verify whether the intended unidimensionality had been achieved. For this purpose, eight final-year students and one graduate of the University of Surrey's 'Tonmeister' degree course were employed as the listening panel. As part of their education the students had received considerable training in the detection of small changes and impairments in sound quality. Also, all of them had participated in psychoacoustic tests before. The rationale for using such experienced and critical listeners was that if their responses did not contain references to any unwanted differences, the stimuli would almost certainly be free of unwanted artefacts. Hence, unidimensionality of the simulation could be inferred, thereby authorising its use for training programmes. All subjects participated on a voluntary basis, i.e. none of them was remunerated for their time. Before the start of the test it was made sure that no questions about the experimental procedure remained. However, no information about the nature of the investigation was given until the experiment had been completed. Only one type of source material was experimentally verified, the acoustic guitar material being arbitrarily selected.

5.5.1 Experimental design

As evident from the discussion of different methods allowing the identification of the perceivable attributes of a stimulus set given in Chapter 3 and the outline of the general approach adopted for this work provided in Section 5.3, the validation of the attribute simulations was based on MDS. To reiterate, MDS requires each stimulus in a given group to be compared with every other stimulus of the same group. Since it is an attribute-free technique, participants do not make these comparisons with respect to highly subjective and hence potentially misinterpreted verbal descriptors of a certain quality. Rather, all stimulus pairs are assessed in terms of their overall similarity. This is beneficial in that subjects do not have to try to understand and adopt pre-specified attribute scales. The collected similarity judgements are then transformed into Euclidean distances, which in turn are represented in multidimensional (Euclidean) space [Hair et al., 1998]. For instance, the rated degree of dissimilarity of stimuli $i$ and $j$ is mathematically modelled as:

$$d_{ij} = \left[ \sum_{r=1}^{R} (x_{ir} - x_{jr})^2 \right]^{1/2}$$

where $d_{ij}$ stands for the Euclidean distance between stimuli $i$ and $j$, $x_{ir}$ is the co-ordinate of stimulus $i$ on the $r$-th dimension and $R$ is the number of dimensions in the Euclidean space.
As a result of the absence of attribute scales, MDS avoids the problems of ambiguity that such semantic descriptors can cause (e.g. see [Mason et al., 2001]). That is why it has a reputation for being a relatively bias-free method for measuring human perception. Nevertheless, MDS techniques are limited insofar as the unravelled psychological structures need to be interpreted. That is to say that MDS cannot provide the meanings or labels for these perceptual dimensions. Instead, they have to be found by other means. In order to address this shortcoming, it was decided to gather supplementary verbal responses after the completion of the MDS exercise. In particular, all listeners were asked to verbally express the differences they had perceived between the stimuli. A questionnaire was provided for that purpose instructing respondents to write down words and descriptions differentiating between stimuli in order of priority. Before the start of the verbal reporting stage, listeners could listen back to the sound excerpts if they felt they had to.

Based on the experimental design outlined above, it was hoped that a single strong dimension would be revealed, thus allowing the unidimensionality of the stimuli to be concluded. Since there were nine (see Section 5.5.2) stimuli to compare, each subject made a total of $9(9-1)/2 = 36$ gradings. A different order of presentation was created for each listener to minimise carry-over effects. The test was subdivided into three groups of 12 trials. After the completion of each group subjects were offered a short break in an attempt to reduce listener fatigue.

The user-interface implemented in the listening test software is shown in Figure 5.5. As can be seen, an undifferentiated line scale was provided for the subjects to indicate their perceptions, the scale ends being labelled 'Same' and 'Most different'.

Figure 5.5: User-interface employed for the MDS experiments

Using written instructions, participants were informed that they had to make global similarity judgements taking into account any and all detected differences when grading a pair of sounds.
addition, they were told to give the 'Most different' grading to those sounds that appeared to be the two most dissimilar ones out of the whole group. In order for them to get an idea of the range of differences, subjects were given the opportunity to listen to all nine sounds before and halfway through each group of 12 trials. Hence, it was hoped that they would be able to familiarise themselves and refresh their memories with respect to the magnitude of possible differences between the stimuli and, as a result, would grade the stimuli in a consistent manner.

5.5.2 Results

Analysis of similarity data

To be able to analyse the similarity judgements from the validation experiment, the data from each subject were converted into proximity half-matrices and then entered into the statistical analysis software package SPSS. The chosen layout of the analysis was based on statistical requirements imposed by the format of the collected similarity data.

The number of stimuli used during the data collection phase determines the maximum dimensionality a meaningful MDS solution can portray. As a rule of thumb, for a given solution to be stable more than four times as many stimuli as dimensions are required [Hair et al., 1998]. While this does not prevent the experimenter from executing analyses for higher dimensionalities, violation of this guideline means that results have to be regarded as very tentative until replicated with more stimuli [Kruskal & Wish, 1978]. Since for this experiment listeners had compared a set of nine stimuli, 1- and 2-dimensional, statistically robust solutions could be derived, therefore permitting the unfolding of a second perceptually relevant factor. This was sufficient for the purpose of this study, i.e. to either approve or disapprove the envisaged unidimensionality of the sound excerpts.

The decision of metric vs. non-metric analysis has to be based on the quality level of the input measures of similarity. As the listeners had judged the absolute magnitude of the similarities between the stimuli, the data can be assumed to be at the ratio level of measurement. However, Anderberg [1973] recommended against declaring similarity data to be of quantitative scale quality. He suggested that in a full array of paired comparisons subjects might not be able to make very precise judgements as to the amount of difference between each pair. In this context, Borg and Groenen [1997] pointed out that solutions from ratio and ordinal MDS analyses are almost always very similar in practice, the positions of the stimuli in ordinal MDS being practically just as unique as they are in ratio MDS, unless very few stimuli are used. Furthermore, they emphasised that “a weak scale level makes it easier to approximately represent the essential information in an MDS space of low dimensionality”. This is because approximate representations smooth out the errors (or noise) contained in similarity judgements (e.g. due to measurement imprecision or unreliability), which is why they are more robust, reliable, replicable and substantively meaningful than those that are formally perfect.
On these grounds, non-metric MDS analyses were carried out. To assess dimensionality the 'measures of fit' calculated by the MDS procedure were examined. Measures of fit are non-statistical parameters, which express how well a given model represents a set of raw data. In the case of the 'alternating least-squares algorithm' (ALSCAL) implemented in SPSS [Norusis, 1994] these are 's-stress' and 'RSQ'. S-stress ranges from 1 (worst possible fit) to 0 (perfect fit). RSQ, the squared correlation index, can be interpreted as the proportion of variance accounted for by the MDS model [Hair et al., 1998]. Although it is desirable to maximise this parameter for a given solution, the maximal number of dimensions taken into account needs to be limited, especially if the increase in explained variance per dimension is less than ~0.05 [Martens & Zacharov, 2000]. This is because dimensions with a low contribution to the explained variance are difficult to explain and are likely to be associated with noisy data.

In Figure 5.6 a so-called 'scree plot' is shown, displaying s-stress as a function of dimensionality. For the sake of completeness, a second 'badness-of-fit' parameter, 'stress', has also been included, which differs from s-stress in that it is reported in terms of linear rather than squared distances. As can be seen, both s-stress and stress decrease monotonically with each additional dimension. However, this comes as no surprise, since, as a matter of fact, both parameters will always get smaller if dimensionality is increased – even if the conditions for a stable analysis are not satisfied.

![Scree plot derived from non-metric MDS analysis (source distance)](image)

Borg and Groenen [1997] also pointed out that a non-metric MDS model can yield a degenerate solution, i.e. a configuration that is almost perfect in terms of the measure of fit criteria, even though the MDS distances do not represent the data properly. Such a misleading result can be prevented by choosing a metric representation of the data. Therefore, in order to make sure that degeneracy was not a problem, all sets of similarity judgements collected for this thesis were submitted to metric MDS analyses as well. Since only minor differences between each pair of solutions were found, it was decided to go ahead with the non-metric representation for the reasons outlined above.
In spite of this problem, the scree plot is commonly inspected to see whether a point is apparent beyond which the decrements in the chosen badness-of-fit metric begin to be less pronounced. Several researchers have argued that such a "knee" corresponds to the dimensionality that should be chosen (e.g. [Schiffman et al., 1981; Martens & Zacharov, 2000]). The reasoning for this choice is that the knee marks the point where MDS uses additional dimensions to essentially only scale the noise in the data after having succeeded in representing the systematic structure in the given dimensionality [Borg & Groenen, 1997]. Yet, this statement is believed to be oversimplified, because an apparent knee at 2-D cannot rule out unidimensionality of a set of stimuli since a real knee located at 1-D would not be identifiable as such. Hence, to be precise, an apparent knee at the second dimension on its own is unable to resolve whether a data set contains one or two discrete perceptual characteristics. To the best of this author's knowledge this has not been highlighted in the relevant literature so far.

It follows that for assessing the dimensionality of a solution the (s-)stress measure can be problematic. Therefore, it is also unclear which model should be chosen for representing the perceptual structure of the SD stimuli. Fortunately, however, a decisive clue can be obtained by evaluating the scree plot results in parallel with RSQ. Table 5.1 shows the RSQ values computed by the 1-D and 2-D MDS models. As can be seen, the 1-D solution is characterised by a high RSQ value that increases by only a small amount (i.e. less than 0.05) in the case of the 2-D solution. This indicates that a strong first dimension exists in the similarity data.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80</td>
</tr>
<tr>
<td>2</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Hence, by conjointly examining the two measures of fit, a firm indication is given that the panel identified and employed a single perceptual factor when comparing the sounds. The trustworthiness of the MDS model's representation with regard to the psychological structure of the stimuli is also reflected in the 'stimulus space' (Figure 5.7). This plot is the result of aggregating the subjects' dissimilarity judgements and depicting them graphically as the 'psychological distances' between the stimuli. Those stimuli that the subjects rated to be similar appear as points close to each other whereas those stimuli judged to be dissimilar are distant from one another.
Knowing that 'a' was synthesised to appear to be the closest and 'i' the most distant stimulus, it is evident that the sounds were perceived in the order intended. The spacings between each pair of stimuli are not constant, which might be due to inaccuracies during the generation stage and/or the inability of the subjects to be consistent in their judgements. However, all sounds appear to have a different intensity with regard to a particular perceptual factor, thereby enabling the listeners to rank them correctly.

Thus, on the whole, it can be concluded that the MDS analysis managed to successfully uncover the envisaged perceptual organisation of the SD stimuli from the panel's responses.

Analysis of verbal data

Seeking confirmation for the apparently unidimensional structure of the distance samples the verbal responses were examined. As pointed out in Chapter 3 and Section 5.3, the motivation for collecting additional verbal data was to establish a basis on which to identify the fundamental dimension(s) used for evaluating the stimuli. Apart from encouraging them to think of terms differentiating between stimuli and to note these terms in order of priority, the subjects had not been instructed to comply with any particular response format in order not to bias their responses. As a result of this free verbalisation approach, the data were fairly diverse and hence needed to be structured. For this purpose, VPA (see Section 3.2.1) was employed. At the first level of analysis the data were separated into two categories – one for holistic and one for analytical terms. The distinction was based on whether the subjects' responses were directly related to a perceptual phenomenon as a whole (i.e. high-level descriptors) or whether they described more specialised and/or technical aspects related to signal properties influencing the formation of a certain perception (i.e. low-level descriptors). Since the experimenter
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

has to interpret the meaning of the subjects' verbalised perceptions, there is a risk of biasing the outcome by condensing the data into fewer and fewer groups. By limiting the classification process to two stages, an attempt was made to restrict distortion of the meanings of the responses as much as possible. To obtain an index of perceptual importance for each group of terms, a total weight factor was calculated. As the subjects had written down the perceived differences in order of priority, each term was weighted according to its position in its list. To illustrate, the first descriptor in a given list was assigned the value 1, the second 1/2, the third 1/3 and so forth. The values of all verbalisations within each group were then added up and the result was divided by the number of subjects that had taken part in the listening test, resulting in a maximum possible weight factor of 1.

With regard to the holistic terms, it was found that the first classification stage (i.e. holistic vs. analytical) structured the data in a logical manner. Five groups were identified, the details of which are shown in Table 5.2. The first group contains terms strongly related to SD perception such as "Distance of source", "Proximity to instrument", "Distance/perspective", "Perceived distance of sound source", "Proximity of instrument location in room", "Apparent closeness of sound image", "Perceived distance of guitar" or "Closeness of the sound source". This perceptual effect was noted by all nine participants, seven of which wrote it down in first place, resulting in a strong weight factor of 0.87.

<table>
<thead>
<tr>
<th>Holistic groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source distance</td>
<td>9</td>
<td>0.87</td>
</tr>
<tr>
<td>Image width</td>
<td>2</td>
<td>0.06</td>
</tr>
<tr>
<td>Room size</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>Presence</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>Source depth</td>
<td>1</td>
<td>0.06</td>
</tr>
</tbody>
</table>

References were also made to 'Image width' ("Focus of sound image" and "Frontal width"), 'Room size' ("Perceived size of recreated room ambience"), 'Presence' ("Sense of space") and 'Source depth' ("Depth of bass notes"). Yet, for none of them was the agreement as uniform between the subjects as for source range, which is why their perceptual weights are a fraction of the one obtained for the 'Source distance' bin.

---

21 This was also the case with respect to analysing the verbal data from all the other validation studies carried out for this thesis.
The results from the classification of the analytical terms are depicted in Table 5.3. Verbal descriptors related to 'Spectral changes' were stated most often. More precisely, four out of the seven terms in this category address variations in HF content (e.g. “Loss of HF”, “Sharpness” or “Brightness”), two LF content (“Loss of LF” and “Relative volume of bass line to upper parts”), whilst the remaining one does not specify any particular frequency region. The relative weight factors of these three subgroups are 0.14, 0.04 and 0.06, denoting that changes in the HF part of the frequency spectrum were most detectable. This may be the result of simulating air absorption effects. Moreover, three listeners made specific references to the D/R, which of course is a parameter that was specifically varied as part of the SD simulation. The ‘Definition’ category is comprised of the terms “Distinctiveness of bass notes”, “Loss of attack”, “Clarity of attacks” and “Attack of notes”. These terms can probably be explained by the fact that a lower ratio of the early to the reverberant sound energy tends to decrease the degree of clarity or definition of both simultaneously and successively played sounds [Beranek, 1996]. The last group contains terms that were only brought up once and that do not fit into any of the other bins. It includes “Perceived loudness of guitar” (which can be considered a sub-attribute of the SD percept (see Section 0.4)), “Width of reverb”, “Ringing sounds in rear speakers” and “Delay of sounds coming from the rear relative to the front”. The last two descriptors are somewhat peculiar since LS and RS were not used for this simulation. In any case, the weight factors of the analytical groups are generally much lower compared to the one of the holistic ‘Source distance’ category, implying that they were perceptually not important.

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral changes</td>
<td>7</td>
<td>0.24</td>
</tr>
<tr>
<td>D/R</td>
<td>3</td>
<td>0.22</td>
</tr>
<tr>
<td>Definition</td>
<td>4</td>
<td>0.17</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Finally, it is worth pointing out that as part of their education all listeners had received thorough training in analytical listening skills. Notwithstanding, seven of the nine panellists expressed their perception of distance changes in the form of a holistic descriptor first before resorting to analytical terms. This can be interpreted as strong evidence for the successful unidimensional simulation of the perceptual phenomenon of source distance.

INDSCAL analysis
To gather further evidence for the non-existence of a second meaningful dimension the data were submitted to an INDSCAL analysis. INDSCAL can be considered a derivative of MDS, because it also calculates the co-ordinates of a group of stimuli on a number of perceptual dimensions common to a set of similarity judgements. The result is then displayed in the stimulus space (e.g. see Figure
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

5.7). However, in contrast to MDS, INDSCAL acknowledges that subjects may differ in how they form their verdicts and therefore tries to take such individual differences into account. More precisely, it models inter-subject agreement as well as disagreement, separating those factors common to a group of subjects from those in which the subjects differ [Borg & Groenen, 1997]. That is why INDSCAL can provide a quantitative characterisation of the individual differences that exist within a panel, which are captured as subject-specific weights placed upon each of the INDSCAL dimensions [Martens & Zacharov, 2000]. These weights are commonly portrayed in the ‘subject space’. Mathematically speaking, INDSCAL is very similar to MDS, the distance between stimuli \(i\) and \(j\) for subject \(n\) being defined as follows:

\[
d_{ijn} = \left[ \sum_{r=1}^{R} \frac{w_{nr}(x_{ir} - x_{jr})^2}{\sum_{r=1}^{R} w_{nr}} \right]^{1/2}
\]

where \(x_{ir}\) is the co-ordinate of stimulus \(i\) on dimension \(r\) and \(w_{nr}\) is the weight (required to be non-negative) for dimension \(r\) associated with subject \(n\).

In Figure 5.8 the 2-dimensional subject space of a non-metric INDSCAL analysis executed on the listeners' dissimilarity judgements of the SD stimuli is displayed. In interpreting this (or any other) subject space it is important to note that the origin of this space is not arbitrary, but has a fixed meaning [Carroll & Chang, 1970]. The distance of a subject from the origin corresponds, at least roughly, to the variance accounted for in the data from that subject. This means that if a subject's point is precisely at the origin, no variance at all is accounted for. The direction of a subject from the origin relates to the pattern contained in the data from that subject. Therefore, two subjects who lie on the same straight line issuing from the origin would have identical configurations except for a single overall scale factor. One subject's being closer to the origin on that line would indicate simply that less of the variance in his/her data is accounted for by that common configuration, either because his/her data are noisier or because additional dimensions are needed to explain the subject's judgements fully.

Inspecting the SD subject space, one can see that the weight placed upon Dimension 1 is greater for all nine listeners, denoting that it was perceptually more important to them than Dimension 2. Thus, this provides confirmation for the previous finding that the participants detected and used the same (main) difference for assessing the sound excerpts in terms of their similarity. As to the second dimension, the previous analyses gave the impression that no meaningful, additional perceptual factor is contained in the data. To ascertain whether this is true or not, one can look for correlations between the listeners' verbal responses and their positions along Dimension 2. Basically, the verbal data are mapped onto the subject space to see if any inter-listener agreement exists.
Fortunately, this is not the case. Subjects 7 and 9, for instance, verbalised very similar effects when asked to specify which differences they had detected between the stimuli, i.e. “Perceived distance of sound source” and “Level of reverb with respect to direct sound” (Subject 7) and “Perceived distance of guitar”, “Ratio of reverb to dry guitar sound” and “Perceived loudness of guitar” (Subject 9). Despite the large common ground between these verbalisations, the two listeners are far apart relative to the area occupied by the panel as a whole. This is especially the case for their positions along the second dimension. Counterintuitively, it is Subject 7 who is higher up along Dimension 2 even though he noticed one effect less than Subject 9. Thus, it is reasonable to assume that his larger weight for the second dimension does not point to another distinctly perceivable factor, which would be extremely difficult to account for anyway, because of the close relatedness of his two chosen descriptors. These tendencies are also apparent for other listeners, e.g. Subject 6 came up with seven perceptual effects, whereas Subjects 2, 3, 4 and 7 did not write down more than three differences each. Since most of these terms are either directly related to SD perception or at the least reflect the applied processing, one can deduce from this analysis that Dimension 2 has no collective meaning.

Therefore, interpretability is made the final criterion for evaluating dimensionality. It has been proposed elsewhere that dimensions, which cannot be interpreted, probably do not exist [Schiffman et al., 1981]. While it cannot be claimed that no further perceptual factors exist on an intra-listener level for the SD stimuli, there are plenty of indications that the panel as a whole did not identify a second perceptual dimension. Hence, it can be concluded that unidimensionality of the SD simulation was achieved, which is why the stimuli should be suitable for training purposes.
5.6 Simulation of source width

Having successfully simulated source distance, it was decided to tackle source width next. As the summary of relevant psychoacoustic research in Section 2.5 showed, lateral reflections with delay times of up to 80ms relative to the DS have been associated with the occurrence of a source broadening. However, this temporal cut-off point may well be inappropriate as there is also evidence that SW perception is only affected if these reflections arrive at a receiver whilst the DS is (still) audible. In contrast, if they arrive during purely reverberant sound segments, the width of the environment will be perceived to vary. To make matters worse, changes in the properties of both the source signal and the reflections have been found to have a direct impact on the perceptibility of these effects. Hence, as already argued in Section 2.5.1, manipulating SW perception in an unequivocal manner appears to be a complex venture. The procedure devised for producing a set of SW samples is outlined in the next section.

5.6.1 Creation of source width stimuli

Since real-time control over discrete physical parameters had proven to be very valuable when simulating the attribute of source distance, it was decided to approach the creation of the width sounds in a similar manner. Thus, the Lexicon reverberator and the Sony mixing console were used again for emulating perceptually salient sound field characteristics. Informal experimentation with the directional and temporal properties of synthesised reflections confirmed that up to a certain level reverberant energy arriving within ~80ms of the DS from lateral directions generally caused a spreading of an auditory event. Yet, when the level of the reflections exceeded a certain value other perceptual characteristics (both spatial and non-spatial) changed simultaneously, thereby prohibiting the envisaged unidimensionality. What is more, the magnitude of the width effect was certainly not large enough to allow the generation of nine stimuli, each exemplifying a different intensity in terms of that attribute. Consequently, it was concluded that the creation of SW stimuli cannot be achieved solely through manipulation of early lateral reflections, thus necessitating the search for another method.

In the hope of achieving extended control over the breadth of a sound source, established width-processing techniques (see Chapter 4) were scrutinised. However, initial attempts were disappointing since the magnitude of the accomplishable width changes was still too small. For instance, when making use of the mixing desk’s built-in divergence control, a spatially very confined image could easily be produced by sending a monophonic input signal to the centre channel. Conversely, the maximum source broadening achievable by means of divergence was insufficient to be able to create seven other stimuli lying in-between the two extremes of the width scale. The opposite was the case with regard to stereo shuffling, i.e. whilst 2-channel programme material could be made to sound very wide, restricting it to a narrow point source proved to be impossible (phantom mono being too defocused). The attempt to take only one channel of such a stereo recording and route it to the centre speaker so as to produce the spatially confined source striven for turned out to be problematic, too.
This was because the resultant sound was timbrally very different to those stimuli constructed using both channels of the recording (i.e. those relying on phantom images). Besides, if reflections were contained in the single channel, they appeared to cluster around the source. Not only did this sound unnatural, but spatially different as well (i.e. the source appeared to be farther away and the width of the reproduced environment was reduced) compared to those stimuli for which both channels of the programme material were needed to produce a source having ample width. Due to these deficiencies cross-fading between a narrow and a wide source resulted in unwanted perceptual artefacts occurring in parallel. These findings for artificially processed sounds were more or less the same for those recordings made with different microphone techniques, including X-Y, M-S and A-B (e.g. see [Rumsey, 2001] for details).

Based on the informal experimentation described above, two conclusions were drawn:

1. To enable the creation of a well-defined, highly localised as well as a wide, diffuse sound source, which would be adequately different from each other to allow seven stimuli to be placed in-between, 3-channel programme material was needed.

2. All SW stimuli would have to be generated using L, C and R to be able to keep any unwanted spatial and timbral differences between the sounds to a minimum.

Although it should have been possible to produce suitable programme material with the help of 3-channel recording techniques, a viable alternative was to synthesise a centre channel from existent 2-channel source material. Various matrix-based methods have been developed for such upmixing purposes (e.g. see [Rumsey, 2001]). Arguably, the best results can be achieved with a technique known as *Trifield* that was proposed by Gerzon and others as a means of deriving a centre channel. In essence, the *Trifield* technique is a psychoacoustically optimised 3-channel panpot that enables a centre channel to be derived from and integrated into a (2-channel) sound image in a coherent fashion. In [Gerzon, 1992d] a detailed insight into the functionality of this algorithm was provided, including a description of the frequency-dependent processing that takes place to better replicate the properties of human hearing at different frequencies. In this context, Gerzon also stressed the importance that such a decoder be energy-preserving, as this leads to far less audible colouration over a wide variety of recording techniques (both time- and intensity-based) compared to decoders with marked variations in energy gain for different components of the input signal. Moreover, an energy-preserving decoder maintains the widest stereo images, whereas other systems cause a reduction in image width.

For this study, the *Trifield* algorithm implemented in a *Meridian 565 Digital Surround Processor* was employed. The algorithm was fed with programme material that had been created based on the anechoic monophonic sound excerpts used for the creation of the distance stimuli, i.e. the acoustic guitar and cornet. These were spatialised by simulating an A-B recording (with a microphone spacing of 35cm and a microphone-to-source separation of 1m) in the auralisation software *CATT-Acoustic*. Care was taken to ensure that the resultant (2-channel) programme material contained only small
amounts of reverberant energy so as not to mask any subsequent processing. For that reason, the room modelled in CATT-Acoustic had an average RT of less than 1s. This auralisation approach was preferred to making real recordings, because it allows one to trace back the reflection patterns picked up by the microphones and it was felt that this might be beneficial at a later stage of this work. A spaced microphone technique was chosen because of the more diffuse sound images such recording set-ups tend to create compared to coincident ones, for example [Eargle, 2001]. With the help of a synthesised centre channel sufficiently large differences in source width were hoped to be feasible.

In addition to the processing outlined above, it was decided to use the reverberation unit (combined with the mixing console) to construct an apparent space around the sound source. The reasoning behind modelling an acoustic enclosure was to increase the ecological validity of the SW simulation. After all, in a typical sensory evaluation situation subjects are not very likely to assess the spatial properties of reflection-free stimuli. Moreover, in/decreasing the levels of early lateral reflections had previously been found to broaden/narrow a sound source slightly (see above), which is why the space modelling was expected to offer supplementary control over the simulation of source width.

To emulate a space, the following generic impulse response was designed with the help of the reverberator:

1. Direct sound: $t = 0\text{ms}$
2. Early reflections: $40\text{ms} < t < 70\text{ms}$
3. Reverberation tail: $t > 60\text{ms}$

The reverberator was also fed with the A-B recording. Similar to the SD simulation, a single ER pattern was used for the creation of all SW stimuli, which consisted of six reflections having the same level and spectral content (a first-order low-pass filter with a cut-off frequency of ~4.5kHz being used again). The time window of 40ms to 70ms turned out to be adequate for making small adjustments in perceived source width, and since the A-B recording also contained some reflected sound, the relatively large initial time gap of 40ms was not noticeable. The six reflections were divided into two groups of three and amplitude-panned between $L$ and $LS$ or $R$ and $RS$, respectively. The two channels of decorrelated reverberation had a length of 1s and were routed to $L$ and $R$. A pre-delay of 60ms was employed to restrict interference with the early reflection set. In order not to affect room-related spatial percepts, no alterations were made with respect to the diffuse reverb. Figure 5.9 depicts the generic impulse response of the reverberator.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

The perceptual effect of changes in the width of a source (positioned directly in front of the listener) was produced by (in order of importance):

- Spreading the synthesised centre channel across L, C and R using divergence,
- Changing the relative level between the L/R and C outputs of the Trifield by ±1.5dB and
- Varying the level of the early lateral reflections by up to 1.5dB.

The ‘tuning’ (i.e. the conjoint adjustment of all variables so as to achieve unidimensional variation) of the SW simulation was done by ear. In this context, it has to be emphasised that the applied processing (which admittedly was even more artificial than that used for the SD simulation) had to be pushed to its limits in order to be able to generate nine SW stimuli, and yet the achieved differences were still small. Again, an attempt was made to attain linear spacings between the sound excerpts. However, this undertaking was complicated due to the difficulty of determining the left and right boundaries of especially the wider, de-focused sound sources. A block diagram of the algorithm deployed for the SW simulation is shown in Figure 5.10.
The entire processing took place in the digital domain and the stimuli were recorded to hard disk as 5-channel .aiff files with a sampling rate of 44.1kHz and a resolution of 16 bit.

5.7 Validation of source width stimuli I
To determine if unidimensionality of the SW simulation had been achieved another validation experiment was performed, which was identical to the SD test in terms of the physical set-up and the experimental design. Five of the nine listeners from the previous study served as the ‘measurement instruments’ for this study. Again, only the guitar stimuli were empirically scrutinised.

5.7.1 Results
Since the experimental procedure had not been modified, both the similarity and verbal data were treated and analysed in the same way as before. The results are presented below.

Analysis of similarity data
Figure 5.11 shows the behaviour of s-stress and stress as a function of dimensionality. The first thing to note is that both measures are significantly higher compared to the results of the SD experiment (see Figure 5.6), meaning that there is much poorer correlation between the subjects’ verdicts and the MDS solutions. Although the (s-)stress metric is affected by several experimental parameters such as the number of stimuli, the dimensionality of the model or the level of measurement, “an MDS solution may have high stress simply as a consequence of high error in data” [Borg & Groenen, 1997]. In consideration of the fact that the experimental procedure was identical for the SD and SW tests, noisy subject responses seem likely. Also, the lack of a knee in the scree plot implies that there are no prominent dimensions in the similarity data, thereby lending support to this view.
Confirmation for the bad fit between the similarity scores and the fitted model is available from the RSQ results (see Table 5.4). Whilst the 1-D MDS solution of the SD experiment had been found to feature a high RSQ value of 0.80 (see Table 5.1), the corresponding SW model can only account for 35% of the data variance. Even more striking is the lack of an increase in RSQ when a 2-D analysis is applied. Thus, it appears that extending the MDS model by an extra dimension cannot help unfold a pattern contained in the subjects’ responses either. This could mean that the listeners evaluated different perceptual factors and/or that the MDS algorithm was not able to identify common perceptual dimensions due to the data set being noisy.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>0.35</td>
</tr>
</tbody>
</table>

The poor fit can be traced back to the transfer of the width stimuli. When the sound excerpts were played back in the listening room, a lot of the width cues introduced into the sounds in Studio 3 disappeared. This is very bewildering if one takes into account the fact that Studio 3 and the listening room are very similar in terms of their acoustical properties. Indeed, the listening conditions in the listening room are even more critical than those in Studio 3 (see Section 5.4.1). This indicates that while the distance stimuli seem not affected by the different acoustical and playback conditions of the listening room, the SW sounds do not ‘travel’ easily from one listening environment to another. Feedback given by the subjects confirmed these findings. They reported that whilst they had heard clear changes between the SD samples, the differences between the width stimuli were very small.

The lack of obvious width variations is also evident from the associated 1-dimensional stimulus space (Figure 5.12). The stimuli basically cluster into two groups (‘a’ being the narrowest and ‘i’ the widest stimulus). This is likely to be the result of the subjects judging the sounds in each group to be highly similar or even the same. Also, stimuli ‘f’ and ‘g’ are shown in reversed order, which presumably is due to the listeners’ inability to distinguish reliably between sounds close to each other along the width scale.
Overall then, it has to be accepted that the MDS algorithm did not discover the envisaged psychological organisation of the SW stimuli in the similarity judgements. Moreover, it could not make any more sense of the data by means of a 2-D model either. This strongly suggests that they do not contain a systematic rating pattern. Intending to gain an insight into the audible effects of the created sound excerpts when reproduced in the listening room, the verbal responses were examined next.

Analysis of verbal data

The holistic categories that were obtained from structuring the verbal responses are summarised in Table 5.5. As can be seen, there is no single strong group of holistic terms, which is in line with the findings of the MDS analysis. Interestingly, though, two references were made regarding changes along the lateral plane, i.e. ‘Image shift’ (“Slight image shift”) and ‘Source width’ (“Slight variations in source width”). However, neither is there agreement between the subjects with respect to either of these two groups, nor do they have large weight factors. This implies that their perceptual relevance is low, thus corroborating that the differences between the width stimuli were negligible when auditioned in the listening room. It is also worth noting that two listeners detected changes in the range of the sound source (“Perceived distance of sound source”, “Presence: some sounds appear closer”). Other researchers [Martin et al., 1999; Berg & Rumsey, 2001] have highlighted the difficulty associated with defining the SW attribute, there being a risk of confusion between narrower source width, increased source distance and less spread of LF content. This might be an explanation for the two references to source proximity.
Table 5.5: Groups of holistic terms and their relative weights (source width, listening room)

<table>
<thead>
<tr>
<th>Holistic groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image shift</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>Source distance</td>
<td>2</td>
<td>0.13</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Envelopment</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

The groups of analytical terms are shown in Table 5.6. It has to be pointed out that the only meaningful subgroup that could be identified is 'Spectral changes'. Overall, this category has a large weight factor, but this is due to the aggregation of all descriptors referring to changes anywhere in the (audible) frequency spectrum. In particular, three references were made to HF content ("Brightness", "Sparkle of top notes" and "Slightly less HF in some extracts"), two to LF content ("Lightness" and "Fullness of sound") and another one to "Thinness (mid frequencies)". The relative weight factors of these three subgroups are 0.25, 0.25 and 0.20. On the whole, the range of the perceptual weights for both the holistic and analytical groups is therefore small, 0.05 being the lowest and 0.25 the highest value. Hence, it seems plausible that because of the lack of clearly discernible differences between the stimuli, each listener may have focused on specific details of the reproduced sound fields, thereby giving rise to the diversity of terms that is apparent. Thus, the indication given by the MDS results that there is hardly any inter-listener concordance is also evident from the verbal responses, which is why it can be inferred that the collected data are noisy and therefore meaningless.

Table 5.6: Groups of analytical terms and their relative weights (source width, listening room)

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral changes</td>
<td>6</td>
<td>0.70</td>
</tr>
<tr>
<td>Clarity of sound</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>Early reflection time</td>
<td>1</td>
<td>0.20</td>
</tr>
<tr>
<td>Slight phase shift</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>Differences in overall sound</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>Length of reverb tail</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>EQ of reverb tail</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Some instrument resonances stronger in certain extracts</td>
<td>1</td>
<td>0.07</td>
</tr>
</tbody>
</table>

INDSCAL analysis
Even though the previous analyses had revealed the uselessness of the collected data for deriving a meaningful sensory profile of the SW simulation, the similarity scores were submitted to an
INDSCAL analysis. The resultant 2-D subject space (Figure 5.13) also documents the absence of a rating pattern and hence the lack of clearly identifiable stimulus attributes. In comparison with the SD subject space (see Figure 5.8), the subjects’ dimension weights are generally low, Subject 4 featuring the largest one (i.e. ~0.65 on Dimension 1). Incidentally, he was the one to perceive “Slight variations in source width”, albeit not as the dominant stimulus difference. Subject 5, who noticed the “Slight image shift”, seems to be particularly poorly represented by the 2-D INDSCAL model as evident from his closeness to the origin of the subject space. Whilst this could be due to the employed model having too low a dimensionality, his verbal responses (“Slight image shift”, “Slight phase shift perhaps?”, “Some instrument resonances seemed stronger in certain extracts” and “Slightly less HF in some extracts”) indicate that he could not detect any strong perceptual variations, hence producing noisy data.

Apart from these few observations, however, no other new insights can be gained from evaluating the subject space. Therefore, as the results of the validation experiment do not allow the auditory effect(s) of the width stimuli to be determined, it was decided to repeat it in Studio 3 where the sounds had been generated.

5.8 Validation of source width stimuli II
In order for the results of the second validation study to be directly comparable to the ones from the first test, it was ensured that the experimental conditions were as similar as possible. Unfortunately, it was not feasible to position the computer monitor in front of the subjects because of the mixing console installed in Studio 3. Also, due to the loudspeakers being larger than the ones in the listening
room and the different physical layout, only the three loudspeakers in front of the listening position could be concealed with the help of the acoustically transparent curtain. A diagram of the physical set-up is given in Figure 5.14.

![Diagram of experimental set-up (Studio 3). The loudspeakers were positioned at 0°, ±30° and ±110°.](image)

The mixing desk was configured in such a way that the channels in use were not displayed and all level meters were covered to make sure that no visual cues were available to the subjects. When aligning the loudspeaker levels, care was taken that the absolute SPL at the listening position matched the one measured in the listening room. All five listeners who had participated in the first validation experiment also completed the test in Studio 3. Endeavouring to avoid the occurrence of any confounding perceptual artefacts due to the non-optimal location of the computer monitor, they were instructed to restrict their head movements as much as possible.

### 5.8.1 Results

#### Analysis of similarity data

The scree plot derived from the MDS analysis of the similarity data is shown in Figure 5.15. Compared to Figure 5.11, (s-)stress is much lower this time, which denotes a better match between the new similarity ratings and the corresponding MDS solution. What is more, a knee is apparent in the two curves at 2-D. Even though this cannot rule out a sharper but invisible knee at 1-D (see Section 5.5.2), it indicates at least that the MDS algorithm managed to find a structure in the listeners' similarity judgements collected in Studio 3. It is true that there is also a slight kink in the stress curve at 4-D, but since SPSS optimises s-stress and not stress [Norusis, 1994] evaluation of the former index should be more reliable.
As regards the results for RSQ (see Table 5.7), the value for the 1-D solution obtained this time (0.62) is much higher than the one from the first test (0.35). Whilst it is clearly less than the one for source distance (0.80), it is beyond the 0.60 mark and can therefore be considered acceptable [Hair et al., 1998]. Moreover, the increase in explained variance due to the addition of a second dimension is minimal again (i.e. less than 0.05), which suggests that only one major dimension is present in the width sounds.

Table 5.7: RSQ results from non-metric MDS analysis (source width, Studio 3)

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Inspection of the 1-D stimulus space (Figure 5.16) reveals that the stimuli do not cluster as much into two groups as before (see Figure 5.12). Again, two stimuli (‘g’ and ‘h’) appear in reversed order, but this is comprehensible if one acknowledges that the differences between adjacent stimuli are very small (see Section 5.6.1).
In summary, the outcome of the second MDS analysis looks much more promising than the previous one with respect to the desired unidimensionality of the SW simulation. Admittedly, the results are not as convincing as the ones for source distance, but an improvement in terms of the stimuli’s envisaged perceptual structure is clearly visible. Hoping that this would also be the case for the verbal responses, a VPA was carried out on the semantic data.

Analysis of verbal data
Table 5.8 displays the outcome of scouring the verbal responses for holistic descriptors. At first glance, the results of the classification process appear to contradict the explanation found for the MDS solution. Two holistic categories with almost the same weight factor have emerged from the analysis, implying that two major dimensions exist, i.e. ‘Image width’ and ‘Source position’. Regarding the former, five verbalisations pertinent to this group were obtained (“Width of instrument”, “Central focus”, “Width of stereo image”, “Width of source” and “Image width”), which compares favourably with only one such mentioning after the first listening test (see Table 5.5). Undisputedly, for some of these responses it is not clear whether they describe a source- or an environment-related auditory aspect (hence the ‘Image’ prefix for this group), thereby highlighting a deficiency of verbal language. As to the second category, three listeners perceived changes in the position of the sound source. However, similar to the terms in the ‘Image width’ bin, there may be differences regarding the exact spatial variation that these verbalisations express. One listener specified that the positional changes took place in the lateral plane (“L-R position of source”), whereas the other two subjects did not indicate whether the source appeared to vary along the left-right axis and/or the front-back axis (“Stereo position” and “Positioning of starting notes”). Thus, it seems that in order to make sense out of such descriptions, additional information is needed that can help resolve these kinds of ambiguities.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

Table 5.8: Groups of holistic terms and their relative weights (source width, Studio 3)

<table>
<thead>
<tr>
<th>Holistic groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image width</td>
<td>5</td>
<td>0.67</td>
</tr>
<tr>
<td>Source position</td>
<td>3</td>
<td>0.60</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>Source depth</td>
<td>1</td>
<td>0.04</td>
</tr>
</tbody>
</table>

The remaining two holistic groups (‘Source distance’ and ‘Source depth’) are each due to a single mentioning and consequently have significantly lower weights than ‘Image width’ and ‘Source position’. Thus, while there is some common ground between the data sets obtained from the two SW validation studies, the subjects seem to be much more in agreement as far as the stimulus differences detected during the second listening test are concerned.

The structuring of the analytical terms produced the groups shown in Table 5.9. As can be seen, the total number of analytical descriptors is almost half that of the terms collected as part of the first experiment (see Table 5.6), resulting in fewer categories. Again, there is a group containing spectral descriptors, which comprises one reference to LF content (“Very slight loss of bass frequencies”), one to HF content (“Fullness of high notes compared to low notes”) and a third one not specifying any particular frequency range. The relative weight factors of these three subgroups are 0.10, 0.07 and 0.07, denoting that none of them was perceptually important. The same statement can be made with respect to the other four categories, which contain terms that were elicited only once. All in all, the range of the weight factors of the analytical groups is very small indeed, 0.05 being the lowest and 0.10 the highest value. Taking into account this finding and the ones made for the holistic groups, it can be deduced that this time the differences between the stimuli were far less ambiguous than they had been in the first SW validation test. Accordingly, a clear dominance of two holistic descriptors, a reduction in the total number of elicited terms and much stronger inter-assessor agreement can be certified.

Table 5.9: Groups of analytical terms and their relative weights (source width, Studio 3)

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral changes</td>
<td>3</td>
<td>0.24</td>
</tr>
<tr>
<td>Front-back first echo time</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>Reverb time</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Phase shift</td>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>Level</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>
INDSCAL analysis

When analysing the subject space for the new similarity data (Figure 5.17), it appears that the listeners are better represented by the INDSCAL model this time (most notably so Subject 5). This had been evident already from the much higher RSQ value for the second MDS analysis (see Table 5.7). Only Subject 1 has not improved in terms of the explained variance, his weight for Dimension 1 having decreased slightly. In general, however, a clearer rating pattern in the similarity scores can be attested. Besides, three of the five panellists (i.e. Subjects 3, 4 and 5) appear in close proximity, which indicates that they perceived the stimuli in almost the same way. The terms that these listeners wrote down first are all related to the concepts of ‘Image width’ and ‘Source position’ (i.e. “Stereo position”, “L-R position of source” and “Image width”). Similar descriptors were elicited from Subject 1 (“Central focus”) and Subject 2 (“Positioning of starting notes”) who, like the other three subjects, lie on virtually the same straight line from the origin. Yet, their weights for the first INDSCAL dimension are not larger than the ones for Dimension 2. Thus, not much can be said regarding the meaning of the two INDSCAL dimensions.

Overall then, no clear picture emerges regarding the dimensionality of the SW stimuli. The fact that the RSQ results hint at a reasonably strong single dimension whereas the subject space and the verbal data indicate that two dimensions exist leaves unanswered the question of how to interpret the results. A possible explanation might be that the subjects mistook the changes in source width for changes in its position. This could be explained by the more complex nature of the notion of SW. In everyday life, humans make sense of their environment by constantly evaluating the position of the sources distributed around them. Hence, source position is a very familiar concept, and therefore listeners can easily relate to it. Conversely, source width is a more unfamiliar perceptual construct that is likely to
be less meaningful to listeners. It is probably a fair assertion to make that most humans hardly ever consciously assess the width of a single (real) source based on auditory information only because of the dominance of visual cues. It is also possible that due to the lateral positioning of the computer monitor the subjects' head movements might have induced an apparent shift in the position of the source within the stereo field. Although an attempt had been made to reduce the influence of this variable onto the subjects' verdicts as much as possible, the likelihood that head movements could be eliminated completely is small.

In light of the obvious uncertainties, it was decided to conduct a confirmation experiment to help determine the sensory properties of the SW samples, the details of which are presented in the next section.

5.8.2 Confirmation experiment

The confirmation experiment consisted of two parts. For the first part, a classification approach was adopted. In particular, subjects were asked to make paired comparisons of a selected number of width stimuli. Next, they had to choose a spatial attribute out of a given list that could be used to describe the dominant difference they detected for each stimulus dyad. Both acoustic guitar and cornet stimuli were included in this test to see whether they might lead to different results. Yet, to limit the duration of the listening test only four sounds per source material were used, resulting in eight pairwise comparisons per listener. Stimuli from the extreme ends of the SW scale were picked, as they were most suitable for evoking the stimuli's predominant effect in the listeners. The spatial attributes offered to the subjects were based on the holistic categories obtained as part of the second SW validation study (see Table 5.8). In order to restrict the number of attributes to a reasonable minimum, only three types were chosen, i.e. 'Source width', 'Source position' and 'Source distance'. It is acknowledged that in view of the unclear width verbalisations (see above) the choice of 'Source width' may seem controversial. Yet, informal feedback given by the listeners after the completion of the second validation test had indicated that they were not referring to the width of the overall sound scene, which is why the 'Source width' label was chosen for this experiment. In addition, an 'Other' category was included to give subjects the option to express a different sensation, should they find none of the three provided attributes appropriate for successfully describing the dominant difference they might perceive. Selection of this category caused a pop-up menu to appear, asking the listener to specify his perception by entering a short description. Hence, 'no difference' verdicts were permitted, too. To avoid misunderstandings, the participants were provided with the following attribute definitions:

1. **Source width**: Does the width of the sound source appear to change, e.g. does one of the two sounds resemble a narrow, well-defined, focused point source whereas the other is wider, ill-defined, more diffuse?
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

2. **Source position**: Does the apparent location of the sound source appear to change, e.g. do you perceive a shift of the source from a central position towards the left or right or vice versa?

3. **Source distance**: Does the perceived range between you, the listener, and the sound source appear to in/decrease, i.e. does one of the two sources seem closer to/farther away from you than the other?

For the second part of this experiment all listeners had to rank a number of sounds in terms of perceived source width. The objective of this ranking exercise was to verify whether or not the listeners found this concept to be meaningful if instructed to apply it to the judgement of a number of SW stimuli. For that reason, the definition of source width employed for the classification test was presented to the subjects again. As the sounds had proven to be difficult to discern when all nine stimuli were presented, only five guitar stimuli were included in this test (i.e. sounds ‘a’, ‘c’, ‘e’, ‘g’ and ‘i’). Consequently, the differences between adjacent stimuli were slightly bigger. The user-interface (Figure 5.18) comprised five sliders, which could be set to values ranging from 1 to 5, and listeners were told to assign each value only once.

The listening test took place in Studio 3 again. The physical set-up was identical to the one of the second SW validation experiment (see Figure 5.14), except that this time no effort was made to hide the frontal loudspeakers. As the previous listeners had been exposed to some of the chosen sound excerpts before, they were likely to be preconditioned with regard to the sounds’ auditory
characteristics. That is why four different listeners were asked to take part in this study. Two of them were first-year and the other two second-year ‘Tonmeister’ students. All listeners were instructed to face forward when switching between the sounds.

Analysis of classification data
The stimulus pairs presented to the subjects during the classification test as well as the corresponding accumulated listener responses are shown in Table 5.10. Evidently, SW changes were perceived for more than half of all stimulus dyads, which is why this forms the strongest category. Notwithstanding, a significant proportion of all judgements was made in favour of source position. This attribute came up as the next strongest group, thereby reflecting the finding of the second SW validation test. Variations in source distance were perceived only a few times, the single ‘Other’ verdict being due to a listener who was unable to detect a difference between one pair of sounds. In general, these observations hold true for both sets of stimuli, i.e. no relationship between SW perceptibility and type of source material is apparent. Hence, while these results allow deducing that the stimuli’s sensory differences are generally not experienced as distance changes, confusion between source width and source position is still apparent. Therefore, even though source width turned out to be the dominant attribute for this test, the results cannot prove the unidimensionality of the variation in the sound excerpts. In fact, it is very likely that the inclusion of smaller differences between the stimuli would have spread out the responses more evenly.

Table 5.10: Accumulated attribute classification judgements from confirmation experiment (source width, Studio 3)

<table>
<thead>
<tr>
<th>Stimulus pair</th>
<th>Width</th>
<th>Position</th>
<th>Distance</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guitar a-h</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Guitar a-i</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Guitar b-h</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Guitar l-b</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cornet a-h</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cornet l-b</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Cornet l-a</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Cornet h-b</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

Analysis of ranking data
To evaluate the ‘degree of wrongness’ of each ranking sequence relative to the envisaged response, the square of the Euclidean distance (see Section 5.5.1) was applied. Hence, a ranking sequence that is ‘close’ to the correct response yields a small squared Euclidean distance (SED), whereas a sequence
that is ‘far away’ gives rise to a large value. In the case of this particular example, a response with a SED of 40 corresponds to the ‘worst case scenario’, i.e. the exact reversal of the correct sequence. A value of 2 implies that an adjacent pair of sounds has been inverted and an SED of 4 denotes the reversal of two pairs of sounds. The listeners’ ranking sequences and associated SEDs are displayed in Table 5.11.

Table 5.11: Ranking sequences and associated SEDs from confirmation experiment (source width, Studio 3)

<table>
<thead>
<tr>
<th>Listener</th>
<th>Rank sequence</th>
<th>SED (max. = 40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>e, a, c, g, i</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>a, c, e, i, g</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>c, a, e, g, i</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>c, a, e, g, i</td>
<td>2</td>
</tr>
</tbody>
</table>

As can be seen, three of the four participants came very close to the correct sequence, reversing only two adjacent sounds. This is encouraging because it indicates (1) that the subjects could apply the notion of source width to the set of stimuli and hence rank them almost correctly and (2) that the stimuli vary along a common perceptual dimension that was pre-defined to be source width. It is interesting to note that Listener 1, who performed worse than the others, was one of the two first-year ‘Tonmeister’ students that took part in this test. Perhaps he found it harder to detect and assess the SW changes because of his limited experience in critical listening exercises.

To summarise the SW validation findings, there is no doubt that the results obtained for the SW stimuli are not as persuasive as the ones for the SD simulation with regard to the intended unidimensionality – probably because of the relatively poor robustness of source width as a psychological construct. Nevertheless, the MDS solution implied unidimensional variation in the similarity ratings, and since there are strong indications that changes in source width dominated the listeners’ perceptions of the SW samples, the stimuli should still be adequate for training purposes.

5.9 Simulation of ensemble width

In contrast to SD and SW perception, the EW attribute has up to now been overlooked by psychoacousticians. Assumably, this is because of its close connection to source direction hearing, which has been extensively studied (see Section 2.1). Whilst one might argue that further research into the EW percept is therefore superfluous, it is at a different level in Rumsey’s scene-based paradigm (see Section 1.3) and hence a higher-level construct, which considers groups of sources. In order to be able to achieve unidimensionality in variation, a better understanding of its sensory properties is thus deemed necessary. That is why a preliminary study into the perception of ensemble width was conducted, which is to be outlined in the next section. This will be followed by a description of a
processing environment that was implemented so as to facilitate the modelling of ensemble-level spatial attributes. In this context, a new multichannel panning technique will be discussed and compared to a well-known pan law. Finally, the methodology employed for simulating the EW attribute shall be depicted.

5.9.1 Pilot study into the perception of ensemble width

As is evident from the research summarised in Chapter 1, the various elicitation studies have managed to break up the 'perceptual conglomerate' of spatial quality into multiple discrete, qualitative factors. Nevertheless, as their independence has yet to be fully verified (see Section 1.2), it cannot be ascertained if these attributes should be seen as some kind of standard and hence be used for subjective testing purposes or if more refinements are required first. Thinking about the concept of EW more carefully, it occurs that this descriptor needs further specification in order to describe the corresponding auditory effect completely. In particular, it is unclear whether it implies that sources spread out along a straight line perpendicular to the line between the listener and a central source or whether the sources move on an arc maintaining constant perceived distance to the listener. Figure 5.19 and Figure 5.20 illustrate the differences between the notions of 'linear EW' and 'constant distance EW' graphically. Although perhaps slightly pedantic, such a precise distinction is believed to be a valuable step towards an unambiguous description of spatial quality, leading to clear definitions of attribute scales that are much more likely to result in meaningful and consistent responses from listeners (see Section 1.3). It is therefore considered important to differentiate between the two concepts. Yet, this poses the question of which of them should be made the subject of this study. Or, more specifically, which notion is more likely to result in an unequivocal, artefact-free representation of changes in the desired quality that will enable effective listener training?

Figure 5.19: Graphical illustration of the concept 'linear ensemble width'. Note the additional source distance component (see text).
To resolve this issue a pilot experiment was conducted. Five researchers at the Institute of Sound Recording (IoSR) – all very experienced in evaluating reproduced sound – were asked to compare two groups of musical stimuli illustrating the two different EW concepts (for information on the stimulus generation process see Section 5.9.4). The experimental design was very similar to the one employed for the validation of the SD and SW samples, i.e. both MDS and verbal reporting were used again. It was found that the MDS solutions of the two scenarios were almost identical, pointing towards a single strong dimension in each case. However, it should be noted that these sensory profiles are somewhat misleading. A second spatial characteristic is bound to be present in each simulation scenario, which can be described as 'source distance' in the case of 'linear EW' (see Figure 5.19) and as 'ensemble depth' in the case of 'constant distance EW' (see Figure 5.20). Whilst these artefacts were not revealed by the MDS analyses, references to source range were made during the verbal elicitation phase. That is, not only did subjects specify perceiving lateral source separations in the case of 'linear EW', but also changes in the relative distances of the sources. On the contrary, not a single listener mentioned depth changes in the case of 'constant distance EW'. This may be the result of these stimuli being conceived more holistically because of the listeners' preconditioning. In a concert situation, musicians are much more likely to position themselves in a curved arrangement in order to facilitate eye contact and thus musical interaction. Since all five listeners have a musical background, it is surmised that they perceived 'constant distance EW' as being more natural, taking for granted the additional depth variations. Another possible explanation for this might be that humans in general are more accustomed to listening for cues to how far away a sound event is than for cues to its depth. This is analogous to what was said previously about source width being a much less familiar percept than source position (see Section 5.8.1). Irrespective of the exact reason for the absence of depth descriptors in the listeners' verbalisations, ensemble depth can be considered a sub-attribute (see Section 0.4) of the 'constant distance EW' percept. That is why, from the point of view of accomplishing unidimensionality, it is not a problem that the MDS algorithm failed to unravel it.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

other sensory evaluation situations, however, such 'negligence' may not be tolerable (see also Section 5.10.1).

Hence, because of the absence of any additional spatial constructs in the listeners' verbal responses, the 'constant distance EW' approach was chosen for the remainder of this study as it seemed more promising with respect to achieving unidimensionality.

5.9.2 Spatial attribute processing platform

In the field of reproduced sound, the most commonly used processing technique for controlling the direction of a source within an auditory scene is the panpot. Normally, this device features on most channels of a typical mixing desk. It is true that such a mixing console was used in conjunction with other commercially available hardware devices for the creation of the SD and SW stimuli. However, this approach was deemed too cumbersome and restrictive for the synthesis of ensemble-level spatial attributes, because of the need to control numerous parameters for several component sources simultaneously and the fact that the Lexicon 480L reverberator is only equipped with a 2-channel input and a 4-channel output. Therefore, in order to facilitate the rendering of sound scenes containing multiple sources, a processing platform was implemented with the help of the object-oriented programming language MaxMSP. The devised software runs on an Apple Macintosh G4 computer and in its current configuration allows for the simultaneous processing of up to seven monophonic input signals. For each of these, 12 early reflections can be generated and individually adjusted in terms of level, delay and angle of incidence. This corresponds to all first- and second-order specular reflections in the horizontal plane. Since the height dimension is not taken into account by most reproduction systems and the relevance of floor and ceiling reflections to perceived sound quality has yet to be investigated in detail, it was decided to ignore them because of the need for CPU 'housekeeping'. As regards the generation of specular reflections only, Martin [2001] emphasised the perceptual benefits of including a diffusion control in such a system. Yet, he also acknowledged a resultant steep rise in computational cost, which is why this issue has been neglected so far. For each reflection order a biquad\footnote{‘Biquad’, which is short for ‘bi-quadratic’, is a common name for a two-pole, two-zero digital filter (e.g. see [Smith, 1985] for more information).} filter can be inserted into the signal path to simulate the effects of wall absorption. A 4-channel decorrelated reverberation stream is also computed, the level and decay time of which can be independently controlled in three separate, adjustable frequency bands. The reverb processing is based on a slightly modified version of an algorithm developed by Jot [1997]. Different techniques can be employed for directionally encoding the DS and/or the ERs. Currently, 5-channel pairwise constant power panning (PCPP) and a novel ambisonic-based technique (see Section 5.9.3) are supported. For reasons outlined in Section 5.2 an ITU set-up is used at the reproduction end. A block diagram of the processing platform is shown in Figure 5.21. Due to its modular structure modifications and extensions can be accomplished fairly easily as different requirements arise and more processing power becomes available. Ultimately, the platform is intended to serve as the framework for a stand-
alone ear-training system offering real-time control over each of the chosen spatial characteristics. As such, it represents a novel approach to the design of a spatial sound processor.

5. Unidimensional simulation and validation of spatial attributes of reproduced sound

Figure 5.21: Block diagram of implemented processing platform

5.9.3 A new surround panner

Since for this study an artefact-free panning technique is virtually mandatory if unidimensionality is to be achieved, an evaluation of selected surround-panning approaches was undertaken. This was necessary because two independent studies [West, 1998; Martin et al., 1999] have scrutinised PCPP in the context of multichannel audio in terms of a number of qualitative criteria and identified various perceptual deficiencies. While the differences in terms of localisation accuracy between PCPP and more esoteric panners (e.g. vector-optimised ones [Gerzon, 1992b]) were found to be small, other artefacts were detected as a function of azimuthal positioning. For instance, variations in apparent source width and distance were noticed. What is more, neither of these two studies made use of an ITU-compatible loudspeaker layout, i.e. the spacing between the front and rear speakers was smaller in both cases23. Therefore, it is possible that these artefacts will be more prominent if PCPP is used in an ITU-conformant playback situation.

In the hope of overcoming some of the aforementioned side-effects, a novel fourth-order 'ambisonic' pan law was considered. This algorithm was developed by Peter Craven of Algol Applications in

23 In [West, 1998] the left and right loudspeakers were located at ±45° instead of the more conventional ±30°, and in [Martin et al., 1999] an 8-channel, radially symmetrical (horizontal-only) set-up was used, resulting in loudspeaker spacings of 45°.
association with Meridian Audio Ltd. and has been optimised for 3/2-stereo reproduction in accordance with energy and velocity localisation theories. Thus, it was anticipated that as a result of the optimisation unwanted artefacts would be less audible. Seeking confirmation for this assumption, the new panner was compared to ITU-compatible PCPP using analytical as well as subjective testing methods.

Analytical comparison of new panner and 5-channel PCPP

To be able to describe the new algorithm in more detail it is helpful to assess it in terms of Gerzon's [1992b] design aims for multichannel panners. These are based on a theoretical model for the psychoacoustics of directional hearing so as to satisfy as many different auditory mechanisms as possible and hence to achieve reliable, stable and natural sound image localisation.

Regarding the functionality of the ambisonic panner, for a given source signal five gains, one for each speaker, are calculated as a linear combination of nine circular harmonics [Craven, 2002], i.e.:

\[
1, \cos(\theta), \sin(\theta), \\
\cos(2\theta), \sin(2\theta), \\
\cos(3\theta), \sin(3\theta), \\
\cos(4\theta), \sin(4\theta),
\]

where \(\theta\) is the desired panning angle. This results in a total of 45 coefficients, but due to the left-right symmetry of the ITU layout the number of independent coefficients is reduced to 23. These are then adjusted numerically in order to optimise a weighted sum of the following five psychoacoustic criteria:

1. The reproduced energy should be substantially independent of the panning angle;
2. The perceived angles \(\theta_v\) and \(\theta_e\) derived by the velocity and energy localisation theories should be closely matched;
3. The perceived angles \(\theta_v\) and \(\theta_e\) should be reasonably close to the panning angle \(\theta\);
4. The velocity vector length \(r_v\) should be close to unity;
5. The energy vector length \(r_e\) should be as large as possible.

For a useful summary of the localisation parameters \(\theta_v, \theta_e, r_v\) and \(r_e\) the reader is referred to [Gerzon, 1992b]. Although each localisation parameter was evaluated for panning angles covering 360°, frontal sources were weighted most and rear sources least strongly to produce a pan law compatible with general industry practice [Craven, 2002]. Further details about the design process are given in [Craven, 2003].
In Figure 5.22 the resulting speaker feeds are shown for $-180^\circ \leq \theta \leq 180^\circ$, to be compared with the corresponding gains for 3/2-stereo compatible PCPP (Figure 5.23).

Since for this work the sound quality of frontal images is most relevant (see Section 5.9.4), the remainder of this discussion will address the differences between the two techniques for $-60^\circ \leq \theta \leq 60^\circ$. In Figure 5.24 and Figure 5.25 $\theta_v$ and $\theta_E$ are plotted against the panning angle $\theta$ for the new panner and PCPP, respectively.
As can be seen, for the new panner $\theta_v$ and $\theta_e$ trace each other almost perfectly as long as $\theta$ is within $\pm 30^\circ$. For PCPP, on the other hand, the $\theta_e$ curve 'oscillates' around the $\theta_v$ curve, meaning that there is little agreement between the sound image position predicted by the velocity and energy localisation theories. As was to be expected, the $\theta_e$ curve is 'flat' when $\theta$ corresponds to a loudspeaker position. This implies that the energy image is stationary whilst the velocity image continues to move with $\theta$. Thus, the energy localisation tends to 'hug' the loudspeakers.
Figure 5.26, Figure 5.27 and Figure 5.28 compare $r_E$, $r_V$, and $(\theta_V - \theta_E)$ respectively for the new panner and for PCPP. For each of these plots the independent variable is the reproduced angle $\theta_V$. 

Figure 5.26: $r_E$ for new panner (solid) and PCPP (dot-dash) (reproduced from [Craven, 2002] with permission)

Figure 5.27: $r_V$ for new panner (solid) and PCPP (dot-dash) (reproduced from [Craven, 2002] with permission)
Even though the two panners perform similarly in terms of the length of the energy vector $r_E$ (see Figure 5.26), the ambisonic technique has a less severe reduction in $r_V$ for $\theta < -35^\circ$ and $\theta > 35^\circ$ (see Figure 5.27). Besides, from Figure 5.28 it can be seen that the new algorithm achieves a much lower directional error $(\theta_V - \theta_E)$ across the frontal sound stage compared to PCPP.

Overall then, the new pan law seems to offer an improvement for sources panned between L, C and R compared to PCPP. However, since it cannot be ascertained from this analysis whether the improvement is in the form of a better localisability or whether other qualitative features are (also) positively affected, an informal listening test was carried out.

Subjective comparison of new panner and 5-channel PCPP

In order to get an indication of the perceivability of the aforementioned side-effects, four spatial audio researchers at the IoSR made informal A-B comparisons of the sound image quality produced by 5-channel PCPP and the ambisonic algorithm. As it turns out, the new panner was clearly preferred by all listeners for its superior performance with respect to timbre as well as image width constancy across the frontal sound stage.

Martin et al. [1999] have pointed out, for PCPP, a potential singularity in terms of active (i.e. sound-radiating) loudspeakers and related this to changes in perceived image width. In order to be able to describe this more accurately, Craven [2002] proposed the ‘chorus’ parameter as a measure of the effective number of active loudspeakers. He defined it mathematically as:

$$\text{Chorus} = \frac{(\text{feed}_L^2 + \text{feed}_R^2 + \text{feed}_C^2 + \text{feed}_{LS}^2 + \text{feed}_{RS}^2)^2}{\text{feed}_L^4 + \text{feed}_R^4 + \text{feed}_C^4 + \text{feed}_{LS}^4 + \text{feed}_{RS}^4}$$
where $\text{feed}^2_{L}$ is the square of the signal for loudspeaker $L$. Thus, chorus takes the value 1 if only one speaker is driven and the value 2 if just two speakers are driven equally. In Figure 5.29 chorus is plotted for each of the two panners. As anticipated, for PCPP chorus falls to the critical value of unity whenever the sound is panned to a loudspeaker direction.

This situation will not arise with ambisonic techniques, because in order to reconstruct a given wavefront as faithfully as possible, all speakers will radiate sound energy irrespective of the encoded direction. Closely tied into this observation is the assertion that the improved timbral constancy also stems from Ambisonics’ democratic approach regarding the number of active channels. In [Martin, 2002] the audibility of interchannel interference at the listening position of a 3/2-stereo set-up was investigated. It was reported that for audio signals with a high degree of interchannel correlation (e.g. pairwise amplitude-panned signals) the perceptibility of comb-filtering (as caused by head rotation) increased with smaller interchannel level differences and smaller spatial separation of the undelayed and delayed signals. It is speculated that in the case of the new panner this kind of acoustic interference evens out more compared to PCPP because of the use of five (as opposed to just two) output channels for assigning a certain direction to a sound image. Thus, this may be responsible for the improved timbral constancy of the new pan law compared to PCPP.

Admittedly, the new panner cannot remedy well-known shortcomings of the 3/2-stereo layout (e.g. unstable lateral phantom images) and possibly does not create as wide a listening area as PCPP. Notwithstanding, for this study neither had the sounds to be panned well beyond the frontal loudspeakers nor was off-centre listening required. That is why these deficits were of no concern.

---

24 In [West, 1998] PCPP was found to perform slightly better in terms of localisation accuracy for off-centre listening positions compared to an "optimal", vector-optimised panning technique. As expected, this was not the case for sweet spot listening.
5.9.4 Creation of ensemble width stimuli

Having devised a suitable spatial processing engine, a set of test stimuli could be created. In order to evoke the cognitive cues necessary for conjuring up the impression of an ensemble rather than several discrete sources, non-musical signals seem less suitable. Undoubtedly, the term 'ensemble' bears strong musical connotations. Nonetheless, five anechoic speech recordings (two male, three female) in four different languages (English, French, German and Danish), taken from the Archimedes and SQAM [EBU, 1988] CDs, were used as the source material. Human speech is a very familiar and hence critical test signal, which is why it was applied during this experiment so as to reveal all perceivable attributes. Also, its broadband spectrum and syllabic envelope characteristics allow the hearing mechanism to collect multiple localisation cues [Blauert, 1997], thereby creating more stable phantom image positions compared to narrow-band, continuous signals. As regards the concurrent use of five contextually unrelated voice recordings, the resultant auditory scene was reminiscent of the 'cocktail party effect' [Bronkhorst, 1999] – an acoustic scenario that hearing humans are very familiar with. By using speech in four different languages, it was hoped that this would encourage the subjects to listen to the stimuli more holistically, rather than trying to understand each speaker separately. The exclusive use of the subjects' mother tongue would probably have been an incentive for analytical listening, thus perceiving each talker as an individual source, rather than as a constituent of a group of speakers.

Using the ambisonic panner in conjunction with digital reverberation and mixing techniques, a set of stimuli was synthesised. The sounds altered between all five sources being clustered around C and the sources being separated widely whilst maintaining constant perceived distance at all times (see Figure 5.20). In particular, one female voice would always remain panned to C whereas the other sources would gradually spread out in a symmetrical manner with their relative angular separations in- or decreasing proportionally. For the two widest stimuli the outermost (female) voices could be panned beyond L and R with a maximal perceived angular width of about ±40° without having to jeopardise phantom image stability or timbral constancy.

Besides, it was found that with increasing displacement the four non-static sources appeared to get slightly louder. There may be a number of explanations for this observation. For instance, it may be traced back to the psychoacoustic phenomena of directional loudness perception and spatial unmasking. With regard to the former, there is an increase in the sensitivity of the human hearing system in the horizontal plane as a source is displaced laterally, reaching its peak in the ipsilateral ear at about 60° [Zacharov et al., 2001]. This is the result of the directional dependency of the HRTFs [Blauert, 1997]. Spatial unmasking implies that when one of two sources located directly in front of a listener is moved off-centre, changes in the relative source levels at the ears will be evident due to the introduction of head shadowing [Shinn-Cunningham et al., 2001]. (Also, different interaural time and level differences will be caused by the source displacement, helping to increase speech intelligibility.) Bearing in mind that the human hearing system is used to these effects, one may claim that the loudness changes do not need to be addressed as part of this study. However, this argument has only
limited applicability as the created acoustic scenario is very unnatural, its choice being somewhat
dictated by the requirements for effective listener training. A more plausible explanation for the need
to adjust loudness might be that higher-level cognitive factors are involved in evaluating the sound
quality of the panned sources. More precisely, the positional variations in the voices resulted in very
large changes in speech intelligibility (e.g. see [Shinn-Cunningham et al., 2001]). Hence, the outer
sources were perceptually emphasised, which could have led listeners to base their judgements solely
upon the behaviour of the two outermost voices.

From the above, it is evident that it is unclear which underlying mechanisms are responsible for the
occurrence of this phenomenon. Hence, further research would have to be carried out before cogent
answers as to its origin could be given. Nevertheless, the levels of the non-static sources were
gradually reduced by maximally 2.8dB for the outer and 1.2dB for the inner two voices as they were
panned off-centre in order to counterbalance the audibility of this side-effect.

In addition to the direct sounds (i.e. the anechoic recordings), diffuse reverberation was employed to
produce a sense of space and hence to increase the naturalness as well as the ecological validity of the
rendered auditory scenes. To avoid source broadening no discrete ERs were included in the simulation
(see Section 5.11.3). It is true that reverberation leads to a degradation of directional hearing
performance, because it distorts interaural time and level differences as well as spectral cues [Shinn-
Cunningham, 2002]. Likewise, it can impair speech intelligibility [Everest, 2001]. Yet, since this
study did not aim to ensure optimum localisation accuracy and informational exchange (four different
languages were used after all), the perceptual significance of these effects was small and easily
outweighed by the gain in experienced realism – a conclusion also drawn by Shinn-Cunningham.

Moreover, the reverb was used for making small corrections in the apparent closeness of each source.
Effectively, the D/R was exploited for each voice to craft the desired range perception (see Section
2.2.4). Since the DS level was altered as a function of encoded direction (see above), the level of the
added reverberation had to be adjusted accordingly to make ‘constant distance EW’ possible.

In the case of this study, the problem of inhomogeneous distance perception was slightly aggravated
due to some voices having been recorded at short distances (~40cm) from a unidirectional
microphone, manifesting itself in the form of the proximity effect [Josephson, 1999]. As pointed out
in Section 2.2.1, under natural listening conditions a LF boost will be apparent for nearby sources.
Therefore, the proximity effect could have introduced unwanted distance cues, which is why the
concerned recordings were high-pass filtered before integrating them into the mix.

Based on the methodology depicted above, nine stimuli were carefully synthesised, each
demonstrating a different magnitude of ‘constant distance EW’. The entire processing took place in
the digital domain and the sound excerpts were recorded to hard disk as 5-channel .aiff files with a
sampling rate of 44.1kHz and a resolution of 16 bit.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

5.10 Validation of ensemble width stimuli

As had been the case for the SD and SW stimuli, a validation experiment was performed in order to assess the degree of success achieved in trying to simulate 'constant distance EW' in a unidimensional manner. With the help of the implemented processing platform (see Section 5.9.2) it had been possible to create the stimuli in the listening room. Hence, by conducting the validation test in the same location, potential problems due to transferring the sound examples to a different reproduction environment could be avoided. 14 final-year 'Tonmeister' students acted as the listening panel. Unlike in the previous studies, all subjects completed a short training session prior to the proper test, so that they could acquaint themselves with the task, the user-interface and the scale. It was hoped that this would help minimise the error variance in their judgements. The training comprised a comparison of nine stimuli that were different from the ones to be validated afterwards, followed by six paired comparisons. Refinements were also made with regard to the previously adopted validation methodology so as to make the verification process even more reliable. The various modifications are outlined in the next section.

5.10.1 Experimental design

As the discussion of the concepts of 'linear EW' and 'constant distance EW' in Section 5.9.1 indicated, there are instances when MDS will fail to reveal attributes that subjects could detect during a similarity rating exercise. To reiterate, the MDS solutions of the two EW simulations had both pointed towards a single strong dimension in the data despite the fact that in each case two separate spatial attributes were (inevitably) varied. This is because for each simulation the attributes changed in tandem, thereby prohibiting the listeners to grade them independently. Thus, it seems that MDS can only uncover those dimensions that are orthogonal with respect to each other, whereas perceptual factors that are directly correlated will be compressed into a single statistical dimension. While this aspect of (multi)collinearity has been discussed in the context of other statistical techniques (e.g. multiple regression analysis [Hair et al., 1998]), it appears to have been neglected as far as MDS is concerned. Thus, highlighting this problem constitutes a novel contribution to the field of sensory evaluation.

In addition to MDS' insensitivity to distinctly perceivable attributes varying in parallel along a single dimension, another limitation has to be pointed out. The classical MDS (and INDSCAL) model assumes that a stimulus set possesses a small number of common attributes. Yet, such a model may not always be appropriate, for in addition to these shared dimensions subjects may be able to detect perceptual factors specific to a particular stimulus. Such discontinuous dimensions cannot be revealed by traditional MDS analyses either. More recently, however, new techniques have been employed in timbre perception research that promise to complement the classical ones in various ways, e.g. one can analyse for discontinuous, stimulus-specific dimensions or 'specificities'. For instance, in the case

25 It is assumed that such factors would also have to be linearly related for this to be true. A, say, linear increase for one and a quadratic increase for another factor should be disclosed by the MDS solution in the form of a second salient dimension.
of the ‘extended two-way Euclidean model’ with common and specific dimensions the distance between stimuli $i$ and $j$ is given by [Winsberg & Carroll, 1989]:

$$d_{ij} = \left[ \sum_{r=1}^{R} (x_{ir} - x_{jr})^2 + s_i + s_j \right]^{1/2}$$

where $R$ is the number of dimensions in the common space shared by all stimuli, $x_{ir}$ is the co-ordinate of stimulus $i$ on the $r$-th common dimension and $s_i$ is the square of the (non-zero) co-ordinate of stimulus $i$ along the dimension specific to that stimulus. A specificity may therefore be thought of as the square of the perceptual strength of a feature possessed by the associated stimulus [McAdams et al., 1995].

Thus, the extended two-way Euclidean model seems to be a useful addition to a sensory evaluation strategy designed for checking the unidimensionality of an attribute simulation. Yet, although there could be, in principle, more than one specific dimension for each stimulus, this formulation is mathematically indistinguishable from the case of exactly one specific dimension per stimulus [Carroll & Winsberg, 1995]. Hence, parsimony dictates assuming only one specificity per stimulus. Irrespective of this limitation, the technique should at least be able to verify if individual stimuli are artefact-free or not. Unfortunately, however, “the interpretation of these specificities is, at this point, intuitive and needs to be substantiated by ... further perceptual tests” [Krumhansl, 1989]. That is to say that there are no established guidelines yet as to what magnitude of a specific dimension is considered perceptually salient.

From the above it follows that for the applied validation methodology to be as reliable as possible these shortcomings should ideally be addressed. As for attributes changing concurrently along single orthogonal MDS dimensions, help is fortunately at hand. The reader may remember that in the case of the ‘linear EW’ vs. ‘constant distance EW’ example some listeners had mentioned source distance changes. Hence, a workaround to the problem of revealing perceptual factors changing in parallel is the collection of additional verbal data. Not only is such information needed to interpret common, continuous dimensions, but it may also serve as a means of identifying artefacts that MDS on its own cannot deal with. That is why, for this validation experiment, the MDS results were supplemented with verbalisations again. In addition, it was decided to collect non-verbal responses, too. Other researchers have looked at the advantages and disadvantages of these two types of data [Mason et al., 2001]. They pointed out that when trying to elicit the spatial features of certain stimuli, non-verbal representations can circumvent many limitations of verbal language. For instance, semantic descriptors become less useful if a given stimulus is more complex and therefore difficult to describe. In contrast, drawing enables a subject to delineate auditory space more faithfully. In this respect, graphical elicitation techniques were found to be particularly useful for investigating the spatial attributes of image width, location and skew [Ford et al., 2001]. For that reason, following the
attribute-free comparison required for the MDS analysis, the listeners were asked (1) to verbally express all differences they had perceived between the stimuli and (2) to depict their verbal responses graphically (whenever relevant).

In the case of (1) a questionnaire was provided again. Similar to before, listeners were encouraged to write down words and descriptions that were differential in nature. However, whilst they had previously been instructed to note these perceptual changes in order of priority, this time they had to grade them on a scale from 1 to 10. Consequently, the measurement level was increased, which was meant to lead to a more accurate quantification of the perceptual salience of each perceived difference. A ‘1’ was defined to correspond to a subjective effect being just audible whereas a score of ‘10’ was specified to imply that a particular difference was the only subjective effect perceivable between all nine stimuli. Intermediate anchor points were deliberately omitted, since they can cause problems with subject-dependent interpretations and ensuring linear scale increments [ITU-R BS.1116, 1997].

In the case of (2) all participants were given an A4-sized sheet of paper displaying an outline of their surroundings as an indication of scale. The listeners’ task was to draw any spatial changes they had detected as faithfully as possible. Due to the fact that the EW concept cannot be fully described using a few words only (see Section 5.9.1) it was anticipated that the drawings would help them explicate their verbal responses. Crucially, it was only after the completion of both the MDS and verbal reporting stages that subjects were told that they had to draw what they had heard in order not to bias them in terms of the type of perceptual changes detectable between the sound fields.

It has been argued before that in order to be able to guarantee an optimal training effect, an attribute simulation has to be free from confounding sensory interactions. Thus, if a single dimension is identified, it is of great importance that this dimension results from perceived variance in the attribute of interest only. The last ‘quality control’ step therefore was to tell the listeners that the stimuli were intended to vary only in the subjective quality that they had graded highest, and to ask them whether, knowing this, anything “sounded wrong”. In this context, it is worth emphasising that throughout the whole experiment the chosen descriptor of ‘ensemble width’ was never mentioned so as to prevent distortion of the subjects’ wordings. For the same reason, care was taken to use each participant’s own terminology only when discussing his/her responses for clarification purposes.

Finally, small modifications were made to the user-interface implemented in the listening test software. While in the previous studies an undifferentiated line scale had been provided for the subjects to indicate their perceptions (see Figure 5.5), numeric scale labels ranging from 1 to 9 were placed below the (non-quantised) slider this time. During the SD and SW studies subjects had expressed their concern with regard to their ability to remember the magnitude of inter-stimulus differences for a prolonged period of time. Hence, it was hoped that these visual anchors would help them be more consistent.
5.10.2 Results

Since the number of stimuli included in the listening test had not been changed and the format of the collected similarity scores was not affected by the applied modifications outlined above, the MDS analysis was carried out in the same way as before. The results are presented below.

Analysis of similarity data

Figure 5.30 displays the badness-of-fit measures as calculated by the MDS model. In comparison with the scree plot from the SD experiment (see Figure 5.6), the s-stress and stress metrics shown here have smaller magnitudes, especially in the case of the 1-D solution. This could be due to a positive effect of the training exercise and/or the use of scale labels. In addition, both s-stress and stress exhibit a knee at 2-D, thereby denoting that the addition of supplementary dimensions does not sufficiently reduce stress to be worth trying to interpret. Nonetheless, this still leaves the problem of having to decide whether or not the second dimension is perceptually meaningful.

Figure 5.30: Scree plot derived from non-metric MDS analysis (ensemble width)

Examination of the RSQ results (see Table 5.12) shows that the 1-D EW solution sets a new record. Recalling that the 1-dimensional SD model could account for a considerable 80% of the data variance, the EW solution exceeds this value by 3%. This may be seen as an indication that the EW sounds could convey a certain qualitative effect to the listeners in an even clearer manner than the SD stimuli and/or that the panellists were more consistent in their ratings. Furthermore, extending the model by a second dimension only raises RSQ slightly (i.e. by less than 0.05). This suggests a very good match between the similarity data and the fitted 1-D model, which in turn conveys a strong unidimensional effect being present in the stimuli.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

Table 5.12: RSQ results from non-metric MDS analysis (ensemble width)

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Figure 5.31 is the result of depicting the panel’s perception of the stimuli in (1-dimensional) graphical form. Although the stimuli are not spread linearly, they appear in the order intended (‘a’ being the widest and ‘i’ the narrowest sound excerpt). Hence, it seems that the stimuli exhibit different intensities with regard to a particular sensory property, because the listeners could rank them correctly.

On the whole, one can therefore conclude that the MDS analysis managed to disclose the envisaged psychological organisation of the stimuli from the panel’s responses.

Analysis of verbal data

The procedure for analysing the verbal data was identical to before, except that because the subjects had quantified the ‘magnitudes of audibility’ of the perceived effects in a different way (i.e. on a scale from 1 to 10), the ratings had to be treated differently, too. Similar to before, an index of perceptual importance for each group of terms was calculated. Specifically, the scores of all verbal descriptors within each group were added up. The resultant values were then divided by 140 (the number of subjects multiplied by 10), leading to a maximum possible weight factor of 1 again.
As for the holistic terms, four groups were identified, which are shown in Table 5.13. The first category has been labelled ‘Ensemble width’, because it contains terms closely related to this notion. 13 participants perceived and rated it as the most salient difference, resulting in a large weight factor of 0.82. The listeners themselves used descriptors such as “Space between voices”, “Position of voices: all in middle vs. displaced laterally”, “Spatial separation of sources”, “Collective width of voices: clustered together vs. widely spaced apart”, “Stereo width: sources either on top of each other or spread apart”, “Distance between voices”, “Stereo image width varied between very wide and mono”, “Image spread”, “Spacing of sources changed between mono and widely spaced”, “Perceived angle between voices” and “Movement of outer voices from central position to fully left/right”. Whilst it had been relatively straightforward to decide whether or not each response from the SD or SW validation studies conformed with the intended auditory phenomenon, this is not the case for this study. That is to say that by examining the EW responses on their own one cannot conclude that they are equivalent to this author’s concept of ‘constant distance EW’, as it is perceptually much more complex than, say, the distance of single source. Nevertheless, as will be argued in the following, joint evaluation of all types of collected data does allow this inference to be made. It should also be noted that no descriptors related to the ‘ensemble’ as a whole were elicited from the panel. A possible reason for this was already commented upon in Section 5.9.4, i.e. it is likely to be the result of utilising speech as opposed to musical sounds for the source material.

Table 5.13: Groups of holistic terms and their relative weights (ensemble width)

<table>
<thead>
<tr>
<th>Holistic groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble width</td>
<td>13</td>
<td>0.82</td>
</tr>
<tr>
<td>Clarity/Intelligibility</td>
<td>3</td>
<td>0.11</td>
</tr>
<tr>
<td>Distance of widely spaced sources</td>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>Room ambience</td>
<td>1</td>
<td>0.01</td>
</tr>
</tbody>
</table>

The second group contains references to clarity and intelligibility. In particular, three subjects found it easier to pick out the individual voices when lateral source separation increased. Clearly, this perceptual factor is intrinsically linked to the EW attribute due to the phenomenon of spatial unmasking (see Section 5.9.4). Despite the fact that the two occur in tandem the simulation is not flawed. It is just an inevitable side-effect (or sub-attribute), which in itself does not preclude unidimensionality. Interestingly, two other listeners preferred to express their perception of changes in speech intelligibility in the form of a passing comment instead of a distinct difference on the questionnaire. This could be interpreted as an instinctive acceptance of this relationship on their behalf. The remaining two groups comprise spatial attributes again. One listener experienced distance changes in the outermost voices of the widest stimulus relative to narrower ones. Another subject perceived differences in room ambience between wide stimuli only. Both of these artefacts may be
due to imprecise manipulation of the D/R used to control constant apparent distance (see Section 5.9.4). Still, these differences are not present in the responses of any other subject. Also, the scores given to them are small, which is why the corresponding weight factors are a fraction of the one obtained for ‘Ensemble width’. On these grounds, these artefacts can be dismissed as being perceptually insignificant.

The results from classifying the analytical terms are displayed in Table 5.14, showing that references to ‘Spectral changes’ were made primarily. Three out of the six terms concern variations in HF content (i.e. “Changes in HF response”, “Lack of HF for some narrow sounds, but other narrow ones OK” and “Upper mid to HF loss”), one addressed overall spectral changes (“Timbre of the outermost voices for the widest stimuli”), another one LF content (“Loss of LF for narrow sounds”) and the sixth one yet another frequency region (“Differences in low mids when at least one sound was spread out”). Unquestionably, the responses differ greatly with respect to the affected frequency bands and inter­ listener concordance is virtually non-existent. Thus, the total weight factor calculated for this category is somewhat misleading.

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral changes</td>
<td>6</td>
<td>0.17</td>
</tr>
<tr>
<td>Levels of discrete voices</td>
<td>2</td>
<td>0.07</td>
</tr>
<tr>
<td>Panning</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>Phase</td>
<td>2</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Furthermore, two listeners effectively stated that changes in EW were accompanied by level variations in the non-static voices (“Volume of outer voices increased for wide sounds” and “Changes in levels of individual voices”). Hence, it may be that the applied gain adjustments were not accurate enough to subjectively segregate changes in lateral source separation from this factor. The ‘Phase’ category is composed of the terms “Phase cancellation for medium spaced sounds” and “Wide stimuli sounded phasy”. While the former is difficult to account for, the latter could be explained by the negative channel gains commonly encountered with Ambisonics (e.g. see [Rumsey, 2001]). Having said that, neither was phasiness experienced by any of the other listeners nor was it detected during the informal comparison of the new panner to PCPP (see Section 5.9.3). At any rate, similar to the other analytical groups, the ‘Phase’ category exhibits a small weight factor and can therefore be ignored.

Finally, it is emphasised that all but one listener delineated their perception of variance in EW using a holistic descriptor even though the participants had been equipped with analytical listening skills during their education. The other subject preferred the more technical descriptor “Panning: relative
source positions", which constitutes the fourth analytical category. Evidently, panning is directly related to the applied processing.

All in all then, the verbal data seem to imply a successful unidimensional simulation of ‘constant distance EW’. However, because of the difficulty to pin down this percept using language only (see Section 5.9.1), the remaining data have to be examined as well in order to be absolutely sure.

**INDSCAL analysis**

When scrutinising the listeners' positions in the 2-dimensional subject space for the EW similarity judgements (Figure 5.32), it seems that Subject 8 perceived the stimuli somewhat differently compared to all the other panellists. Another explanation for his more isolated position could be that he was less consistent when grading the stimuli. Still, his weight for Dimension 1 is larger than the one for Dimension 2, as is the case for all the other listeners. Thus, it can be concluded that Dimension 1 was perceptually the more important of the two, thereby confirming the earlier finding that the panellists identified and used the same (main) effect for grading the stimuli.

![Figure 5.32: 2-dimensional subject space from non-metric INDSCAL analysis (ensemble width)](image)

As regards the second dimension, an example for its meaninglessness is evident from Subjects 9 and 13. According to their verbal responses, both listeners experienced just a single audible effect ("Separation of sources" and "Spacing between individual voices", respectively). However, their weights along the second dimension are larger compared to all but one listener. This is a strong sign that, at least in the case of these two subjects, Dimension 2 constitutes noise. Thus, in effect, the subject space has the potential to act as some kind of a noise gate that can 'mute' those signal parts
unrelated to the input to the auditory system, i.e. those dimensions, which are the result of listeners making inconsistent similarity judgements. However, another signal is required to set the 'gating threshold' correctly as well as to 'trigger' the gate – a verbal response.

In contrast to Subjects 9 and 13, the majority of the other participants verbalised at least two differences, many of which are directly related to the concept of ensemble width or at least the applied processing. Enough evidence is therefore available to deduce that the panel as a whole did not identify a second perceptual dimension in the EW stimuli.

Analysis of graphical data
The graphical responses were simply scrutinised by means of visual inspection. Specifically, it was checked that no spatial information was contained in the drawings that would have contradicted the implied meanings of the verbal data. Figure 5.33 shows a typical example of a graphical response, the remaining ones being located in Appendix A. The reader should note that for reasons of clearness it was suggested to listeners to sketch the two extremes of their perceptions only. Likewise, they were asked not to draw them on top of each other. Hence, extreme 'A' is slightly displaced from extreme 'B'.

Although the visible surroundings had been delineated on the response sheet to help subjects draw to scale, considerable differences in overall distance are apparent. Nevertheless, this does not come as a big surprise, as humans are notoriously bad at judging absolute source range (e.g. see [Nielsen, 1993]). More importantly, no other spatial variations are apparent for the panel as a whole. Also, the trend for the outer sources to bend in towards the listener is evident from all sketches. Thus, whilst
this intended feature of the stimuli had not been obvious from the other types of data, the graphical responses imply that the listeners perceived the ‘constant distance’ aspect of the simulation.

... but does it sound right?
As mentioned in Section 5.10.1, the final step of the validation strategy was to tell the listeners that the stimuli were intended to vary only in the subjective feature that they had perceived as being the most prominent difference. Equipped with this knowledge they were essentially asked to reflect upon their aural experience and to decide whether modification of the sounds was required to ensure total compliance with their own understanding of the concept. All listeners were offered the chance to listen back to the stimuli once more. Satisfyingly, only one subject felt the need to comment. In particular, the widest sound image was perceived as being asymmetrical, i.e. the voices on the right-hand side appeared to be closer to the central source than the ones on the left. Besides, it was criticised that there were only clear changes in the positions of the two outermost voices. Even though symmetry of movement had been aimed for during the generation process, it is not a critical feature of the notion of ‘constant distance EW’. Also, the comparatively bigger angular increments for the outermost voices were included deliberately so as to maintain equal spacings between each pair of voices for a given stimulus. The fact that the other listeners did not mention anything in this respect does not necessarily mean that these artefacts were not present in the simulation. It could be that because the artefacts did not interfere with the subjects’ understanding of the concept, they found it unnecessary to comment. Thus, by and large, the stimuli were found to “sound right”.

In summary, as the analysis of the various types of data showed, a single common dimension was detected and used by the sensory panel in comparing the created EW stimuli. While there might exist additional perceptual factors for the listeners individually, plenty of evidence has been presented to conclude that the panel as a whole did not identify a second dimension. Hence, unidimensionality of the ‘constant distance EW’ simulation was accomplished, thereby suggesting it to be suitable for training applications.

5.11 Simulation of ensemble depth I
The ED construct resembles its width sibling in that it is a rather unexplored spatial percept. As pointed out in Section 2.3, some researchers have attributed its perceivability to the presence of early lateral reflections in a reproduced sound scene. However, little formal evidence has been gathered and published so far to test and further specify this assertion. That is why it was considered appropriate to conduct a preliminary investigation into the perceptual effects of various finer details of ER patterns before attempting an ED simulation. An extensive description of this pilot study will be given first, after which the generation of a set of ED stimuli is to be delineated in the usual manner.
5.11.1 Pilot study into the perceptual effects of early reflection pattern characteristics

Ideally, spatial audio systems should be capable of recreating a truly stereophonic sound field so as to enable spatial distinction of sources in all three dimensions. However, while exercising control over the lateral plane of an auditory scene is usually fairly straightforward (e.g. by using suitable panning techniques), a vivid sense of perspective is much harder to come by. Undoubtedly, the perceptibility of front-back separation in a recording can greatly enhance the listening experience and can lead to an improvement in perceived realism. Thus, the question arises which factors are responsible for the hearing of a depth dimension.

If one assumes that the ‘early lateral reflections’ predication is true, then the design of ER patterns, both in the temporal and spatial domain, may be crucial for the achievement of depth imaging. Since hardly any research results appear to be available as far as ED hearing is concerned, one is tempted to rummage for information in the SD literature, because of the intrinsic link between these two spatial concepts. Unfortunately, as pointed out in Section 2.2.5, the identification of finer details of ER patterns that are (allegedly) salient to perceiving SD changes has often been based on the sole examination of physical models rather than on psychological verification. Results from the few psychoacoustic studies that have been conducted indicate that, for SD hearing at least, finer distance-dependent, binaural cues are only marginally important (see Section 2.2.5). Yet, these findings are contrasted by Kendall and Martens’ [1984] postulate that under natural listening conditions the exact spatio-temporal distribution of ERs plays a dominant role in the evaluation of apparent distance. However, no experimental results were presented to back up this statement. Thus, on the whole, it is unclear whether the information the hearing system deduces from such acoustical parameters is located at the top end of a hierarchy of perceptual distance/depth cues or not.

Be that as it may, the fact that control over perceived depth in a recording has been difficult to achieve so far strongly suggests that the underlying mechanisms are not sufficiently understood. This difficulty may partly be due to the multidimensional complexity of (1) the human response to spatial auditory stimulation and (2) the acoustic range (and hence depth) stimulus [Martens, 2001]. As is evident from Section 2.2, SD hearing depends on a multitude of independent physical parameters – even in the case of a static sound source. Since for ED perception at least two sources have to be dealt with, the modelling of such spatial changes is likely to be more complicated. Hence, it may well be that numerous psychophysical factors need to be closely concerted in order to allow for the perceivability of ED in a reproduced sound scene. Moreover, the perceptual integration of such cues with cognitive processes might have to be considered as well to be able to evoke the envisaged spatial impression. Finally, it also has to be borne in mind that ERs can affect several psychological dimensions simultaneously (see below), which is why one may have to make sure that these additional percepts conform with the overall ‘picture’. Clearly, this does not make a unidimensional ED simulation any easier.
Acoustical parameters and test conditions
Hoping to shed some light on some of the issues outlined above, a psychoacoustic experiment was
designed to determine the auditory effects of a number of finer details of ER patterns. More precisely,
the aim was to investigate various physical characteristics of sets of reflections to get an indication of
their impact not just on depth hearing, but also on any other subjective phenomenon that they might
give rise to. Since ERs have been related to several auditory constructs such as source distance (see
Section 2.2.5), source width (see Section 2.5), environment width (see Section 2.7) or timbre (e.g. see
[Beranek, 1996]), it was anticipated that the findings of this pilot test would prove useful in terms of
achieving an unambiguous ED simulation subsequently. Thus, in an unprecedented manner, the
following four ER pattern parameters (with their associated test conditions printed in *italics*) were
scrutinised:

1. Left-right time differences: *Yes* vs. *No*
2. Source-specific (as opposed to common) reflection patterns: *Yes* vs. *No*
3. Left-right azimuth differences: *Yes* vs. *No*
4. Azimuthal distribution: *Accurately panned* vs. *Folded into nearest loudspeaker*

Due to the fact that not all mathematically possible combinations of these four acoustical
characteristics are meaningful or even practicable (see below), only seven permutations or *test
materials* were investigated. These are shown in Table 5.15.

<table>
<thead>
<tr>
<th>Test material</th>
<th>L-R time differences</th>
<th>Source-specific ERs</th>
<th>Azimuthal distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Yes</td>
<td>Yes</td>
<td>Accurately panned</td>
</tr>
<tr>
<td>B₁</td>
<td>Yes</td>
<td>Yes</td>
<td>Folded into nearest loudspeaker</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test material</th>
<th>L-R time differences</th>
<th>Source-specific ERs</th>
<th>L-R azimuth differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>B₂</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>C</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>D</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>E</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>F</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>G</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

The above table has deliberately been split into two to highlight the difference between the concepts
of ‘Azimuthal distribution’ and ‘Left-right azimuth differences’ (see last column). Regarding the
upper half, the perceptual salience of accurately panning ERs was put to the test. It was hypothesised
that accurate encoding of the angles of incidence of ERs (test material ‘A’) would be indistinguishable
from rendering them from their nearest loudspeakers (test material ‘B₁’). With respect to the lower
half of the table, the auditory consequences of spatially correlating and decorrelating a sound field
were investigated by means of the 'Left-right azimuth differences' parameter. For this purpose it was
thought sufficient to just route each reflection to its closest speaker again (test materials 'B2' to 'G'). It
should be noted that test materials 'B1' and 'B2' were physically identical, but since they were utilised
to test slightly different hypotheses this differentiation seemed appropriate. Therefore, on the whole,
seven distinct test materials were generated.

In the context of finding suitable combinations of parameters, it is possibly worth pointing out that
accurate panning of ERs is feasible only if all ER patterns are source-specific. Consequently, the
number of test materials shown in Table 5.15 for which the panning of reflections would make sense
is limited anyway. Along a similar line, for those test materials that include left-right time differences,
the 'Yes vs. No' distinction for left-right azimuth differences becomes meaningless. That is why test
condition 'Yes' for the 'Left-right time differences' parameter was subjectively evaluated in
conjunction with 'Left-right azimuth differences – Yes' only.

In order to check whether the effects of these parameters are depth-dependent, three levels of ED were
scrutinised for each test material, i.e. 'flat', 'medium' (med) and 'deep' (see below). This resulted in a
total of 21 test materials, each of which was compared to all those other test materials that were
identical except for one parameter. Since the second part of this study into the spatial construct of ED
will be concerned with the detailed validation of the chosen simulation strategy (see Section 5.12), it
was deemed unnecessary to compare different degrees of ED to each other (such as 'A flat' with 'A
deep') at this stage. Thus, the time and effort required on behalf of the subjects to complete this test
were substantially reduced. Tabulation of all other possible combinations resulted in a total of 24 test
items. These are listed in Table 5.16 together with their associated test materials and ER pattern
characteristics.
Table 5.16: The 24 test items, their associated test materials and the corresponding physical ER characteristics. The one parameter varied in each case is printed in italics.

<table>
<thead>
<tr>
<th>Test item</th>
<th>Test materials</th>
<th>Early reflection parameter settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 flat</td>
<td>A + B₁ (flat)</td>
<td>L-R time differences – Yes</td>
</tr>
<tr>
<td>1 med</td>
<td>A + B₁ (med)</td>
<td>Source-specific ERs – Yes</td>
</tr>
<tr>
<td>1 deep</td>
<td>A + B₁ (deep)</td>
<td>Azim. distrib. – Accurately panned vs. Folded</td>
</tr>
<tr>
<td>2 flat</td>
<td>B₂ + C (flat)</td>
<td>L-R time differences – Yes</td>
</tr>
<tr>
<td>2 med</td>
<td>B₂ + C (med)</td>
<td>Source-specific ERs – Yes vs. No</td>
</tr>
<tr>
<td>2 deep</td>
<td>B₂ + C (deep)</td>
<td>L-R azimuth differences – Yes</td>
</tr>
<tr>
<td>3 flat</td>
<td>B₂ + D (flat)</td>
<td>L-R time differences – Yes vs. No</td>
</tr>
<tr>
<td>3 med</td>
<td>B₂ + D (med)</td>
<td>Source-specific ERs – Yes</td>
</tr>
<tr>
<td>3 deep</td>
<td>B₂ + D (deep)</td>
<td>L-R azimuth differences – Yes</td>
</tr>
<tr>
<td>4 flat</td>
<td>C + F (flat)</td>
<td>L-R time differences – Yes vs. No</td>
</tr>
<tr>
<td>4 med</td>
<td>C + F (med)</td>
<td>Source-specific ERs – No</td>
</tr>
<tr>
<td>4 deep</td>
<td>C + F (deep)</td>
<td>L-R azimuth differences – Yes</td>
</tr>
<tr>
<td>5 flat</td>
<td>D + E (flat)</td>
<td>L-R time differences – No</td>
</tr>
<tr>
<td>5 med</td>
<td>D + E (med)</td>
<td>Source-specific ERs – Yes</td>
</tr>
<tr>
<td>5 deep</td>
<td>D + E (deep)</td>
<td>L-R azimuth differences – Yes vs. No</td>
</tr>
<tr>
<td>6 flat</td>
<td>D + F (flat)</td>
<td>L-R time differences – No</td>
</tr>
<tr>
<td>6 med</td>
<td>D + F (med)</td>
<td>Source-specific ERs – Yes vs. No</td>
</tr>
<tr>
<td>6 deep</td>
<td>D + F (deep)</td>
<td>L-R azimuth differences – Yes</td>
</tr>
<tr>
<td>7 flat</td>
<td>E + G (flat)</td>
<td>L-R time differences – No</td>
</tr>
<tr>
<td>7 med</td>
<td>E + G (med)</td>
<td>Source-specific ERs – Yes vs. No</td>
</tr>
<tr>
<td>7 deep</td>
<td>E + G (deep)</td>
<td>L-R azimuth differences – No</td>
</tr>
<tr>
<td>8 flat</td>
<td>F + G (flat)</td>
<td>L-R time differences – No</td>
</tr>
<tr>
<td>8 med</td>
<td>F + G (med)</td>
<td>Source-specific ERs – No</td>
</tr>
<tr>
<td>8 deep</td>
<td>F + G (deep)</td>
<td>L-R azimuth differences – Yes vs. No</td>
</tr>
</tbody>
</table>

Creation of test materials

During preliminary experimental listening, the subtlety of many of the subjective effects caused by variation in the various physical characteristics had become apparent. Hence, to be able to expose subjects to these changes in an adequate fashion, it was decided to make the pilot test as critical as possible. Admittedly, this is likely to result in data more detailed than necessary for an ecologically valid ED simulation. Nonetheless, it was considered preferable to collect too much rather than too little information about the perceptual effects of each of these parameters. Therefore, the number of sources was restricted to the minimum needed for introducing depth changes into an ‘ensemble’ (i.e. two) in order to reduce auditory masking as much as possible. For the same reason, two contextually unrelated, anechoic speech recordings of a Danish woman and a French man were employed as the source material\(^\text{26}\). As has been pointed out before, human speech is a very critical test signal. Besides, due to its syllabic envelope characteristic it facilitates the detection of changes in various spatial qualities. This is because the background spatial impression [Griesinger, 1997] will be audible in-

\(^{26}\) Foreign languages were chosen so as to avoid distraction of subjects due to semantic content.
between the gaps of the DS, thereby enabling the auditory system to move from a loudness to a distance inference, for example.

All test materials were rendered using the processing platform illustrated in Section 5.9.2. As a starting point for the derivation of the various ER patterns an acoustic (image-source) model, produced with the help of CATT-Acoustic, was applied for calculating specular, horizontal-only reflections up to second order. The model was designed in such a way that the first reflection of each source would arrive as late as possible (t > 15ms) relative to the DS to minimise unwanted colouration [Pellegrini, 2002]. The obtained room impulse responses were then truncated at 50ms and the discarded reverberant energy replaced by diffuse reverberation. In order not to compromise the audibility of the ERs, the length and level of the reverb tail were kept suitably small. Nonetheless, care was taken not to depart too much from the acoustical properties of the model.

In the case of the 'flat' set-up, the two sources, separated laterally by 2m, were at an equal distance of ~2.2m from the receiver. For the 'medium' and 'deep' arrangements the male voice was moved backwards by 4m and 8m, respectively. Figure 5.34 depicts the CATT-Acoustic model in graphical form. Thus, the number of reflections arriving within the distance-salient time window of 15ms to 50ms [Theile, 2001] increased from five through to seven to ten as front-back separation of the two voices changed from 'flat' through to 'medium' to 'deep'. To allow for symmetrical ER pairs the impulse responses had to be modified slightly, i.e. the reflections adopted from the model were increased to six, eight and twelve, respectively, and then compressed into the chosen time window. As a consequence, the reflections were closer in time with respect to each other, but global differences (i.e. changes in the number and hence density of reflections) between impulse responses were maintained.

Figure 5.34: Plan view of the (2-dimensional) CATT-Acoustic model showing the positions of the receiver (01), the female voice (A1) and the male voice (B1, B2 and B3, respectively)
For test material ‘A’ the reflections and the DS were panned to directions in accordance with the CATT-Acoustic model. Azimuthal encoding was accomplished with the help of the fourth-order ambisonic panning technique, which was preferred over 5-channel PCPP because of its improved performance with respect to Gerzon’s design aims for multichannel panners. As explained in Section 5.9.3, these are based on a theoretical model for the psychoacoustics of directional hearing so as to satisfy as many as possible different auditory mechanisms and hence to achieve reliable, stable and natural sound image localisation. It therefore represented the best possible approximation to ‘accurate’ reproduction of the angles of incidence of the reflections within the constraints of the 3/2-stereo set-up. For all other test materials the reflections were simply fed to the nearest loudspeaker, whereas the DS would still be panned using the ambisonic technique to avoid timbral and image width changes (see Section 5.9.3). For test materials ‘D’, ‘E’, ‘F’ and ‘G’ left-right time differences were omitted by selecting representative delay times and making the reflections time-symmetrical with respect to the median plane. Left-right azimuth differences were either maintained (‘D’ and ‘F’) by feeding each pair of reflections to opposite loudspeakers (i.e. either L and RS or R and LS) or also removed by rendering them symmetrically relative to the medial plane. Finally, source-specificity was simply manipulated by retaining or eliminating any inter-source differences. With respect to the latter case, for the ‘medium’ and ‘deep’ source arrangements additional pairs of ERs were interpolated into the impulse response of the more distant voice. Hence, maximum possible ‘commonness’ of the ER patterns of the two voices was attained, whilst at the same time replicating distance-relevant changes in the D/R.

The levels of the reflections were essentially taken over from the CATT-Acoustic model. In particular, the first ERs were around −14dB relative to the DS for the ‘flat’ source arrangement. Later reflections would drop by up to 7dB as delay time approached 50ms. An increase in ED would cause all reflections to be attenuated by only a few dB while the level of the DS would follow the inverse distance law (see Section 2.2.3), thus resulting in a lower D/R for the ‘medium’ and especially the ‘deep’ layouts. For a given source set-up (i.e. either ‘flat’, ‘medium’ or ‘deep’) the number of reflections and their levels remained unchanged for all test materials in an attempt to keep overall loudness constant. Yet, due to the psychoacoustic phenomena of directional loudness perception and spatial unmasking (see Section 5.9.4), this does not preclude loudness changes from occurring because the angles of incidence of the ERs were varied. However, informal checks did not reveal any problems.

From the above description it is self-evident that many of the test materials featured highly artificial sets of reflections, which would not be encountered in a typical enclosure. Nevertheless, in using such extreme patterns the intention was to increase the likelihood that all potentially audible artefacts would be excited – even those that, to most listeners, would normally not be detectable.
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

Experimental design
The experiment took place in the listening room. As it was considered appropriate to first establish if a group of experienced listeners could consistently detect any difference at all between each of the 24 test items, the experimental design was based on the ABX subjective testing paradigm. This "double-blind triple-stimulus with hidden reference" method has been found to be especially sensitive and stable and to permit accurate detection of small degradations [ITU-R BS.1116, 1997]. Hence, it was deemed appropriate for assessing the auditory qualities of the various ER characteristics.

Regarding the test procedure itself, subjects were asked to compare three stimuli at a time and, knowing that only one of them was different, to specify the one being identical to the hidden reference. For each trial (or test item) and listener, two test materials were randomly assigned to three buttons labelled '1', '2' and 'X' on a graphical user-interface (see Figure 5.35).

The ABX test was followed by a form of Semantic Differential analysis (e.g. see [Osgood et al., 1975; Solomon, 1958]). The Semantic Differential is a scaling device that usually takes the form of a 7-point bipolar adjectival scale, but other forms are commonly used as well. These usually differ in terms of the number of points on the scale and the degree and type of labelling of these points.

During this second stage of the pilot study subjects were presented with 15 of the 24 test items again (see below) and asked to specify precisely (1) the types of differences that they experienced and (2) their relative magnitudes. For this purpose, a response sheet was provided for each test item containing a list of pre-selected attributes that had been adopted from the scene-based paradigm for spatial audio evaluation outlined in Section 1.3. 12 attributes describing spatial features at the source-, ensemble- and environment-level as well as a 'Timbre' and an 'Other' category were included. The subjects' task was to denote all subjective effects that they perceived to be varying for each test item.
If a particular effect was source-related, listeners were instructed to indicate which voice was affected. Also, they were asked to rate the 'magnitude of audibility' of each detectable change using the following scale:

\[
\begin{array}{ccccccccccc}
& & & & & & & & & & \\
& & & & & & & & & & \\
& & & & & & & & & & \\
& & & & & & & & & & \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\end{array}
\]

Extremely audible

The reason for using a 10-point as opposed to a 7-point rating scale was simply that the former had been employed in other studies, thus perhaps making some subjects feel more at ease with the task.

In addition to rating the magnitude of each auditory construct, listeners had to indicate its 'directionality', meaning that they delineated whether it appeared to be more pronounced in Sound 1 or Sound 2. The rationale for collecting this additional information was to get independent confirmation as to whether the presentation of a particular parameter resulted in an in- or decrease in a certain psychological dimension.

Listening panel
The listening panel comprised seven members of the IoSR who had all taken part in psychoacoustic studies before. All of them were conversant with the multidimensional concept of spatial quality and the attributes defined in Rumsey’s scene-based paradigm. Therefore, it was hoped that they would be able to successfully adopt the descriptors provided or at least to express their perception in a clear manner, should they find none of the pre-defined attributes to be suitable. Six subjects repeated the 24 ABX trials three times. The remaining listener completed six runs, resulting in a total of 30 observations per test item. To reduce listening fatigue the four/six sessions were spread over two days for most participants. The Semantic Differential test was conducted two days afterwards.

5.11.2 Results
Although seven listeners had participated in the test, it was decided to only submit the data of four of them to the statistical analyses. The reason for doing so was that the noisiness of the results was considerably reduced. It is appreciated that post-screening of listeners is generally considered bad practice. However, after the completion of each run-through all participants had given informal feedback as to the percentage of answers they expected to be correct\(^{27}\) and therefore how confident they were in their judgements. As it turns out, there was very good agreement between the actual number of correct answers achieved by these four listeners and their personal assessment of their performance. Besides, two of the rejected listeners openly questioned their own ability to distinguish reliably between many of the test items. Hence, for the sake of clarifying relationships between

---

\(^{27}\) The notion of a correct answer in a subjective test may seem a bit odd and hence demands clarification. Due to the working principle of the ABX paradigm it is possible to code a response as being right if a listener managed to identify the stimulus identical to the hidden reference. However, this does not mean that it would be categorically wrong not to be able to spot a (physical) difference, as it might be below the threshold of audibility, for example.
applied processing and resultant perception post-screening was felt justifiable. This left over a total of 18 observations per test item, which was still above the critical value of 16 scores that are needed for valid ABX testing [Meilgaard et al., 1991].

Analysis of ABX data
In Figure 5.36 95% confidence intervals have been plotted for the means of the aggregated responses of the four listeners for each test item. Note that wrong answers were coded as 0 and right ones as 1.

As can be seen, there are only a few test items for which listeners could consistently differentiate between the two associated test materials (i.e. '3 flat', '4 med', '5 deep', '6 med', '6 deep' and '8 deep'). For the remaining ones large confidence intervals are evident, which indicates that differences were not as clear-cut. To determine the statistical significance of these results a binomial test was executed. In particular, the following null and alternative hypotheses were put to the test:

\[ H_0: \text{The reproduction of parameter } X \text{ does not result in audible differences (i.e. the proportions of right and wrong answers are equal, } \pi_1 = \pi_2). \]

\[ H_a: \text{The reproduction of parameter } X \text{ results in audible differences (i.e. the proportions of right and wrong answers are not equal, } \pi_1 \neq \pi_2). \]

In this context, it should be noted that the post-screening did not change the results apart from making some psychoacoustic connections clearer.
where $X$ stands for any one of the physical characteristics outlined in Section 5.11.1 and $\pi_1$ and $\pi_2$ denote the population proportions for the two categories ‘right answer’ and ‘wrong answer’.

In Table 5.17 the percentages of correct answers obtained for each test item as well as the corresponding results of the (non-directional) binomial test are shown.

<table>
<thead>
<tr>
<th>Test item</th>
<th>Correct answers (%)</th>
<th>P-value (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 flat</td>
<td>72</td>
<td>0.096</td>
</tr>
<tr>
<td>1 med</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>1 deep</td>
<td>67</td>
<td>0.238</td>
</tr>
<tr>
<td>2 flat</td>
<td>89</td>
<td>0.001</td>
</tr>
<tr>
<td>2 med</td>
<td>88</td>
<td>0.001</td>
</tr>
<tr>
<td>2 deep</td>
<td>78</td>
<td>0.031</td>
</tr>
<tr>
<td>3 flat</td>
<td>94</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>3 med</td>
<td>72</td>
<td>0.096</td>
</tr>
<tr>
<td>3 deep</td>
<td>76</td>
<td>0.031</td>
</tr>
<tr>
<td>4 flat</td>
<td>72</td>
<td>0.096</td>
</tr>
<tr>
<td>4 med</td>
<td>100</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>4 deep</td>
<td>83</td>
<td>0.008</td>
</tr>
<tr>
<td>5 flat</td>
<td>89</td>
<td>0.001</td>
</tr>
<tr>
<td>5 med</td>
<td>78</td>
<td>0.031</td>
</tr>
<tr>
<td>5 deep</td>
<td>100</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6 flat</td>
<td>67</td>
<td>0.238</td>
</tr>
<tr>
<td>6 med</td>
<td>100</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>6 deep</td>
<td>94</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>7 flat</td>
<td>67</td>
<td>0.238</td>
</tr>
<tr>
<td>7 med</td>
<td>61</td>
<td>0.481</td>
</tr>
<tr>
<td>7 deep</td>
<td>67</td>
<td>0.238</td>
</tr>
<tr>
<td>8 flat</td>
<td>78</td>
<td>0.031</td>
</tr>
<tr>
<td>8 med</td>
<td>78</td>
<td>0.031</td>
</tr>
<tr>
<td>8 deep</td>
<td>100</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

By definition, the lower a $p$-value the stronger the evidence against the null hypothesis. More precisely, $p$-values equal to or less than 0.05 imply that there is a significant relationship between a given parameter and the result of the associated test item. $P$-values equal to or less than 0.01 indicate that there is a very significant relationship between the two. Hence, from Table 5.17 it can be inferred that listeners were unable to differentiate between the accurately panned ER scenario and the spatially quantised one, irrespective of the degree of ED (i.e. test items ‘1 flat’, ‘1 med’ and ‘1 deep’). Likewise, test items ‘7 flat’, ‘7 med’ and ‘7 deep’ appear to be indistinguishable, which suggests that if left-right time and azimuth differences are already absent the inclusion of source-specific impulse responses will go unnoticed.
From the results displayed in Table 5.17 it can also be deduced that the threshold for obtaining statistical significance lies at about 75%. This means that most of the other test items produced significant values (the exceptions being '3 med', '4 flat' and '6 flat'). Thus, the null hypothesis ('subjects made their judgements by chance') can be rejected at or beyond the 5% level of confidence for them. Yet, neither the percentages of correct answers nor the \( p \)-values allow any conclusions to be drawn about the substantive significance of these findings. That is to say that even though listeners were able to repeatedly detect differences between these test items, for some of them the noticeable changes might have been just above the threshold of audibility, whereas for others they could have been immediately obvious. Since this kind of situation will not be reflected in the ABX results, the responses from the Semantic Differential were analysed to specify their perceptual relevance in more detail.

**Analysis of Semantic Differential data**

As part of the second phase of this pilot test the subjects had been exposed to those 15 test items having \( p \)-values of less than 0.05 once more. As already mentioned above, their task had been to verbalise all perceived changes (using the list of attributes provided if possible) and then to grade their magnitudes of audibility. Due to the fact that considerable inter-listener differences were apparent with regard to the range of the scale that was used, all magnitude judgements were subjected to a z-score transformation. Thus, the responses of the individual subjects were normalised with respect to their means and standard deviations, whilst at the same time retaining the relationships of the original scores [ITU-R BS.1116, 1997].

The verbal descriptors were analysed by means of a basic VPA. As the input to the VPA consisted of a list of (mostly pre-specified) attributes, a sophisticated coding protocol was not needed. The descriptors were simply sorted into categories according to their labels and meanings. In addition, the sums of the associated magnitude judgements were calculated to give an overall perceptual weight factor for each category. Any inter-listener disagreement with respect to the directionality of a certain effect was taken into account by weighting the magnitude scores of any contradictory judgements by 0.5. For ease of interpretation, the obtained weight factors were then re-scaled, resulting in a maximum possible value of 10.

Unfortunately, this categorisation process does not produce a clear picture as to what exactly the audible effects of the various ER parameters are. Generally speaking, the responses are very diverse, leading to several groups of descriptors for all test items. These are presented in Appendix B. A plausible explanation for this may be that the listeners focused on and described different aspects of the auditory scene or that they found different terms suitable for expressing their perception of a certain subjective phenomenon. To give an example, while some listeners might have applied the term 'source direction' to denote a particular effect, others could have opted for 'source width' instead to delineate the same phenomenon. In view of the close interdependency of these two descriptive spatial
properties this seems possible. An alternative explanation might be that an increase in source width detected in one test material could have perhaps been expressed in the form of a source distance change for the other test material. After all, from visual perception humans are accustomed to the narrowing of an image being a corollary of a perceived increase in source distance. On these grounds, one might be tempted to reinterpret the meanings of some descriptors in order to lessen the number of categories and hence to improve the overall transparency of the results. However, due to the speculative nature of this process and the high risk of distorting the subjects’ wordings it was preferred not to tamper with the attribute labels.

Regardless of the diversity of the results, some general trends can be observed:

- The vast majority of changes noticed were spatial. These, in turn, were mostly source-related\(^2\), but the perceived environment was affected as well.
- SW changes were detected in all test items. Likewise, ‘source direction’ is only missing for ‘6 med’ and ‘environment width’ only for ‘2 med’. The three attributes seem to go hand-in-hand, e.g. for almost all ‘deep’ test items at least two of the three categories have a large weight factor (> 5).
- Manipulation of the ‘Left-right azimuth differences’ parameter caused the most obvious perceived differences. This is evident from the VPA categories of the associated test items (e.g. ‘5 deep’ and ‘8 deep’), as they have the highest weight factors.
- For ‘2 flat’, ‘2 med’ and ‘2 deep’ the weight factors (and occurrences) are generally low. Hence, source-specific impulse responses do not seem to make much difference if left-right time and azimuth differences are reproduced.
- Timbral differences were perceived occasionally and occurred either for the upper-mid frequency band or at low frequencies for the male voice. The latter effect always happened in tandem with a strong SW change (‘3 deep’, ‘5 flat’, ‘8 deep’).

Attempting to make more sense out of the data, it was decided to try to compute a measure illustrative at a higher analytical level. Thus, a ‘total difference’ coefficient was calculated for each test item, which involved the summation of all its magnitude judgements. Similar to before, inter-listener disagreement regarding the directionality of a particular change was considered. The resultant values were then divided by the number of categories derived from the VPA and re-scaled so as to have an upper limit of 10 again. Thus, a high coefficient either signifies good subject conformity as to what the audible differences are, large magnitude scores for the VPA categories in general or a combination of these two factors. The results are shown in Table 5.18 and contrasted with the \(p\)-values derived from the ABX experiment.

\(^2\) For all test items the male voice was affected. For ‘4 deep’, ‘5 deep’, ‘6 deep’ and ‘8 deep’ small changes in the direction of the female voice were also apparent.
Table 5.18: 'Total difference' coefficients for the 15 test items applied to the Semantic Differential and associated p-values

<table>
<thead>
<tr>
<th>Test Item</th>
<th>P-value (2-tailed)</th>
<th>'Total difference' coeff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 flat</td>
<td>0.001</td>
<td>3.3</td>
</tr>
<tr>
<td>2 med</td>
<td>0.001</td>
<td>3.9</td>
</tr>
<tr>
<td>2 deep</td>
<td>0.031</td>
<td>3.1</td>
</tr>
<tr>
<td>3 flat</td>
<td>&lt; 0.001</td>
<td>2.9</td>
</tr>
<tr>
<td>3 deep</td>
<td>0.031</td>
<td>6.1</td>
</tr>
<tr>
<td>4 med</td>
<td>&lt; 0.001</td>
<td>5.3</td>
</tr>
<tr>
<td>4 deep</td>
<td>0.008</td>
<td>7.7</td>
</tr>
<tr>
<td>5 flat</td>
<td>0.001</td>
<td>4.8</td>
</tr>
<tr>
<td>5 med</td>
<td>0.031</td>
<td>7.2</td>
</tr>
<tr>
<td>5 deep</td>
<td>&lt; 0.001</td>
<td>10</td>
</tr>
<tr>
<td>6 med</td>
<td>&lt; 0.001</td>
<td>5.1</td>
</tr>
<tr>
<td>6 deep</td>
<td>&lt; 0.001</td>
<td>9.2</td>
</tr>
<tr>
<td>8 flat</td>
<td>0.031</td>
<td>5.7</td>
</tr>
<tr>
<td>8 med</td>
<td>0.031</td>
<td>7.8</td>
</tr>
<tr>
<td>8 deep</td>
<td>&lt; 0.001</td>
<td>6.7</td>
</tr>
</tbody>
</table>

From a logical point of view, one would expect there to be a negative correlation between the magnitudes of the p-values and the ones of the 'total difference' coefficients, but this is not the case. The only meaningful observation that can be made is that (apart from test item '2 deep') an increase in ED will give rise to a higher coefficient. This does not come as a surprise since a lower D/R leads to more pronounced acoustic interference and decorrelation of the ear signals, which in turn might cause a (stronger) stimulation of (more) auditory dimensions.

Analysis of directionality judgements

Since structuring the verbal descriptors from the Semantic Differential had proven difficult already, summarising the directionality judgements in a meaningful manner was bound to be complicated, too. Nonetheless, some information about the audibility of the various subjective effects caused by varying a particular parameter may be obtainable.

For a given test item, the mode of the directionality judgements of all VPA categories was identified, i.e. either (Sound) 1 or (Sound) 2. Each judgement was then codified as 1 or 0, depending on whether it was equal to the mode or not. The resultant values were added up and divided by the total number of directionality judgements. This produces a kind of 'certainty factor' because the better the judgements concur as to which sound was overall more affected by the applied processing, the closer it will be to 1. A value of 0.5 denotes a uniform distribution of the directionality judgements, corresponding to 'maximal uncertainty'. Inevitably, though, this process will partly alter the responses. To illustrate, consider a test item for which an increase in source distance is noticed in Sound 1 and a source widening in Sound 2. Depending on the remaining directionality judgements, either of the two
verdicts will be forced into the 'nonconformity bin' and hence be coded as 0. Yet, although consideration of such attribute-dependent problems would have been possible, the downside would have been a great deal of guesswork again (see above). Fortunately, only a few such quandaries were encountered, which is why it was decided to compute the certainty factors as proposed above. Nevertheless, the reader may want to make up his/her own mind as to its expressiveness.

In Table 5.19 the obtained certainty factors are shown together with the modes of the associated directionality judgements for each test item.

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Certainty factor</th>
<th>Mode of directionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 flat</td>
<td>0.83</td>
<td>1</td>
</tr>
<tr>
<td>2 med</td>
<td>0.75</td>
<td>1</td>
</tr>
<tr>
<td>2 deep</td>
<td>0.6</td>
<td>2</td>
</tr>
<tr>
<td>3 flat</td>
<td>0.83</td>
<td>1</td>
</tr>
<tr>
<td>3 deep</td>
<td>0.6</td>
<td>2</td>
</tr>
<tr>
<td>4 med</td>
<td>0.5</td>
<td>1, 2</td>
</tr>
<tr>
<td>4 deep</td>
<td>0.88</td>
<td>2</td>
</tr>
<tr>
<td>5 flat</td>
<td>0.86</td>
<td>1</td>
</tr>
<tr>
<td>5 med</td>
<td>0.83</td>
<td>1</td>
</tr>
<tr>
<td>5 deep</td>
<td>0.91</td>
<td>1</td>
</tr>
<tr>
<td>6 med</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6 deep</td>
<td>0.57</td>
<td>2</td>
</tr>
<tr>
<td>8 flat</td>
<td>0.78</td>
<td>1</td>
</tr>
<tr>
<td>8 med</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8 deep</td>
<td>0.91</td>
<td>1</td>
</tr>
</tbody>
</table>

While it seems sensible to suppose that a certain parameter will stimulate auditory perception more strongly as ED increases (see above), a 'change of direction' occurring in parallel would be confusing. Yet, this situation is apparent from test items 2, 3, 4 and 6. What is more, each of them has one relatively small certainty factor (≤ 0.6). For test items 5 and 8, on the other hand, no such discrepancy is evident, i.e. the directionality judgements are consistent across test items. In light of the fact that (for the purpose of this analysis) the test material with the most 'Yes' test conditions (see Table 5.16) was rearranged to be Sound 1 for all test items, it is encouraging to see that for Sound 1 changes as a whole were perceived to be more pronounced. This makes sense because the sound field will have been more correlated for Sound 2, thereby reducing the audibility of certain effects.

Recalling that for test items 5 and 8 left-right azimuth differences were in- or excluded when generating the test materials, it may be concluded that this parameter results in the least ambiguous
changes of all the ones investigated. A similar trend was evident from the VPA categories (see above). With respect to the other parameters, it may be that while listeners were able to spot differences between the test items, they were unable to pin down their perception of these changes using verbal descriptors and directionality verdicts, possibly because they were hardly noticeable. This might be a reason for the inconsistency and diversity of these results.

5.11.3 Summary of pilot study
The aim of this pilot experiment was to determine the auditory effects of a number of physical characteristics of ER patterns as a function of ED. In particular, azimuthal distribution, left-right time and azimuth differences and source-specificity of ERs were scrutinised. All participants were experienced in the evaluation of spatial sound reproduction and yet they were struggling (1) to detect changes consistently and (2) to describe and rate these. This strongly suggests that many of the subjective effects caused by variation in the chosen parameters are very subtle. It is true that this cannot be deduced directly from the magnitude scores because of the relative scale that was used. However, conjoint analysis of all types of collected data (as well as informal feedback given by the listeners) indicates that this was the case.

As regards the elicited differences themselves, source width, source direction and environment width changes were heard most often and graded most strongly for almost all test items. Since they usually occurred simultaneously, it is likely that listeners perceived and/or described the same phenomenon in different ways.

It is well known that ERs influence width perception (see Section 2.5). According to the results of this study, of the parameters investigated left-right azimuth differences between ER patterns produce the most obvious changes in perceived width. Additionally, it was found that, regardless of the degree of ED, listeners were unable to distinguish between an ER pattern that comprised accurately panned reflections and one that was physically identical except that each reflection was reproduced by its nearest loudspeaker. This suggests that although spatial differences in ER patterns are perceptually salient, the actual angles of incidence of reflections may not be crucial.

In terms of the applicability and relevance of the results to simulating spatial dimensions, it is probably fair to say that their many-sidedness should be no cause for concern. After all, test conditions were made as critical as possible, which is why for applications that are ecologically more valid – such as the ensuing ED simulation (see below) – it is anticipated that far less artefacts will have to be dealt with.
5.11.4 Creation of ensemble depth stimuli

With the results of the pilot study in hand, the synthesis of appropriate ED training stimuli was approached. This required selecting suitable source material, conceiving a simulation strategy and applying the actual processing. Details for each of these steps are given below.

Source material

In terms of choosing appropriate source material, a number of issues were taken into consideration. An undesirable side-effect of using (contextually unrelated) speech recordings for the EW simulation had been that subjects did not conceive of the speakers as an 'ensemble' (see Section 5.10.2). Despite the fact that unidimensionality could still be achieved, it would be beneficial to this work if subjects not only perceived intra-source characteristics correctly, but also the sources’ 'roles' within their intended frame of reference. This should make the concepts and definitions of the associated attributes more intuitive, hence resulting in a more effective training procedure. Therefore, musical programme material was used for this experiment, hypothesising that this would lead to a more unitary result with regard to the sources' psychological relatedness.

To reduce stimulus complexity as well as processing cost an ensemble comprising four instruments only was envisaged. Yet, this was believed sufficient to evoke the cognitive cues necessary for producing the desired sensation. The ideal candidate for this job appeared to be a string quartet, as it seemed to exhibit all the desirable properties for an ensemble-level attribute simulation. Hence, a 4-bar recording (~8s in length) was used for this study, featuring the standard instrumental line-up of violin 1, violin 2, viola and cello. The instruments had been recorded separately under acoustically 'dry' conditions, thereby lending themselves to the superposition of a synthetically created acoustic context. Figure 5.37 displays the waveforms of the string quartet recording.

Figure 5.37: Waveforms of ‘string quartet’ source material
Simulation strategy

In principle, altering the distance of one component source at a time is sufficient to deepen/flatten an ensemble. However, to increase the likelihood that listeners perceive depth as opposed to just relative SD changes, it might be helpful to move all sources simultaneously. This assumption was based on the experience that when presented with complex stimuli, subjects tended to focus on whatever they perceived to change first rather than 'scanning' the whole sound scene for differences. Suffice it to say that this could lead to an oversimplified representation of the perceptual organisation of the stimuli to be generated and validated as part of this study. Therefore, to counter such auditory complacency the decision was made to have two pairs of sources being displaced from a straight-line (or 'flat') arrangement with one dyad (the outer sources) gradually getting closer relative to the listener and the other one (the inner sources) progressively getting more distant. Even though it was anticipated that this would increase the difficulty of the listeners' task, it was considered an appropriate measure to prevent the auditory system from undertaking a superficial analysis. At the same time, it was felt that the source grouping should enable listeners to detect the pattern more easily, which seemed less likely with all instruments moving independently. Figure 5.38 illustrates the applied notion graphically.

Figure 5.38: Graphical illustration of the concept of 'ensemble depth' as applied for this study

Applied processing

The ED stimuli were created in the listening room with the help of the platform described in Section 5.9.2. The first step in the generation of the sounds involved determining the maximally possible 'dynamic range' of the DS levels. That is, it was ascertained how quiet the inner instruments could be
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

made in relation to the outer ones whilst still being localisable. For this purpose, source material properties were exploited as much as possible, i.e. the first violin and the cello were placed in the centre of the ensemble, since they were melodically (violin 1) and rhythmically (cello) distinctive and thus less susceptible to being masked by the other two instruments. The resultant level difference (16dB and 19dB, respectively) was then subdivided into nine steps as required for the validation phase. As had been the case for the SD simulation (see Section 5.4.1), inhomogeneous treatment of the DS levels was required to ensure subjectively equal step sizes, i.e. the inner instruments' levels were reduced more as they receded into the distance.

In addition, changes were made in terms of the lateral positioning of the instruments to incorporate a perceptual cue in the simulation commonly utilised in 2-dimensional imagery. Converging lines in a 2-D drawing convey parallel lines and hence depth in three dimensions. This linear perspective phenomenon forces the brain to automatically infer a 3-D context on the basis of such information being contained in the 2-D input of the retina [Shepard, 2001]. Attempting to emulate this aspect aurally, a given DS was increasingly lateralised as the corresponding source appeared to get closer. To avoid any extra, unwanted qualitative changes, the ambisonic panner was used for the azimuthal encoding.

In order to provide the listeners with a supplementary distance/depth cue and hence to make the simulation more vivid, a first-order low-pass filter was inserted into the signal paths of the inner instruments. This had the effect of gradually rolling off the HF content of the direct sounds, reaching a maximum cut-off frequency of 8kHz for the deepest stimulus. Thus, the applied processing was broadly in line with the air absorption filter shown in Figure 2.1, whose magnitude response is ~3dB down at 10kHz for a source distance of 10m.

Regarding the design of the ER patterns, reflection levels were adjusted so as not to impair the previously established DS changes. In particular, it was found that for each instrument the ER levels had to follow the one of the associated DS closely. This can be explained with the help of the well-known precedence effect. Reflections arriving within about 50ms after the DS are perceptually combined to allow the human hearing system to localise a source in the direction of the first wavefront. As a result of the sound energy being integrated over this time window, the impression of added loudness arises [Everest, 2001]. Since under natural acoustic circumstances the DS obeys the $1/r$ law whereas the combined energy of all ERs decreases less than 6dB for the same (physical) distance change, the desired increase in (subjective) source range was maintained by reducing the early sound energy more or less in parallel with each DS. The numbers of ERs were chosen, so that the resultant listening conditions roughly resembled a real-world situation. For the closest sources, five reflections were reproduced within the distance-salient time window of 15ms to 50ms, whereas for the deepest stimulus the inner instruments exhibited 12 reflections. By and large, the temporal distribution of the reflections was uniform. As the pilot study had revealed that listeners are unable to distinguish between ERs panned in accordance with a room model and ones that are reproduced by
5. Unidimensional simulation and validation of spatial attributes of reproduced sound

their nearest loudspeakers, the latter approach was applied here for reasons of simplicity. However, overall directional characteristics of ER patterns were still replicated, i.e. the greater the apparent distance of a source, the more reflections were reproduced from in front of the listening position. It is true that no empirical evidence is available, which proves the perceptual salience of this parameter to depth hearing. Nonetheless, since such changes occur under normal listening conditions, their inclusion was expected not to be harmful to this simulation either. Similarly, it was decided to equip each source with a different reflection pattern. Depending on the reflection order the ERs had different frequency spectrums, the two biquad filters in Figure 5.21 being used as (second-order) low-pass filters with cut-off frequencies of ~4.5kHz and ~3.5kHz, respectively.

Finally, diffuse reverberation was added, whereby its level and duration (RT = 1.6s) were adjusted to create a room size impression that complemented the largest envisaged source distance. Depending on the source material, the chosen reverb levels were slightly different for the individual instruments. In particular, the more sustained cello part needed a few extra dB of reflected energy compared to the other sources with a fairly erratic amplitude envelope (see Figure 5.37) for the same SD effect due to its DS tending to mask the background stream information more. Further, minor intra-source adjustments of the reflected sound level were made as the stimuli got deeper. Although a diffuse sound field is characterised by approximately constant reverberant energy irrespective of measurement position [Everest, 2001], the reverb levels had to be balanced against the corresponding direct sounds in order to maintain adequate localisability.

Based on the methodology outlined above, nine reference stimuli were created in a controlled manner, each illustrating a different degree of ED. These were recorded to hard disk as 5-channel .aiff files with a sampling rate of 44.1kHz and a resolution of 16 bit.

5.12 Validation of ensemble depth stimuli I

The experimental design and physical set-up of the ED validation experiment – conducted to check if the intended unidimensionality had been achieved – were identical to the ones employed for the EW study (see Section 5.10.1). This time, 12 final-year students and one graduate of the ‘Tonmeister’ degree course were recruited to constitute the listening panel. All but two of them were acquainted with the experimental procedure by means of a short training session, which was very similar to the one used for the EW experiment (see Section 5.10).

5.12.1 Results

Analysis of similarity data

The scree plot derived from the MDS analysis executed on the ED similarity data is shown in Figure 5.39. As can be seen, the s-stress curve features a slight kink at 2-D, whereas no such indication is given by the stress parameter. Nevertheless, because SPSS optimises s-stress, one is inclined to decide
that more than two dimensions are not needed to model the structure contained in the similarity scores. In order to resolve whether the listeners had based their gradings on one or two discrete perceptual characteristics (and therefore to determine the appropriate dimensionality for the data set), the results for the RSQ measure need to be consulted again.

The RSQ values attained from the 1-D and 2-D MDS models can be found in Table 5.20. Evidently, the 1-D solution is characterised by a high RSQ value of 0.82, which decreases as one goes to a 2-D solution. This indicates that the MDS algorithm struggled to find a systematic structure in the data for a model with more than one dimension.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td>2</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Therefore, by collectively evaluating the different measures of fit one is tempted to deduce that the panel identified and employed one perceptual effect when comparing the nine sounds. Support for this impression is also available from the stimulus space (Figure 5.40). Knowing that ‘a’ is the flattest and ‘i’ the deepest stimulus, it can be seen that the sounds are arranged in the planned sequence. This signifies that they are sufficiently dissimilar in terms of a particular attribute, because the panel could rank them correctly.
All in all then, the MDS results seem to imply that the collected similarity verdicts contain the intended unidimensional rating structure.

Analysis of verbal data
The outcome of extracting holistic terms from the semantic responses is displayed in Table 5.21. Disappointingly, the obtained categories do not reflect the intended subjective effect, because references to changes in the width of the ensemble dominate the picture. Three out of the seven listeners who noticed this change used these very wordings to describe it, the remaining ones preferring the terms “Lateral position of instruments”, “Stereo distribution of instruments”, “Widening of stereo image: violin 2 and viola move outwards” and “Width of overall image”. It is true that the last response could also describe an environment- rather than an ensemble-related auditory aspect, but since the associated graphical data (see below) indicates left-right variations in the source positions, its inclusion in this category is justifiable. In spite of the fact that five listeners perceived ‘Ensemble width’ to be the most obvious subjective difference between the stimuli, an overall weight factor of 0.34 does not imply an unequivocal perceptual effect. The second holistic category concurs with the envisaged phenomenon more, i.e. five panellists detected changes in ‘Relative source distance’. Curiously, for one listener “Relative distance of instruments” was the strongest auditory effect. What is more, he also perceived the converse movement in almost the intended way, as evident from his graphical response (see below). Yet, with a total weight factor of 0.23 this group can hardly be called salient either. What is more, two other participants reported hearing this type of difference only for the cello or mainly for the central instruments, respectively. Parenthetically, only one listener used the descriptor “depth” in this context to delineate her perception of the stimuli. The other three categories are composed of terms that were either mentioned only once or that were given low ‘magnitude of
audibility' scores. In effect, one listener stated that the ensemble as a whole appeared to get closer or farther away ("Distance of quartet"), thus failing to work out the diametric movement of the two instrument dyads. Regarding the other bins, it is speculated that 'Room size' is related to the applied reverb processing, which may also be blameable for the 'Source width' category, because, as the pilot study showed, a decrease in the D/R will tend to de-focus a source and vice versa.

Table 5.21: Groups of holistic terms and their relative weights (ensemble depth, experiment 1)

<table>
<thead>
<tr>
<th>Holistic groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble width</td>
<td>7</td>
<td>0.34</td>
</tr>
<tr>
<td>Relative source distance</td>
<td>5</td>
<td>0.23</td>
</tr>
<tr>
<td>Ensemble distance</td>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>Room size</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>0.04</td>
</tr>
</tbody>
</table>

With respect to the analytical terms, five groups were identified, which are listed in Table 5.22. The first group ('Relative level of sources') constitutes the strongest and hence psychologically most significant one out of all the categories that were found for this experiment. Ten participants identified and used this difference for making their similarity judgements, four of which rated it as the perceptually dominant effect, resulting in an overall weight factor of 0.48. Incidentally, this value is fairly low compared to the highest ones obtained for the previous studies for which unidimensionality was accomplished, i.e. 0.87 in the case of source distance (see Table 5.2) or 0.82 in the case of ensemble width (see Table 5.13). It is also worth pointing out that six listeners perceived the level variations of the two pairs as being antipodal, while two subjects specified hearing balance changes only for the outer instruments and the other two exclusively for the cello.

Table 5.22: Groups of analytical terms and their relative weights (ensemble depth, experiment 1)

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative level of sources</td>
<td>10</td>
<td>0.48</td>
</tr>
<tr>
<td>Audibility of reverb</td>
<td>4</td>
<td>0.15</td>
</tr>
<tr>
<td>Attack/punchiness/edginess of notes</td>
<td>3</td>
<td>0.15</td>
</tr>
<tr>
<td>Reverb level of sources</td>
<td>2</td>
<td>0.05</td>
</tr>
<tr>
<td>Overall level</td>
<td>1</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The second group is similar to the first one in that it also seems to delineate the applied processing. It contains the terms "Audibility of reverb" (two mentions), "Level of reverb" and "Audibility of auditory space" (and therefore may be related to the holistic 'Room size' bin). However, it is
characterised by a small weight factor of 0.15, as is the third group, which comprises terms and phrases describing properties of individual notes ("Attack of notes", "Punchiness: violin 2 and viola sound as if played more/less violently", "Attack of violin quicker/more edgy for louder notes"). It is surmised that these perceptual effects go hand-in-hand with the first one, i.e. an increase in source level leads to more 'punchy' or 'edgy' notes. The last two categories ("Reverb level of sources" and "Overall level") are directly related to the stimulus creation process again, but because of their very low weight factors (0.05 and 0.02, respectively) this is no reason to be cheerful.

To summarise the findings of this section, as the analysis of the verbal data clearly showed the generated stimuli failed to impart the intended quality to the panel. First and foremost, it is striking that there is no single strong category even though this is exactly what one would have expected as a result of the MDS solution. Instead, the above discussion leads to the following inferences:

- Subjects picked out and concentrated on different characteristics of the sound excerpts - most notably the level and direction of the DS - when forming their verdicts, but none of the detected effects really prevailed.
- The various ED simulation 'ingredients' identified by the listeners varied in parallel, which is why a 1-dimensional MDS representation of the stimuli's auditory structure was possible.
- The additional, verbally elicited information was fundamental in laying open the perceptual deficiencies of the simulation. Without this type of data, an undistorted reading of the MDS results would not have been possible.

On the positive side of things, this experiment showed that the employed validation strategy is sensitive to unwanted dimensions; if they 'sneak into' the simulation, they will be revealed. Besides, it is satisfying to see that four listeners made use of the words "ensemble" or "quartet" when verbalising their perceptions, denoting that the choice of musical programme material was helpful in moving away from a 'several discrete sources' perception.

Analysis of graphical data

Since five listeners either had effectively stated that spatial differences had been secondary to their similarity verdicts or had not noticed any spatial changes at all, they were not asked to depict their perceptions visually. The eight graphical responses that were obtained (see Appendix C) were scrutinised to see whether the spatial information contained in the drawings agreed with the implied meanings of the verbal data. Interestingly, those five listeners who had graded EW changes most strongly drew the ensemble in such a way that the sources spread out on an arc maintaining approximately constant distance to the listening position. In other words, they subconsciously included depth changes in their graphical depictions. In Figure 5.41 one example of such a response is shown. When questioned about this, some subjects replied that they would expect a string quartet to
be arranged like this in a performance situation. The possibility of larger source spacings along the front-back axis, on the other hand, was ruled out in this respect. This suggests that consideration of cognitive aspects can also be harmful if the aim is to stimulate a specific impression. Due to their preconditioning subjects might unintentionally suppress any potentially noticeable deviations from their internal portrayal of a particular construct. Furthermore, a sixth listener also included a clear front-back separation of the instruments in his sketch, which was absent from his verbal responses. This example (once again) illustrates the usefulness of graphical elicitation, because it can help subjects explicate their perceptions, especially when being presented with a fairly-complicated set of stimuli like the one used in this case. Yet, even though the above observations make the results look a bit more promising, the variations in ED evident from the drawings are admittedly smaller than the ones in perceived ensemble width.

![Graphical response sheet displaying the visible surroundings and a typical listener response (ensemble depth, experiment 1)](image)

... but does it sound right?

Similar to the EW study, it had been planned to tell the subjects that the sounds were meant to vary only in the auditory feature that they had perceived as being the most prominent difference and to essentially ask them if they considered the simulation to be successful or not. This step had to be modified because of the considerable discrepancies between the envisaged and perceived result apparent from the verbal data of most listeners. Therefore, they were questioned about their listening strategies so as to find out what had handicapped depth detection on their behalf. As it turns out,
several listeners had based their similarity judgements solely on the behaviour of the two outer instruments, not paying any attention to the centre of the sound images. A possible explanation for this may be that the four string instruments were too similar in terms of their physical as well as musical characteristics (see Figure 5.37), thus making their perceptual segregation more difficult. Due to the fact that the outer sources were generally louder, they probably became the listeners’ reference points within each auditory scene. For the same reasons, it is assumed that subjects did not hear the reverberation enough, causing them to deduce level rather than distance changes. This would explain the predominance of verbal descriptors related to changes in the direct sounds (i.e. direction and level) over responses taking into account reflected sound energy.

Having obtained all this feedback, it was decided to create new stimuli. This was imperative because the 'string quartet' examples had turned out to be totally unsuitable for conveying the ED attribute to a group of experienced listeners in an unambiguous fashion.

5.13 Simulation of ensemble depth II

Rather than starting from scratch, the experience gained from the first experiment was capitalised upon by refining the ED simulation. This basically involved two steps: finding better source material and modifying the applied processing. As for the simulation strategy, it was decided to maintain the contrary SD changes together with the source groupings. Despite the first simulation effort being a failure, this feature was still believed to be beneficial for helping listeners identify the depth of the band being the variable for the same reasons as outlined in Section 5.11.4. The rationale for changing the source material and modifying the processing as well as the actual refinements are described below.

5.13.1 Creation of ensemble depth stimuli

Source material

Regarding the choice of suitable source signals, it was decided to stick with musical programme material as the string quartet stimuli had made the subjects listen in a perceptually more integrative manner compared to the speech sounds used for the EW simulation. Yet, in an attempt to improve the localisability of the individual sources and therefore depth changes, different types of instruments were picked this time to increase the sources’ dissimilarity with respect to their physical properties. In particular, an 8-bar recording (~14s in length) of a funk band featuring a tenor saxophone, a double bass and two electric guitars was employed. Again, the instruments had been recorded separately and contained very little reflected sound. By selecting a fairly unusual combination of instruments it was hoped that subjects would be less prejudiced in terms of the expected instrumental layout, thereby perhaps being more tolerant towards an unnaturally large front-back spacing. Special care was also taken that the musical arrangement had sufficient ‘space’, i.e. that the instruments were not all
constantly active to guarantee the audibility of reverberant decays. Figure 5.42 shows the waveforms of the ‘funk band’ recording.

![Figure 5.42: Waveforms of 'funk band' source material](image)

**Applied processing**

The procedure for generating the new sound examples was generally very similar to the one described in Section 5.11.4. A welcome side-effect of choosing miscellaneous instruments was that it allowed an increase in ‘dynamic range’ by at least 3dB per instrument. Consequently, the step sizes in the DS levels (and hence ED) could be made slightly bigger. Since the double bass and saxophone had the busiest parts, they were placed in the centre of the ensemble so as to draw the listeners’ attention to the distance changes whenever there would be a ‘gap’. Also, the DS step sizes of the guitars were slightly reduced for the three deepest stimuli, i.e. when the guitars came closest. This was meant to hinder them from being perceptually too prominent. It was feared that otherwise the loudness changes would dictate the listeners’ perceptions again and hence impede the detection of a depth dimension. Likewise, since subjects had turned out to be very receptive to lateral source position changes, variations in the directions of the direct sounds were minimal this time.

To further raise the panel’s awareness with respect to the variations in source proximity, the RT was increased to 2s. As a result, reflected energy was longer audible and hence more difficult to ‘miss’. Additionally, the reverb levels of the two guitars were gradually raised by ~2dB as they appeared to get closer. Otherwise, their direct sounds would have been psychologically overwhelming, causing the guitars’ estrangement with respect to the otherwise reverberant conditions. It is supposed that this was the case for the first experiment during which subjects had not noticed any range but only loudness changes for the string instruments moving towards the listening position.
Finally, the stimuli were fine-tuned by four expert listeners. Previously, this had only been done by this author and another researcher at the IoSR, but the problematic results from the first experiment and the large spread of responses commonly reported for absolute distance estimation studies [Nielsen, 1993] seemed to demand this measure. Hence, to take no chances the various sound field parameters described above were adjusted until a common ground had been reached between these four judges with regard to the subjective functioning of the envisaged ED notion.

5.14 Validation of ensemble depth stimuli II

Apart from the new stimuli, the second ED validation study differed from the first one only in the subjects that participated. Seven final-year ‘Tonmeister’ students, five members of the IoSR and one professional audio engineer – all experienced in subjectively assessing reproduced sound – served as the sensory panel for this test. 10 of them were trained with respect to the experimental procedure in the same way as before. Due to the fact that five of these 13 listeners had also taken part in the first ED test, one might argue that they were biased. However, it has to be borne in mind that after the completion of the first experiment listeners had not been informed about the aim to achieve a unidimensional ED simulation, because considerable differences between the perceived and intended effect had been found. Also, as the results of the first study clearly showed, the dominant effect detected by the panel was width- rather than depth-related. Therefore, in terms of subject preconditioning it is presumed that in the worst case these listeners may have expected the presentation of spatial sound stimuli.

5.14.1 Results

Analysis of similarity data

When scrutinising the scree plot displayed in Figure 5.43, the 2-D knee in the s-stress curve immediately catches the inspector’s eye. Compared to the kink evident in the s-stress curve from the first ED experiment (see Figure 5.39), it is much more pronounced. Moreover, the stress measure seems to be in better agreement with the s-stress metric this time, i.e. it also hints at a 2-dimensional solution being the appropriate one for the new set of responses. Yet, because a scree plot is unable to reveal a (sharper) knee at 1-D, these observations are of little use as far as verifying the unidimensionality of the ED simulation is concerned. At best, they can show that, for the same dimensionality, the MDS algorithm managed to find a clearer structure in these data compared to the ones from the first ED test.
The results for the explained variance can be found in Table 5.23. Regarding the 1-D solution, there is hardly any difference compared to the RSQ value obtained for the previous ED experiment (see Table 5.20), i.e. it has an almost equally large magnitude of 0.81. This time, though, augmenting the MDS model by a second dimension does not cause the RSQ to decline, but as it does not increase either no incentive is given for exploring its origin.

Table 5.23: RSQ results from non-metric MDS analysis (ensemble depth, experiment 2)

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>RSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Visualisation of the sound excerpts' psychological distances in the form of the 1-dimensional stimulus space (Figure 5.44) reveals that the samples are all sufficiently different from each other in terms of one subjective effect, because they appear in the envisaged order again. Curiously, the stimuli seem to be compressed at the flat (stimulus 'a') end of the simulation. Going back to Figure 5.40, the same tendency is apparent, which suggests that subjects found it generally harder to discern stimuli when all sources had roughly the same D/R. Further work would have to be done in order to trace back the root of this phenomenon.
Overall, the findings of the MDS analyses of the two sets of similarity judgements are very much alike, except that this time the badness-of-fit measures attest a more obvious rating pattern, as apparent from applying the ‘knee criterion’.

Analysis of verbal data
Classification of the verbal responses according to the VPA procedure employed previously turned out to be more difficult than before. This was because listeners sometimes reported the same qualitative phenomenon for different (groups of) instruments. To give an example from the holistic terms (see Table 5.24), some subjects deliberately distinguished between the distance/proximity of the inner/outer (groups of) instruments when asked to enumerate the perceptual differences they had used in forming their similarity verdicts. (This explains the 14 occurrences in the ‘Ensemble depth’ category, even though there were only 13 listeners.) This is likely to be a side-effect of the chosen simulation strategy, i.e. the opposite movement of the two source dyads. Irrespective of this intricacy, an ‘Ensemble depth’ group has emerged from the analysis that is characterised by a large weight factor of 0.84. All but one listener noticed changes pertinent to this category, whereby 10 of them gave them the highest score out of all their verbalisations. More precisely, the following terms were elicited from the panel: “Distance of sax: more distant when less audible” and “Positioning of guitars: closer when more bassy”, “Individual source distance: sax and bass vs. guitars”, “Distance of instruments”, “Closeness of guitars (particularly the left one)” and “Closeness of sax”, “Depth factor: especially bass and sax”, “Sense of perspective: the closer guitars, the farther away sax and bass”, “Relative distance of instruments: guitars vs. sax and bass”, “Relative perspective of sax and left guitar”, “Depth/relative front-back positioning of instruments”, “Relative closeness/distance of guitars compared to sax and bass”, “Positioning of instruments (distance/direction)”. It is satisfying to see
that four listeners, who perhaps perceived the stimuli in a more unitary way, used descriptors such as 
"Sense of perspective" or "Depth" that, in terms of the aims of this study, are clearly preferable to 
source-related wordings. Yet, as had been the case with regard to the EW descriptors, some of these 
phrases are not sufficiently detailed to allow deciding that they express the intended ED concept. 
Therefore, these verbalisations need to be examined in parallel with the other types of data again.

The next category has a familiar look to it. The single listener who had not perceived any ED changes 
heard alterations in the proximity of the ensemble as a whole. The other mentioning of 'Ensemble 
distance' stems from a participant who noticed this effect in addition to "Individual source distance: 
sax and bass vs. guitars" variations. In terms of physical sound field manipulation, the 'Source 
direction' group is clearly related to 'Ensemble width', which represented the strongest holistic group 
of the first ED experiment (see Table 5.21). It is assumed that because the panning was substantially 
reduced and hence less noticeable, listeners verbalised these changes as being source- rather than 
ensemble-related this time. Trailing in fourth place, the 'Togetherness vs. disjointness of ensemble' 
group is a bit awkward to categorise because of its more profound nature. However, it seems to 
circumscribe the envisaged effect in a holistic manner, which is why it has been included here. The 
last two groups are not easily accounted for, but on the grounds of their very small weight factors one 
can dismiss them as being perceptually insignificant.

The analytical groups are displayed in Table 5.25. Similar to the first set of ED stimuli, both intra- as 
well as inter-source level changes were detected again, but this time some listeners described level 
differences for the (outer) guitars and the (inner) saxophone separately. Consequently, as was the case 
with the holistic 'Ensemble depth' group, the associated 'Relative level of sources' category can be 
considered somewhat inflated. Nonetheless, both the occurrences and the weight factor are lower 
compared to the first experiment. Inter-listener disparities were also found again in terms of the 
instruments that were perceived to be influenced by this type of difference; while most listeners did 
not specify any particular instrument(s), one subject noticed it only for the saxophone and three others 
solely for the guitars. The second analytical bin comprises all terms concerning spectral modifications. 
Generally speaking, listeners either stated that the guitars were affected or they did not particularise
whether these differences related to the source-, ensemble- or environment-level. In terms of perceived frequency range, the descriptors "Brightness of individual instruments", "Timbre: dull vs. bright", "Bassiness of guitars" and "Close sounds were harsher" were elicited, suggesting that HF changes were most audible. As was the case with the first analytical group, the third category ('Reverb level of sources') had also been encountered during the first ED study. In terms of perceptual salience, it is slightly more important this time (presumably because the reverberation was more audible due to the longer RT and the new source material), as evident from its marginally higher weight factor. The same applies to the last group ('Overall level'), too.

Table 5.25: Groups of analytical terms and their relative weights (ensemble depth, experiment 2)

<table>
<thead>
<tr>
<th>Analytical groups</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative level of sources</td>
<td>8</td>
<td>0.38</td>
</tr>
<tr>
<td>Spectral changes</td>
<td>6</td>
<td>0.18</td>
</tr>
<tr>
<td>Reverb level of sources</td>
<td>3</td>
<td>0.09</td>
</tr>
<tr>
<td>Overall level</td>
<td>1</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Thus, from the point of view of having to argue the case for unidimensionality, the results are looking good. In this respect, it is also important to point out that the holistic descriptors "sense of perspective" and "depth" were elicited four times during this experiment. This compares favourably with the one mentioning of "depth" for the first ED test (see Section 5.12.1). Besides, two listeners clarified their responses by stating that they detected distance or depth changes only for those pairs of stimuli sufficiently different from each other. In contrast, for two very similar sound examples these types of (holistic) variations were not perceivable as such. Instead, the listeners relied on changes in the directions of the guitars and the reverb level of the saxophone to determine whether the two stimuli were identical or not. This seems to imply that at least some of the analytical terms may have arisen due to the small in/decrements in ED between adjacent stimuli. Put differently, there may be some kind of a perceptibility threshold for an (ostensibly) unidimensional percept like ED.

INDSCAL analysis

Basically, the observations made with respect to the SD and EW subject spaces also apply to the corresponding plot from the (second) ED experiment shown in Figure 5.45. All listeners are placed higher up on Dimension 1, which means that the associated perceptual factor was more salient to the subjects' similarity verdicts than any other potentially audible change(s). Evidently, this is in line with the previous finding that the panel noticed one (main) effect when comparing the stimuli. Further, when consulting the verbal responses for the assessment of the second INDSCAL dimension, there are strong indications that Dimension 2 of this plot is perceptually not meaningful. For example, according to Subject 9, the sound excerpts differ only in terms of one auditory difference, i.e. "Relative distance of instruments: guitars vs. sax and bass". However, the weight of this listener for
Dimension 2 is larger than the one of almost all other panellists who reported between two and six differences each. Again, many of these terms reflect the intended spatial percept or at least the applied processing. It can therefore be inferred that the panel as a whole identified and used only one dimension when comparing the stimuli.

Analysis of graphical data

Even though the visible surroundings had been depicted on the response sheet again to help the subjects draw to scale, clear inter-listener differences in overall ensemble depth are apparent from the graphical responses. As a rule of thumb, the magnitude of the SD variations was drawn to be unexpectedly small\(^3\), albeit much bigger than any other spatial attribute. Notwithstanding, as already stated in Section 5.10.2, absolute distance estimation is not a human strength, which is why this finding should not be given too much weight. More importantly, no additional systematic spatial change is evident from the responses, thus substantiating the apparently unidimensional structure of the similarity and verbal data. According to their graphical depictions, some listeners perceived the inner sources as being closer to the listening position for the (supposedly) flat extreme, while for other subjects it is the other way round. Similarly, some responses show more variation in the apparent range of the saxophone and (occasionally) the bass compared to the guitars, whereas for others it is the exact opposite. An example of such a graphical response is given in Figure 5.46, the remaining 12 drawings being contained in Appendix D.

\(^3\)Bearing in mind that a RT of ~2s and a ~20dB difference in DS level were used, this author perceived (and hence would have also anticipated) a maximal inner source distance well beyond the physical constraints of the listening room.
Further analysis of the sketches shows that, in terms of source locatedness, the double bass was problematic. To be more precise, two listeners did not perceive any changes in its position, while two others depicted its apparent location in the form of a large, fuzzy region. In the case of one response the bass is missing entirely. In retrospect, these results are not really surprising. Since the transient sound information (e.g. the attack portion of a plucked or rattling string) is likely to have been masked when the other instruments were playing, listeners probably only heard the LF content of the bass for most of the time. As a result, an impoverished set of localisation cues may have been available to the panel for this instrument.

... but does it sound right?

Having completed the similarity rating, verbal and graphical elicitation stages, the ED simulation was subjected to the final 'quality control' step. When asked to specify any deficiencies inherent in the set of stimuli as well as how the simulation would have to be improved, three listeners complained about the sporadic nature of the individual musical parts, stating that it made intra-source comparisons difficult. Another subject suggested that detection of the depth changes would have been easier if only one source had moved at a time. Both of these problems had been anticipated when creating the sound excerpts, but it was decided to put up with them for reasons outlined in Section 5.11.4 and Section 5.13. In any case, they do not exclude the achievement of unidimensionality. Two other comments implied that the bass was difficult to localise when distant and that its movement was minimal. Even though congruence in the movement of the bass and saxophone had been aimed for during the
stimulus generation process, it is not essential for being able to demonstrate the concept of ED. Since all but one listener perceived the saxophone and the guitars in the intended way, ED changes could still be conveyed. Therefore, it can be deduced that, by and large, the stimuli "sounded right".

To summarise the results from the ED validation, as the above discussion showed a single auditory effect, closely related to the intended ED variation, was perceived by all listeners when evaluating the stimuli. Although it is possible that additional perceptual variations were noticed on an intra-listener level, it can be stated with confidence that the panel as a whole did not identify a second perceptual dimension. Hence, it can be concluded that unidimensionality of the ED stimuli was achieved, which is why they should be suitable for training purposes.

5.15 Concluding remarks

From the validation experiments reported above it is evident that no type of collected data can disclose the dimensionality of the stimuli if examined in isolation. Due to the fact that subjects were not in any way restricted when verbalising their perceptions these responses are particularly problematic to interpret, albeit crucial to the outcome of these experiments. It is true that, like the dissimilarity judgements and graphical responses, the semantic data potentially document inter-listener similarities and differences. Yet, as was demonstrated for each spatial attribute simulation, the various elicited wordings appear to have a common underlying meaning with respect to the one perceptual difference identified by each listening panel.

The problem of eliciting and interpreting verbal information from subjects for the sake of attribute identification was also discussed by Berg and Rumsey [1999a; 2000a]. Referring to Shaw and Gaines' [1995] work, they stated that subjects may share only parts of their terminology and conceptual systems. Thus, listeners might use the same term for different concepts, different terms for the same concept, the same term for the same concept, or use different terms and have different concepts. These four possible scenarios are summed up in Figure 5.47. Evidently, the 'Correspondence' quadrant would be pertinent to the findings of the attribute simulation studies reported in this chapter. However, it is surmised that, following training with well-defined reference samples, subjects might move closer to 'Consensus', owing to a clear definition of verbal terminology and its relation to the stimuli.
5.16 Summary

This chapter provided a detailed description of almost all of the practical work that was executed for this thesis. For practical reasons, the four spatial attributes of source distance, source width, ensemble width and ensemble depth were selected and both simulated and validated with regard to their intended auditory effects. In this context, it was decided to employ the 3/2-stereo layout for reproduction purposes – its choice also being based on considerations of practicality.

For the simulation of source distance, it was found that the synthesis of a simple generic impulse response (consisting of three separate time regions that could be individually level-adjusted) gave sufficient control over this spatial percept. The finer structure of (six) ERs turned out not to be important for the creation of a number of (static) sources differing in terms of their positions along the front-back axis. As part of the following validation experiment, an experienced listening panel made pairwise comparisons of the generated stimuli in terms of their similarity. In addition, each subject verbalised all audible differences and indicated their magnitudes of audibility. MDS and VPA analyses of the collected data showed that the generated stimuli could exemplify the envisaged effect of source proximity changes to the listeners in an unequivocal manner. Therefore, it was concluded that the intended unidimensionality of the SD simulation was achieved, which is why the stimuli should make suitable reference samples and/or scale anchors.

Regarding the SW attribute, a more complex processing architecture had to be developed to enable the creation of several stimuli differing in terms of this perceptual construct. Essentially, the devised algorithm combines amplitude panning over L, C and R with the synthesis of early lateral reflections for manipulating the apparent width of a source. Besides, diffuse reverberation is used to keep room-related perceptual characteristics constant. Validation of the generated sounds proved to be very...
longwinded, which was mainly the result of them not being robust to different reproduction conditions. More precisely, the intended width effect was hardly perceivable when the stimuli were played back in another environment. It is speculated that electroacoustical differences between the Genelec loudspeakers (used in the listening room) and the ATC loudspeakers (installed in Studio 3) are responsible for this phenomenon. The directional characteristics of these loudspeakers are known to be different, but further research would have to be conducted to be able to confirm that this is the (primary) cause of the diverging subjective results (see also Section 6.2.2). Even though repeating the validation test in the location where the stimuli had been generated led to a 1-dimensional MDS representation, the fact that both source position and width changes were detected by a group of listeners gave reason for concern. One explanation for this finding may be that the SW simulation was only partially successful. It is also possible that positional variations were induced into the sound images by the subjects’ head movements, which were unavoidable because of the non-optimal placement of the computer monitor. Nonetheless, as part of a confirmatory experiment it was found that listeners were capable of adopting a definition of source width and of applying it for the ranking of a group of SW stimuli after they had been instructed to do so. This strongly suggests that the ambiguous responses were due to the listeners’ inexperience with evaluating the width of single sources. That is why, on the whole, the results were deemed acceptable to allow the stimuli to be used in training programmes. This decision was based on the view that source width is an unfamiliar construct, which subjects have confused before in other studies. It therefore may need to be predefined by expert listeners in order that non-experts can make sense of it.

As part of the simulation of the EW percept, a novel fourth-order ambisonic panning technique was analytically and subjectively assessed and integrated into a processing platform, which was implemented so as to facilitate the rendering of complex spatial sound scenes. This system enabled the synthesis of a set of reference stimuli containing a number of sources that maintained constant distance from the listening position, regardless of the degree of ensemble width. In order to ensure reliable verification of the stimuli’s perceptual organisation, the previously adopted validation strategy was refined. This was necessitated by the finding that MDS on its own cannot guarantee the unravelling of all perceptual factors contained in a set of similarity judgements. Additional verbal data were therefore gathered from a critical listening panel – not only to be able to interpret and label the continuous, orthogonal dimensions revealed by the MDS analysis, but also to discover qualitative factors varying in parallel to these. For this purpose, additional graphical responses were collected as well. Conjoint evaluation of the similarity ratings and the verbal and non-verbal data could overcome the deficiencies of MDS. In particular, the analyses showed that the generated stimuli could demonstrate the intended ‘constant distance EW’ effect to the subjects in an unambiguous manner. Crucially, the panel as a whole did not detect any additional qualitative factors. Thus, the accomplishment of unidimensional EW variation was deduced.

Intending to better understand the importance of early lateral reflections to ED hearing, a psychoacoustic study was conducted to determine the auditory effects of a number of physical
characteristics of ER patterns as a function of ED. In particular, azimuthal distribution, left-right time and azimuth differences and source-specificity of ERs were scrutinised. Results showed that many of the subjective effects caused by variation in the chosen acoustical parameters are very subtle, the least ambiguous ones being source width, source direction and environment width changes. These, in turn, were mainly due to the absence/presence of left-right azimuth differences in ER patterns. Although this finding suggests that spatial differences in ER patterns are perceptually salient, accurate panning of the reflections was found not to be important for the rendering of (stationary) ensembles exhibiting different degrees of ED. Based on these findings, a set of exemplary ED stimuli was produced. In this respect, the choice of source material turned out to have a great impact on the perceptibility of a depth component. More precisely, syllabic on both an intra- as well as inter-source level was needed to ensure the audibility of each instrument’s reflected sound. In contrast, controlled manipulation of the finer temporal and spatial structure of ERs was not required. Subsequent verification of the samples’ psychological organisation showed the proposed validation strategy to be sensitive to any unwanted (independent and correlated) perceptual variations contained in the simulation. Specifically, the listeners’ verbalisations were of utmost importance for laying open any potential problems, which the MDS analysis had failed to portray. Besides, the collection of additional graphical data proved beneficial again, as the drawings helped listeners explicate their perceptions. Graphical elicitation is therefore considered especially useful when investigating a complex (spatial) phenomenon like ED that is difficult to sum up in words. Collective assessment of all types of data disclosed that the generated sound excerpts could convey the envisaged ED effect to a group of critical listeners and that no extra, potentially confounding attributes were discovered by the panel as a whole. Hence, it was inferred that the ED stimuli vary in a unidimensional fashion, which is why they can be recommended for the use in spatial ear-training applications.
6 SUMMARY, CONCLUSIONS AND FURTHER WORK

This final chapter will provide chapter-specific summaries of the research and experimentation documented in this thesis, each including a reiteration of the main conclusions that have resulted from the associated work. Additionally, ways in which this project may be extended in the future will be delineated.

6.1 Summary and conclusions

6.1.1 Chapter 0
Intending to put the following work into its wider context, the opening chapter introduced the reader to the background of the research executed for this thesis. In particular, the need to evaluate 3-D audio systems in terms of their spatial quality by means of subjective testing methods was established. Moreover, a brief overview of training studies performed in the context of timbre perception was given, including evidence that systematic training can lead to listener improvements with respect to specific auditory skills. As for the area of spatial sound reproduction, the lack of an ear trainer was identified. Therefore, the development of such a system was proposed and important requirements regarding its characteristics were ascertained. In accordance with the sensory literature, it was concluded that:

- To guarantee an optimal training effect attribute simulations are required which are free from sensory interactions.

In addition, it was suggested that such simulations should give rise to a unified gestalt perception so as to maximise their intuitiveness and thus to enhance the learning process. Based on these two criteria, the concept of 'perceptual unidimensionality' was formulated. This led to the specification of the aims of this research, the main one being the accomplishment of validated, perceptually unidimensional simulations of spatial attributes of reproduced sound for training purposes.

6.1.2 Chapter 1
Chapter 1 summarised two independent attribute elicitation studies conducted in the area of spatial sound display. These confirmed the existence of several spatial dimensions, most of them concerning descriptive sound scene components such as the distance, depth or width of single or groups of sources. Consequently, the multidimensional nature of spatial quality was established, thereby setting a benchmark for the envisaged ear trainer, which ideally should address all dimensions salient to its domain. A scene-based paradigm, proposed by Rumsey for spatial audio evaluation purposes, was then presented, the design of which was based on these and other related findings. In light of the fact that this paradigm was specifically developed to overcome deficiencies evident in earlier research, it was concluded that:
• Rumsey's hierarchical system of spatial attributes of reproduced sound should be adopted as a conceptual framework for this work.

Furthermore, it was found that since a scale's adequacy as a predictor for listener preference is highly subject- and context-dependent, more research has to be carried out to be able to determine the relation of these attributes to listener preference. As a result, valid inferences with respect to the perceptual salience of spatial attributes of reproduced sound cannot be made as yet.

6.1.3 Chapter 2
In Chapter 2 established correlations between spatial attributes and physical factors were reviewed, expecting that this information would prove helpful for the unidimensional attribute simulations which were to be undertaken. It was found that to date several attributes have been scrutinised and that for each of them at least one perceptually salient acoustical parameter has been identified. Nonetheless, it became also evident that many of the underlying connections are not completely understood or possibly have not even been discovered yet. Therefore, it was concluded that:

• Whilst substantial amounts of research still have to be executed before it will be well understood which acoustical parameters relate to which spatial attributes, the numerous psychoacoustic mappings that have been determined should provide a good basis for further experimentation.

6.1.4 Chapter 3
Chapter 3 dealt with finding a validation strategy suitable for studying in detail the perceptual effects of sound-processing algorithms such as the ones devised as part of this work. Three potentially useful types of subjective evaluation paradigms, developed in the sensory sciences, were outlined and their respective benefits and drawbacks discussed. From this it was concluded that:

• To accomplish valid and reliable verification the most appropriate option for this work is to use a hybrid procedure, which utilises MDS to uncover the number of discretely perceivable attributes of a sound-processing algorithm as well as subject-dependent verbalisations as a means of labelling them in a meaningful way.

Moreover, it was argued that, unlike the other methods considered, MDS is an effective tool for determining and portraying psychological structures of low dimensionality. What is more, it does not specify an attribute elicitation and training protocol. For these reasons, it was concluded that:

• In view of the intended application, the suggested validation procedure is more efficient than the possible alternatives.
6. Summary, conclusions and further work

6.1.5 Chapter 4
In Chapter 4 existing perceptually motivated spatial sound processors were evaluated in terms of their appropriateness for varying spatial attributes unidimensionally. A review of the literature revealed that none of the devised systems has been (thoroughly) validated with regard to (all) its perceptual effects. In addition, several of these systems were found not to simulate spatial attributes of reproduced sound as applicable to this project, but to model room acoustical percepts, for example. Finally, instead of providing control over their respective qualitative changes ‘by the flick of a switch’, many of the investigated processors require the input of (textual) data and/or the adjustment of several faders to produce the envisaged effect. Although it was acknowledged that this aspect does not prohibit the unidimensional variation of attributes, it was emphasised that it risks the accomplishment of an intuitive and user-friendly training system. As a result of all these shortcomings, it was concluded that:

• No perception-based spatial sound processor exists that can satisfy all the requirements for unidimensional attribute simulations.

However, it was surmised that elements of some of these systems might be employable as the basis for the development of algorithms suitable for emulating spatial constructs.

6.1.6 Chapter 5
Chapter 5 focused on the experimental work conducted as part of this research. For reasons of practicality, it was decided to simulate and validate the spatial attributes of source distance, source width, ensemble width and ensemble depth for this thesis.

Regarding source distance (as well as the other attributes), several different intensity levels (or stimuli) were synthesised based on an iterative process involving informal listening and a clear knowledge of theoretical factors that have been found to influence this percept. With the help of a group of critical listeners and the hybrid procedure proposed previously the stimuli were then validated. Analysis of the MDS and verbal data showed that the panel as a whole only heard variations in the apparent closeness of a source. Therefore, it was concluded that:

• Unidimensional variation of source distance can be achieved by adjusting the relative levels of three separate time regions of a synthetic impulse response, without manipulation of finer ER characteristics (this has been published in [Neher et al., 2002]).

As to the attribute of source width, the production of a set of reference samples turned out to be more complicated, because only very small inter-stimulus width differences were achievable. Problems were also encountered with the ensuing verification, i.e. the synthesised SW changes were difficult to
detect in another reproduction environment. Therefore, the validation test was repeated in the location where the stimuli had been created, the analysis of the collected listener responses leading to a 1-dimensional MDS representation. Nonetheless, the elicited verbalisations were found to contain references to both source position and width changes. Whilst this could have been the result of the SW simulation being only partly successful, it is also possible that positional variations — as induced into the sound images by panellists’ head movements — were responsible for this finding. To resolve the uncertainty regarding the simulation’s auditory effect(s) a confirmatory experiment was performed, which showed listeners to be capable of adopting a definition of source width and of applying it for the ranking of a set of SW stimuli after they had been instructed to do so. This was interpreted as a strong sign that the ambiguous responses from the second validation test were due to the subjects’ inexperience with assessing the width of single sources. That is why, on the whole, the results were deemed acceptable, leading to the conclusion that:

- Unidimensional variation of source width can be achieved by using amplitude panning over L, C and R and artificially generated early lateral reflections to control perceived SW as well as diffuse reverberation to keep environment-related spatial percepts constant (this has been published in [Neher et al., 2002]).

Prior to the simulation of ensemble width, a processing platform had to be implemented so as to allow rendering of more complex, multiple-source sound scenes. In this context, a novel multichannel panner was evaluated based on analytical and subjective criteria and integrated into the devised platform subsequently. With the help of this system a set of reference stimuli was created, which were then verified by means of a refined version of the validation strategy adopted previously. More specifically, additional verbal and non-verbal data were gathered. These refinements were necessitated by the finding that:

- MDS on its own is unable to expose attributes varying in parallel with any of the common, continuous, orthogonal dimensions it discloses (this has been published in [Neher et al., 2003a]).

Besides, as had been the case for the SD and SW tests, the extra responses were used to interpret and assign meaningful labels to the MDS dimensions as well. Conjoint evaluation of all types of obtained data could overcome the flaws of MDS. In particular, the generated stimuli were found to impart the intended ‘constant distance EW’ effect to the subjects in an unmistakable fashion. Also, the panel as a whole did not notice any additional qualitative changes. Consequently, the conclusion was drawn that:

- Unidimensional variation of ensemble width can be achieved by applying a panner having smooth transitions and being timbrally neutral to control perceived EW as well as by balancing the levels of the DS and reverberation of each component source to
6. Summary, conclusions and further work

The emulation of the ensemble depth construct was preceded by a psychoacoustic investigation into the subjective effects caused by varying four acoustical parameters of ER patterns: azimuthal distribution, left-right time and azimuth differences and source-specificity of ERs. These were scrutinised as a function of ED. Analysis of the gathered data revealed that the resultant qualitative effects were generally subtle. Also, the responses were fairly diverse, listener agreement only being evident for the attributes of source width, source direction and environment width. These were mainly caused by in- or excluding left-right azimuth differences in ER patterns. Even though this denotes that spatial diversity in ER patterns is perceptually salient, accurate panning of reflections proved indistinguishable from spatially simplified reflection scenarios, irrespective of the level of ensemble depth. All in all, it therefore was concluded that:

- The finer structure of ER patterns is not crucial to ED hearing (this has been published in [Neher et al., 2003a]).

With this finding in mind a set of ED reference stimuli was synthesised. The subjective assessment which followed confirmed the (refined) validation strategy to be sensitive to the presence of any unwanted (independent and correlated) percepts in the ED simulation. Specifically, the subjects’ verbal and graphical responses were all-important to the revelation of audible changes, which the MDS algorithm had not brought to light. Collective analysis of all data showed that the devised simulation could exemplify the ED attribute to a critical listening panel who, importantly, did not detect any unintentional perceptual variations. Hence, it was concluded that:

- Unidimensional variation of ensemble depth can be achieved through the use of DS signals featuring syllabic amplitude envelope characteristics, which are not temporally synchronised between the sources to ensure the audibility of each component source’s reverberant sound energy (this will be published in [Neher et al., 2004]). Out of all the attributes dealt with, the ED simulation involved the most complicated processing, requiring the parallel adjustment of numerous parameters (details can be found in Section 5.11.4 and Section 5.13.1).

6.2 Further work

As the above summary of Chapter 5 showed, the main purpose of this investigation – i.e. the validated unidimensional variation of the SD, SW, EW and ED attributes – was achieved. Nonetheless, a number of tasks have arisen from these studies, which need to be completed before an optimal spatial ear-training toolkit will be available. These fall into one of five categories: (1) simulating and validating other salient dimensions of spatial sound reproduction, (2) investigating the simulations’
transferability to different reproduction environments and rendering systems, (3) quantifying the acoustical characteristics of the simulations more accurately, (4) developing an intuitive and effective ear-training system based on the simulations and (5) assessing the system's suitability for training applications. Below, each group of tasks is described in more detail.

6.2.1 Simulation and validation of other salient spatial attributes
In consideration of the findings of the elicitation studies outlined in Chapter 1, it is evident that the four constructs addressed by this work were taken out of a pool containing many more. Hence, the expansion of the 'arsenal' of reference materials is an important aspect to be looked at. This is because for a training tool to be optimal it has to cater for all salient dimensions of its domain, so that subjects can be exposed to all the perceptual characteristics which they might experience during sensory evaluation exercises. If a panellist's task is to grade a stimulus on a number of scales but (s)he is unsure how to accurately indicate his/her perception of a certain perceptual variation, it is very likely that the detected attribute will find its way into another grading of that stimulus. Such 'crosstalk' in direct attribute ratings is undesirable, as it can lead to a confounding of the resultant judgements and hence to a potential misinterpretation of the data [Stone & Sidel, 1993]. Thus, a good training programme should cover its sensory modality completely in order that subjects can be taught how to indicate their perceptions in an unambiguous and meaningful manner. Further research into the simulation of additional spatial attributes would therefore be required. In conformity with the principles of this work, such simulations would have to go hand-in-hand with empirically checking the achievement of unidimensionality in variation to be able to guarantee the best possible training effect. Reliable verification could be accomplished by means of the validation strategy devised as part of this project.

6.2.2 Transferability of attribute simulations
The transferability of the attribute simulations is another issue that is of major interest to this work. Ideally, different types of (high-quality) sound-reproducing equipment as well as (reasonably critical) listening environments should not affect the ways in which the simulations are perceived so as not to restrict their applicability. Unfortunately, this was found not to be the case for the SW stimuli as the intended effect was less audible when the samples were moved from Studio 3 to the listening room (see Section 5.7). In Section 5.16 it was surmised that this phenomenon was (mainly) due to differences between the loudspeakers used in the two monitoring environments, and information provided by the loudspeaker manufacturers lends support to this assumption. Notwithstanding, to allow full specification of the robustness of the SW stimuli to changes in the playback conditions, the factors responsible for the attenuation of the width effect would have to be identified. Evidently, this would require making (electro)acoustical measurements of the loudspeakers (and possibly the rooms as well). Such an investigation should then also be extended to the other sound excerpts. Likewise, it would be valuable to study the transferability of the generated simulations to different rendering systems. The simulations depend on the characteristics and limitations of 3/2-stereo (see Section 5.2),
and currently it is unknown how unidimensional variation of these attributes could be achieved over headphones or with other loudspeaker paradigms, for example. Whilst there are reasons to believe that binaural reproduction is unsuitable for the unambiguous rendering of some spatial attributes (see Section 5.2), an elaborate loudspeaker-based system like Wave Field Synthesis might be able to reproduce the stimuli without (spatial) distortion. Ultimately, this investigation could lead to the compilation of a 'spec sheet', pointing out any compatibility problems of the simulations with rooms, equipment and rendering systems to a potential user.

6.2.3 Acoustic analysis of created attribute simulations
As was mentioned before, the availability of stimuli varying along single, well-defined perceptual dimensions should help with establishing psychophysical relationships and is therefore also expected to be beneficial to the development of objective predictors of these attributes. That is why once the unidimensional psychological configuration of a set of stimuli has been ascertained, it is important to try to give a physical interpretation. This, of course, necessitates the accurate quantification of the acoustical characteristics of the generated signals. As explained in Chapter 5, the synthesis of the spatial attributes was based on knowledge of salient psychoacoustic relationships, trial-and-error experimentation as well as subjective evaluation (rather than on physical modelling). Hence, even though it is known how each algorithm was configured, the precise magnitudes of the various simulation parameters (or sub-attributes) still need to be determined. Objective measurements of the processing set-ups would therefore have to be made to allow full specification of all signal changes. To give an example from the SD emulation, the levels of the three time regions of the synthetic impulse response (see Section 5.4.1) would have to be quantified for each loudspeaker channel and for each of the nine stimuli. Continuation of this process for all the other simulation parameters would result in a comprehensive data set, which would enable another researcher to replicate the SD samples completely (provided that the same or at least very similar programme material would be used). Furthermore, it would be helpful to measure the corresponding sound fields at the listening position (e.g. by using a dummy head) so as to document interactions between reproduction channels as well as signal distortions due to the replay equipment and monitoring environments. Such detailed information could then be used for identifying systematic relationships between the acoustic properties of the stimuli and their locations in the 1-D perceptual space, thereby leading to a better understanding of correlations between the physical and psychological domains.

6.2.4 Development of an intuitive and effective training system
In Chapter 4 it was argued that for a training toolkit to be user-friendly and intuitive to use, it should be equipped with a high-level (perceptually motivated) interface. Furthermore, Section 5.3 noted that real-time adjustments of attributes are an effective way for subjects to become familiar with the associated qualitative phenomena. As for this project, some progress has been made with respect to the design of a graphical ear-training user-interface (see Figure E.1). In terms of real-time controllability, however, this work is still in its infancy as the current implementation only supports...
playback of the created stimuli, based on which a number of discrimination and ranking exercises can be completed by the user (see Appendix E for details). Real-time control would be valuable in that it would enable active (rather than just passive) training tasks to be incorporated into a training programme. Thus, instead of just listening to pre-specified attribute intensity levels subjects could be asked to manipulate the various spatial dimensions themselves. It is anticipated that this would facilitate listener training, especially with respect to more difficult tasks such as direct attribute ratings for which panellists would have to learn to scale the perceived magnitudes of discrete dimensions independently of other perceptual changes (see also Section E.1).

To enable real-time control and ease of use, it would be necessary to implement the SD and SW algorithms that this work has produced into a single processing environment. The rendering platform devised for the EW and ED simulations (see Section 5.9.2) should prove an adequate starting point for this purpose. Some modifications to its structure would be required, e.g. additional processing techniques (such as the divergence control employed for emulating SW) would have to be programmed. However, as a result of the simulations being integrated into software the need for specialised hardware would be eliminated, thereby making them available to a larger user base. Bearing in mind that for this thesis each attribute was dealt with individually, it might also be necessary to check if the created simulations can be combined in an orthogonal fashion, i.e. if the aim was to allow for simultaneous manipulation of the modelled attributes. From the descriptions of the stimulus creation processes given in Chapter 5, it seems that the simulations may not be truly compatible with each other. To give an example, indirect sound was needed for all four modelled percepts, either to be able to achieve the intended effect (i.e. for SD, ED and, to a lesser degree, SW) or to equalise unwanted spatial changes (i.e. for ‘constant distance EW’). Thus, conflicts in terms of reflection and reverberation requirements are likely to be encountered. That having been said, it might be possible to fine-tune the algorithms for some pre-selected source material, so that unidimensional variation of these attributes could still be feasible.

To further increase the versatility and effectiveness of the system, automatic diagnosis and analysis routines could be implemented. This idea is not new. The spectral ear trainer developed by Olive [2001], for instance, calculates on-the-fly quantitative metrics of a listener’s ability to identify and rate linear distortions added to programme material. Hence, panellists can be directly monitored in terms of their performance and, if necessary, be screened. Such data could also be exploited for self-administration purposes. It is well known that subjects can differ considerably in terms of their learning rates (see Section E.2), which is why a training programme has to be adjusted to each individual’s needs to ensure the best possible learning effect. Provided the software would be capable of ‘interpreting’ listener performance data, it could also be configured to optimise the training procedure accordingly, thus resulting in an ‘adaptive training environment’. In an ideal case, training sessions would be completely self-administered, so that the presence of the experimenter would become unnecessary, thereby leading to substantial savings in terms of time and hence cost.
6.2.5 Evaluation of suitability of system for training purposes

In reference to [Olive, 1995], in Section 0.3 it was stated that the investment in training only makes sense if (1) it can be implemented in an efficient and cost-effective manner and (2) it can be demonstrated to produce quantifiable improvements in the reliability of the listeners' ratings. Assuming that a spatial audio ear trainer with all the aforementioned features would eventually be available, a comprehensive assessment of the complete system's suitability for training purposes should follow. More precisely, such a study should verify the toolkit's ability to improve certain listening skills through quantification of the achievable training effect. For this purpose, a pre-/post-test paradigm could be employed to compare, say, listener performance and response time with respect to a particular exercise before and after the completion of some training sessions. Incidentally, in Appendix E a pilot study into the training of listeners for spatial audio evaluation will be presented, which made use of this methodology. More specifically, some of the sound excerpts created by this research were used for trying to increase the sensitivity of a group of naive listeners with respect to the associated attributes. Improvements were evident from the results, but because of the small number of participants no valid conclusions could be drawn. Thus, more research would be required not only to scrutinise the usefulness of the created simulations for training subjects, but also to show that the complete system has a positive effect on the discriminatory ability of listeners, for example.

6.3 Summary

This concluding chapter summed up the research and experimentation reported in the preceding chapters of this thesis. To start with, the background and motivation to this work were depicted and the multidimensional nature of spatial quality was established. In addition, the findings of psychoacoustic studies concerned with spatial attributes were presented, which was followed by a discussion of extant perception-based spatial sound processors. Consequently, a theoretical framework was instituted, based on which the simulation of the spatial attributes of source distance, source width, ensemble width and ensemble depth was tackled. For each attribute a processing algorithm was developed, aiming to provide unidimensional variation of the associated, well-defined perceptual phenomenon. In order to allow detailed testing of the resultant auditory changes, a validation strategy was designed, which was then used for deriving sensory profiles of the created algorithms. These showed that the source distance, ensemble width and ensemble depth simulations could unambiguously illustrate the envisaged effects to groups of experienced listeners who, crucially, did not hear any extra, probably confounding changes. The SW validation proved more difficult, which was traced back to listener unfamiliarity with and poor robustness of the psychological construct of source width. However, since only one major dimension was disclosed by the analyses, the results were deemed acceptable. Overall, it can therefore be concluded that unidimensional variation of spatial attributes of reproduced sound is possible and that it can be achieved (for four such constructs) by means of the synthesis techniques discussed in this thesis. Furthermore, reliable validation of such simulations is accomplishable with the help of a devised methodology that combines MDS techniques
6. Summary, conclusions and further work

with the elicitation of subject-dependent qualitative data. This chapter then finished with an outline of possible avenues for future research.
This appendix contains the 14 graphical responses that were elicited as part of the validation of the 'constant distance ensemble width' simulation (see Section 5.10.2 for details).
Appendix A: Graphical responses from ensemble width validation

Figure A.7: Graphical response from Listener 7 (ensemble width)

Figure A.8: Graphical response from Listener 8 (ensemble width)

Figure A.9: Graphical response from Listener 9 (ensemble width)

Figure A.10: Graphical response from Listener 10 (ensemble width)

Figure A.11: Graphical response from Listener 11 (ensemble width)

Figure A.12: Graphical response from Listener 12 (ensemble width)

Figure A.13: Graphical response from Listener 13 (ensemble width)

Figure A.14: Graphical response from Listener 14 (ensemble width)
APPENDIX B: ATTRIBUTE GROUPS FROM PILOT STUDY INTO EARLY REFLECTION PATTERN CHARACTERISTICS

This appendix contains the results of a Verbal Protocol Analysis applied to the data from the Semantic Differential for each of 15 test items scrutinised as part of the ED pilot study into the auditory effects of a number of ER pattern characteristics (see Section 5.11.2 for details).

Table B.1: Attribute groups, occurrences and weight factors obtained for '2 flat' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>Environmental envelop</td>
<td>1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Table B.2: Attribute groups, occurrences and weight factors obtained for '2 med' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source width</td>
<td>2</td>
<td>2.9</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>2.3</td>
</tr>
<tr>
<td>Source direction</td>
<td>1</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Table B.3: Attribute groups, occurrences and weight factors obtained for '2 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>2.3</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>2.0</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Source direction</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table B.4: Attribute groups, occurrences and weight factors obtained for '3 flat' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>2</td>
<td>2.1</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>1.7</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Timbre (upper mids)</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table B.5: Attribute groups, occurrences and weight factors obtained for '3 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source width</td>
<td>3</td>
<td>6.6</td>
</tr>
<tr>
<td>Source direction</td>
<td>3</td>
<td>6.0</td>
</tr>
<tr>
<td>Environment width</td>
<td>2</td>
<td>4.2</td>
</tr>
<tr>
<td>Environmental envelop</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td>Timbre (LF)</td>
<td>1</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Table B.6: Attribute groups, occurrences and weight factors obtained for '4 med' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>2</td>
<td>4.4</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>3.3</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>2.2</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Table B.7: Attribute groups, occurrences and weight factors obtained for '4 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>3</td>
<td>7.7</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>5.1</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>2.5</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table B.8: Attribute groups, occurrences and weight factors obtained for '5 flat' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source width</td>
<td>3</td>
<td>4.8</td>
</tr>
<tr>
<td>Source direction</td>
<td>1</td>
<td>2.7</td>
</tr>
<tr>
<td>Presence</td>
<td>1</td>
<td>2.3</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>Timbre (LF)</td>
<td>1</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table B.9: Attribute groups, occurrences and weight factors obtained for '5 med' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment width</td>
<td>2</td>
<td>4.7</td>
</tr>
<tr>
<td>Source direction</td>
<td>2</td>
<td>4.0</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Table B.10: Attribute groups, occurrences and weight factors obtained for '5 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment width</td>
<td>3</td>
<td>10.0</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>8.0</td>
</tr>
<tr>
<td>Source direction</td>
<td>3</td>
<td>6.2</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>3.8</td>
</tr>
<tr>
<td>Environmental envelopment</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td>Reverb</td>
<td>1</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Table B.11: Attribute groups, occurrences and weight factors obtained for '6 med' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source width</td>
<td>3</td>
<td>5.6</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>1.4</td>
</tr>
<tr>
<td>Timbre (upper mids)</td>
<td>1</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table B.12: Attribute groups, occurrences and weight factors obtained for '6 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source width</td>
<td>3</td>
<td>7.3</td>
</tr>
<tr>
<td>Source direction</td>
<td>2</td>
<td>4.0</td>
</tr>
<tr>
<td>Environment width</td>
<td>2</td>
<td>3.9</td>
</tr>
</tbody>
</table>
### Table B.13: Attribute groups, occurrences and weight factors obtained for '8 flat' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>4</td>
<td>5.0</td>
</tr>
<tr>
<td>Source distance</td>
<td>3</td>
<td>3.9</td>
</tr>
<tr>
<td>Source width</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>1.8</td>
</tr>
</tbody>
</table>

### Table B.14: Attribute groups, occurrences and weight factors obtained for '8 med' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>3</td>
<td>7.3</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>5.6</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>3.3</td>
</tr>
<tr>
<td>Timbre (upper mids)</td>
<td>1</td>
<td>2.9</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>2.3</td>
</tr>
</tbody>
</table>

### Table B.15: Attribute groups, occurrences and weight factors obtained for '8 deep' test item as part of pilot study into ER pattern characteristics

<table>
<thead>
<tr>
<th>Attribute group</th>
<th>Occurrences</th>
<th>Weight factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source direction</td>
<td>3</td>
<td>9.7</td>
</tr>
<tr>
<td>Source width</td>
<td>2</td>
<td>6.8</td>
</tr>
<tr>
<td>Timbre (LF)</td>
<td>2</td>
<td>5.5</td>
</tr>
<tr>
<td>Source distance</td>
<td>1</td>
<td>3.3</td>
</tr>
<tr>
<td>Environmental envelopment</td>
<td>1</td>
<td>3.1</td>
</tr>
<tr>
<td>Source envelopment</td>
<td>1</td>
<td>2.6</td>
</tr>
<tr>
<td>Environment width</td>
<td>1</td>
<td>2.5</td>
</tr>
</tbody>
</table>
APPENDIX C: GRAPHICAL RESPONSES FROM ENSEMBLE DEPTH VALIDATION I

This appendix contains the eight graphical responses that were elicited as part of the validation of the first ensemble depth simulation (see Section 5.12.1 for details).

Figure C.1: Graphical response from Listener 2 (ensemble depth, experiment 1)

Figure C.2: Graphical response from Listener 3 (ensemble depth, experiment 1)

Figure C.3: Graphical response from Listener 5 (ensemble depth, experiment 1)

Figure C.4: Graphical response from Listener 7 (ensemble depth, experiment 1)

Figure C.5: Graphical response from Listener 9 (ensemble depth, experiment 1)

Figure C.6: Graphical response from Listener 10 (ensemble depth, experiment 1)
Appendix C: Graphical responses from ensemble depth validation I

Figure C.7: Graphical response from Listener 11 (ensemble depth, experiment 1)

Figure C.8: Graphical response from Listener 13 (ensemble depth, experiment 1)
APPENDIX D: GRAPHICAL RESPONSES FROM ENSEMBLE DEPTH VALIDATION II

This appendix contains the 13 graphical responses that were elicited as part of the validation of the second ensemble depth simulation (see Section 5.14.1 for details).
Appendix D: Graphical responses from ensemble depth validation II

Figure D.7: Graphical response from Listener 7 (ensemble depth, experiment 2)

Figure D.8: Graphical response from Listener 8 (ensemble depth, experiment 2)

Figure D.9: Graphical response from Listener 9 (ensemble depth, experiment 2)

Figure D.10: Graphical response from Listener 10 (ensemble depth, experiment 2)

Figure D.11: Graphical response from Listener 11 (ensemble depth, experiment 2)

Figure D.12: Graphical response from Listener 12 (ensemble depth, experiment 2)

Figure D.13: Graphical response from Listener 13 (ensemble depth, experiment 2)
APPENDIX E: PILOT STUDY INTO TRAINING OF LISTENERS FOR SPATIAL AUDIO EVALUATION

As evident from Chapter 0, the idea to train humans for sensory evaluation purposes is not new and has been applied in a wide range of disciplines. In Section 0.2 it was pointed out that even though various audio researchers have successfully dealt with the training of listeners in timbre perception, a lack of such studies is apparent in the relatively new field of spatial sound reproduction. Hence, as part of this work a preliminary investigation into the training of listeners for spatial audio evaluation was conducted, which will be presented in the following sections.

E.1 Experimental design

For this pilot study a paid listening panel was set up, which consisted of five music students of the University of Surrey. All panellists were queried regarding their experience and skills in listening to sound in an analytical way. Since none of them had received any form of training in this respect, they were all considered naïve listeners. Nevertheless, two subjects reported an interest in music technology and sound recording, one of which had also participated in a listening test before.

To be able to quantify the effect of the adopted training method a pre-/post-test methodology was employed. Whilst all five participants completed the pre- and post-test, only three of them proceeded to the training stage that took place in-between. The other two listeners did not receive any training, thereby acting as a control group. Therefore, a comparison of the intra- and inter-listener performances and thus an assessment of the training effect was possible. The pre- and post-test were identical for each subject but varied across listeners to balance any interaction effects related to the order of presentation. The listeners' task was to rank five (SW) stimuli in terms of perceived source width and nine (SD) stimuli in terms of perceived source distance. Hence, each candidate's ability to discriminate several levels for each of these two spatial attributes was verified. The results of the pre- and post-test were then compared in terms of the correctness of the response. In addition, response time was taken into account when assessing the suitability of the chosen methodology. To allow comparability of the results, an effort was made to schedule the last training session and the post-test of each trained listener in such a way that the intermediate time period was about equal to the one between the pre- and post-tests for the untrained listeners. The time gaps ranged from three to five days.

Due to the preliminary nature of this study it was decided not to take the training any further than the ranking exercises outlined above. Yet, ideally, the training of a listening panel should go beyond simple detection of differences between sounds and the ordering of stimuli that differ in their intensities with respect to a specific perceptual property. In particular, an expert panel would also be able to quantify perceived differences in absolute and not just relative terms. Further, the ability to
discern and describe or categorise a particular auditory feature in the presence of other ones is a highly desirable characteristic for a good panel. However, it has to be borne in mind that the objective of this pilot study was simply to get an indication of the suitability of the adopted method for training a listening panel in spatial sound perception. To economise the procedure it was therefore decided not to address scaling and categorisation at this stage.

An essential aspect of any training programme is to provide a structured framework for learning in order to allow listeners to develop both skills and confidence [Meilgaard et al., 1991]. In the case of this study the programme was split into two parts, i.e. one for source distance and one for source width, the structures of which were very similar. Each training session was scheduled to last for no more than 30min, because it was felt that a longer duration would increase the likelihood of listener fatigue. During the first training session each listener had to listen to two sounds taken from the extreme ends of each attribute scale and to verbally describe the perceivable differences. The reasoning for doing so was to elicit terms that the listeners found suitable for describing their perceptions rather than to impose on them potentially non-meaningful wordings. This was found to be especially useful in the case of source width where it turned out that listeners preferred to use descriptors such as “Out of focus – Focused” or “Fuzzy – Defined” rather than “Wide – Narrow”. The second step of the training programme was 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. By drawing their attention to specific features of the sound that changed as a result of the applied processing, they were taught what to listen for when making their judgements.

After the completion of these two initial steps, the familiarisation phase of the programme began. This involved exposing all listeners to a large array of exemplary stimuli that could serve as a frame of reference. Depending on the spatial attribute, stimulus sets were used that contained two to three (for SW) or three to five (for SD) different intensity levels. Once the listeners felt they had a good grasp of a given spatial attribute, they were given a couple of training exercises where their ability to detect a difference between two sounds was tested. In general, sounds that exhibited large differences in intensity were selected, so that the panel could gain confidence as well as learn the basic listening skills. The familiarisation phase was then repeated until the subjects had listened to the full range of intensities of each attribute and successfully completed the associated exercises.

The true training phase built on the knowledge that the listeners had acquired during the previous stages. This time, they completed various exercises that were based on the following tasks:

1. Discrimination
2. Pairwise ranking
3. Multi-stimulus ranking
Appendix E: Pilot study into training of listeners for spatial audio evaluation

Each exercise consisted of 10 trials. If a panellist achieved 80% or more correct answers in a given exercise, (s)he was moved up one level in the training hierarchy. In case a listener gave four wrong answers the exercise was terminated and the listener had to go back and start again. If a subject failed to complete a particular exercise twice, (s)he had to go back a complete stage until listening skills and confidence were restored. Exercises were made progressively harder by (1) choosing a more difficult task, (2) making the differences in intensities between the stimuli smaller and (3) increasing the number of stimuli presented to the subjects in the case of the multi-stimulus ranking exercises. Once a subject had successfully completed the final multi-stimulus ranking exercise (i.e. the one with five SW or nine SD stimuli, respectively) the training programme was considered finished. Table E.1 and Table E.2 present details with respect to the number of sessions each listener needed as well as the total time period over which the training took place.

<table>
<thead>
<tr>
<th>Table E.1: Details of source distance training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table E.2: Details of source width training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

The whole training programme was implemented in MaxMSP. In addition to providing the processing framework for the training procedure, the software automatically collects and saves details of each exercise for further analysis purposes. This includes all user interactions with the application as well as the total time spent on completing the various tasks. A ‘performance’ window is included in the implementation, which displays whether the listeners’ responses are correct or not so that learning can take place. To strengthen the learning effect and to make the training software more fun to use, cartoon characters supply the user with instant aural and visual feedback. The ‘performance’ window also summarises the listeners’ achievements during a particular exercise, so that subjects know “how they are doing”. To supplement the training effect the auditory changes are also depicted in graphical form in the ‘viewer’ window during the familiarisation stage of the programme. Screenshots of some of the various software components are given in Figure E.1.
All pre- and post-tests (as well as training sessions) were run in Studio 3. The experimental preparations were identical to the ones of the SW confirmation experiment (see Section 5.8.2). The user-interfaces for the two ranking tasks were based on the one shown in Figure 5.18. Having been shown the SD and SW definitions used for the confirmation experiment, all listeners were thoroughly instructed in how to indicate their perceptions. In addition, they were informed of the importance of restricting head movements as much as possible. No limit regarding the maximum duration of the tests was set.

E.2 Results

As had been the case for the confirmation experiment, the SED was used to estimate the 'degree of wrongness' of each ranking sequence relative to the correct sequence. The SED results for the SD test from the trained (i.e. Listeners 1 to 3) and untrained (i.e. Listeners 4 and 5) group are shown in Figure E.2 and Figure E.3, respectively. Evidently, for both the trained and untrained subjects the post-test SED is lower, the exception being Listener 2 who ranked all nine sounds correctly during both tests. However, the untrained subjects obtained much larger SEDs than the trained subjects. Admittedly, the SED of both Listener 1 and Listener 4 reduced by the same amount (i.e. 6). Yet, Listener 1 managed to put the nine sounds into the correct order resulting in a SED of 0, whereas Listener 4 was still far away from that goal. Generally speaking, the trained group came very close to achieving a perfect match between the responses and the correct answer during the post-test.

31 Perhaps a word of caution is appropriate at this point. The SED has the disadvantage that the results from the SD and SW rank tests are not directly comparable to each other. This is because the former has a maximum SED of 240, whereas in the case of the latter the maximal SED is 40. While normalising the two sets of data could have easily rectified this problem, this was considered not beneficial because the original SED values directly reflect the closeness between the subjects' responses and the correct sequence. For instance, a SED of 2 corresponds to the reversal of an adjacent pair of sounds and so forth.
If one looks at the trained and untrained listeners' response times from the SD rank test (Figure E.4 and Figure E.5, respectively), a significant decrease in the time taken to complete the task is apparent for the trained group. For each listener response time dropped by more than 50%, whereas no such trend is detectable for the results from the untrained subjects. More precisely, Listener 5 almost had identical response times, whilst Listener 4 even needed slightly longer for the post-test. Interestingly, Listener 2, who ranked all sounds correctly from the start, took considerably more time than all other subjects to complete the pre-test and was still the slowest trained subject after the completion of the training programme. It may be argued that because of continued use of the software during the training sessions the trained group became more familiar with its operational aspects and therefore managed to execute the post-test faster. However, the magnitude of the reduction in response time seems too big to justify this improvement merely on the basis of the listeners learning how to use the software.
Looking at Figure E.6 and Figure E.7, the results from the SW ranking tests display similar trends to the ones identified for SD. Again, Listener 2 achieved a remarkably low SED during the pre-test, which remained the same for the post-test. Hence, no training effect can be inferred from his results. However, for Listeners 1 and 3 a clear drop in SED is evident from their post-test data, thereby pointing towards an effective training programme. This impression is reinforced by the fact that the results from the untrained group do not reveal any similar tendencies. For instance, Listener 2 (trained) and Listener 4 (untrained) obtained the same large SED during their pre-tests, but while Listener 1 performed much better during the post-test the performance of Listener 4 improved only slightly. Conversely, Listener 5 achieved the second lowest pre-test SED of all five listeners but obtained a worse result during the post-test. It is also worth pointing out that no listener achieved a SED of 0, even though only five SW sounds were used compared to nine stimuli for the distance attribute. This suggests that subjects found ranking the SW samples much harder than the SD sounds.
When examining the response times from the SW rank tests as a whole (Figure E.8 and Figure E.9), it can be seen that they were generally lower across all subjects compared to the tests on source distance, which was to be expected as there were fewer stimuli to compare. Notwithstanding, while all trained listeners improved considerably in terms of the time needed to rank the SD stimuli, this is not the case for SW. It is true that Listeners 2 and 3 managed to complete the SW post-test slightly faster, but not so Listener 1 who needed slightly longer. Likewise, Listener 4 was slower during the post-test as well, but in her case response time increased by about 3min compared to only 1min for Listener 1. Similar to the SD ranking test, Listener 5 achieved almost the same response time for the two tests on SW. It should also be noted that Listener 2 took longest again out of all trained participants to finish both pre- and post-test.
Overall then, slight decreases in the time needed to complete the task can be seen for the trained but not the untrained group. Hence, the results indicate that even though the training had a positive effect on the correctness of the subjects' responses, little improvements were made regarding speed. This seems to be in line with the finding made in Chapter 5 that the SW differences are harder to detect and judge than the ones for source distance, which probably led to the small changes in the subjects' response times even after the completion of several training sessions.

A final look at the training details presented in Table E.1 and Table E.2 provides further evidence. Both Listener 1 and 3 needed almost twice as many SW training sessions to reach a similar level of performance compared to SD. Watson [1980] addressed the aspect of training programme duration

---

The fact that Listener 2 only needed three sessions for both source distance and source width to achieve the desired level of performance warrants further commenting. During a conversation it turned out that in addition to his interest in music technology he also had nearly perfect pitch. Thus, it is speculated that Listener 2 had a stronger 'perceptual awareness', leading to superior performances during the pre-tests. On the other hand, the fact that he needed significantly longer to complete the tasks somewhat weakens this argument. At the least, his results show that subjects can differ considerably in terms of their learning rates – an aspect discussed in detail by Bech [1992].
in more detail and showed that the rate of learning of subjects is highly dependent on the complexity of the perceptual task, whereby an increase in complexity results in a prolonged training period.

Hence, the above supports the conclusion drawn from the results of the SW validation and confirmation experiments, namely that source width is an unfamiliar perceptual construct, which is much more difficult to assess than source distance. Nonetheless, the results from the training programme indicate that it should be possible to train listeners in its evaluation.

E.3 Summary

This appendix presented a pilot study into the training of a group of naïve listeners for spatial audio evaluation purposes. A hierarchy of computer-assisted training exercises was devised that were adjusted to each individual’s level of performance so as to achieve an optimal training effect. Likewise, the software was designed to give the user instant feedback to facilitate the learning process. By adopting a pre-/post-test paradigm in combination with a control group (i.e. a group of untrained listeners), results could be compared on both an intra- as well as inter-listener level, thereby enabling quantification of the training effect. For both tests subjects were asked to rank a number of SD and SW stimuli in terms of their perceived intensities. Results showed that the trained subjects improved in terms of the correctness of their responses for both SD and SW. Moreover, the response times of the trained subjects reduced significantly in the case of distance, but only slightly for source width. The difference between SD and SW response times was traced back to the higher level of stimulus uncertainty for the SW sounds, which had been identified during the associated validation experiments (see Section 5.8). For the control group no unanimous improvements were detectable. Thus, despite the fact that the small number of participants prohibits drawing any valid conclusions, the results suggest that the training of listeners in assessing spatial sound reproduction is possible.
### Glossary

This glossary section will define the meanings of the abbreviations and some of the specialised terms as used within the context of this thesis.

#### Abbreviations and terms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>.aiff</td>
<td>Audio Interchange File Format. A common <em>Apple Macintosh</em> file format for storing uncompressed audio.</td>
</tr>
<tr>
<td>AES</td>
<td><em>Audio Engineering Society</em></td>
</tr>
<tr>
<td>ASW</td>
<td>Apparent or auditory source width. A spatial concept stemming from concert hall acoustics research, which commonly refers to the phenomenon that a sound source is perceived as being wider than its visual contours suggest (see Section 2.5 for more information).</td>
</tr>
<tr>
<td>Attribute</td>
<td>A perceived characteristic of a stimulus that can be employed to describe it, e.g. pitch and loudness are two well-known auditory attributes of a sound stimulus.</td>
</tr>
<tr>
<td>Attribute scale</td>
<td>An experimental device for measuring the perceived intensity of a certain sensory characteristic (or attribute) in quantitative form.</td>
</tr>
<tr>
<td>C</td>
<td>Centre loudspeaker of the 3/2-stereo layout (see Figure 0.1).</td>
</tr>
<tr>
<td>CA</td>
<td>Cluster Analysis. A form of non-verbal scaling for classifying a set of stimuli into a (smaller) number of meaningful subgroups (or clusters) based on judgements of similarities among the stimuli. These clusters then have to be interpreted (see Section 3.3.1 for more information).</td>
</tr>
<tr>
<td>CPU</td>
<td>Central processing unit. The control unit of a computer which performs logical and arithmetic operations.</td>
</tr>
<tr>
<td>DA</td>
<td>Descriptive analysis. A class of sensory tests in which a trained panel rates specific attributes of a stimulus on scales of perceived intensity [Bech, 1999] (see Section 3.1 for more information).</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>D/R</td>
<td>Direct-to-reverberant sound ratio. A measurement of the proportion of direct to reflected sound energy at a particular receiver location, usually expressed in dB.</td>
</tr>
<tr>
<td>DS</td>
<td>Direct sound. Sound which travels directly from source to receiver without undergoing reflection.</td>
</tr>
<tr>
<td>DSP</td>
<td>Digital signal processing. The manipulation of signals in the digital domain so as to analyse, modify or extract information from them.</td>
</tr>
<tr>
<td>EBU</td>
<td>European Broadcasting Union</td>
</tr>
<tr>
<td>Ecological validity</td>
<td>Describes the extent to which an experimental situation compares to the context and circumstances it is supposed to represent [Rumsey, 2002]. For product evaluation purposes it is therefore more appropriate to use music, speech etc. rather than conventional test signals.</td>
</tr>
<tr>
<td>ED</td>
<td>Ensemble depth. An ensemble-level attribute from Rumsey's scene-based paradigm (see Section 1.3 for more information).</td>
</tr>
<tr>
<td>ER(s)</td>
<td>Early reflection(s). The sound reflected from nearby surfaces that arrives at the receiver within the first few (usually ~50) milliseconds of the DS radiated by a source. In concert hall acoustics this early time period has traditionally been extended to 80ms. The (discrete) ERs are normally followed by diffuse/late reverberation.</td>
</tr>
<tr>
<td>EW</td>
<td>Ensemble width. An ensemble-level attribute from Rumsey's scene-based paradigm (see Section 1.3 for more information).</td>
</tr>
<tr>
<td>FCP</td>
<td>Free-Choice Profiling. A particular method for the elicitation of individualised attribute scales that does not specify an elicitation protocol (see Section 3.2.1 for more information).</td>
</tr>
<tr>
<td>HF</td>
<td>High frequency</td>
</tr>
<tr>
<td><strong>Glossary</strong></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>HRTFs</strong></td>
<td>Head-related transfer functions. Fourier-transformed, head-related (binaural) impulse responses, which fully describe the modifications undergone by a sound wave propagating from a source in the free field to some point within the ear canals for the case of a stationary source and receiver [Blauert, 1997].</td>
</tr>
<tr>
<td><strong>IEC</strong></td>
<td><em>International Electrotechnical Commission</em></td>
</tr>
<tr>
<td><strong>IHL</strong></td>
<td>In-the-head locatedness. A phenomenon commonly encountered with deficient binaural sound reproduction where the auditory event is not perceived as external (see Section 2.4 for more information).</td>
</tr>
<tr>
<td><strong>ILD(s)</strong></td>
<td>Interaural level difference(s). For a non-medial source a level difference between the ear-input signals will occur at mid to high frequencies due to the contralateral ear signal being attenuated as a result of head-shadowing effects.</td>
</tr>
<tr>
<td><strong>Independence (of attributes)</strong></td>
<td>Attributes are said to be independent if they do not interact [Rumsey, 1998]. The degree of independence between attribute ratings can be assessed using statistical techniques such as correlation analysis, wherein low correlation between attributes is indicative of greater independence.</td>
</tr>
<tr>
<td><strong>INDSCAL</strong></td>
<td>INdividual Differences SCALing. A derivative of MDS which models not just inter-subject agreement but also disagreement (see Section 5.5.2 for more information).</td>
</tr>
<tr>
<td><strong>IoSR</strong></td>
<td><em>Institute of Sound Recording</em>. Part of the Department of Music &amp; Sound Recording at the University of Surrey, hosting research and teaching in audio-related subject areas.</td>
</tr>
<tr>
<td><strong>IRCAM</strong></td>
<td><em>Institut de Recherche et Coordination Acoustique/Musique</em></td>
</tr>
<tr>
<td><strong>ISO</strong></td>
<td><em>International Organization for Standardization</em></td>
</tr>
<tr>
<td><strong>ITD</strong></td>
<td>Interaural time difference. For a non-medial source a time difference between the ear-input signals will occur due to the</td>
</tr>
</tbody>
</table>
emitted sound wave having to travel farther (i.e. around the head) to reach the contralateral ear.

**ITU**

International Telecommunication Union

**L**

Left loudspeaker of the 3/2-stereo layout (see Figure 0.1).

**LEV**

Listener envelopment. A spatial concept stemming from concert hall acoustics research, which commonly refers to the subjective impression of being enveloped by reverberant sound in a hall (see Section 2.8 for more information).

**LF**

Low frequency

**Listening room**

ITU-R BS.1116-compatible listening environment at the IoSR.

**LS**

Left-surround loudspeaker of the 3/2-stereo layout (see Figure 0.1).

**MDS**

Multidimensional Scaling. A form of non-verbal scaling, which transforms judgements of similarity of a set of stimuli into Euclidean distances represented in X-dimensional (Euclidean) space. All stimuli are then placed in such a way that best reflects these distances. The resultant perceptual map shows the relative location of all stimuli along the X common, continuous, orthogonal dimensions uncovered from the similarity data, which then have to be interpreted (see Section 3.3.2 for more information).

**med**

medium

**M-S**

Mid-Side. Invented by Blumlein [1931], the M-S (or sum-and-difference) concept can be seen as an extension to conventional 2-channel stereo, whereby M is the mono sum of the two stereo channels and S is the difference between them.

**MUSHRA**

MUlti Stimulus test with Hidden Reference and Anchor (see [ITU-R BS.1534, 2001] for more information).
<table>
<thead>
<tr>
<th><strong>Glossary</strong></th>
<th><strong>Definition</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthogonality (of attributes)</td>
<td>Attributes that are truly independent are said to be orthogonal and can be represented as orthogonal dimensions on a multi-dimensional scalogram [Rumsey, 1998].</td>
</tr>
<tr>
<td>PCPP</td>
<td>Pairwise constant power panning. The well-known sine/cosine panning law for 2-channel stereo, which is often simply extended to more loudspeakers for multichannel set-ups. For a given pair of channels the gain of one channel varies as the sine of the panning angle $\theta$ and the gain of the other varies as the cosine of $\theta$, thereby maintaining constant total power. The term 'pairwise' denotes that non-zero gain is applied to the two speakers adjacent to the phantom image location only.</td>
</tr>
<tr>
<td>R</td>
<td>Right loudspeaker of the 3/2-stereo layout (see Figure 0.1).</td>
</tr>
<tr>
<td>RGT</td>
<td>Repertory Grid Technique. A particular method for the elicitation of individualised attribute scales that specifies an elicitation protocol (see Section 3.2.2 for more information).</td>
</tr>
<tr>
<td>RS</td>
<td>Right-surround loudspeaker of the 3/2-stereo layout (see Figure 0.1).</td>
</tr>
<tr>
<td>RT</td>
<td>Reverberation time. The time in seconds for the reverberant sound in a room to decay 60dB after its source is abruptly turned off. This corresponds to a decrease in intensity to one millionth of its original value.</td>
</tr>
<tr>
<td>SD</td>
<td>Source distance. A source-level attribute from Rumsey’s scene-based paradigm (see Section 1.3 for more information).</td>
</tr>
<tr>
<td>SED</td>
<td>Squared Euclidean distance. The square of the Euclidean distance between two points located in Euclidean space (see Section 5.8.2 for more information).</td>
</tr>
<tr>
<td>Sensory evaluation</td>
<td>A scientific discipline used to evoke, measure, analyse and interpret subjects' reactions to stimuli based on the senses [Bech, 1999].</td>
</tr>
</tbody>
</table>
SPL

Sound pressure level. A sound pressure expressed in dB relative to the reference sound pressure of 20μPa.

Stimulus

A perceived object, e.g. a sound field listened to.

Studio 3

Multichannel control room at the Department of Music & Sound Recording at the University of Surrey (see Section 5.4.1 and Section 5.8 for more information).

SW

Source width. A source-level attribute from Rumsey's scene-based paradigm (see Section 1.3 for more information).

Unidimensionality (of attributes)

An attribute is said to be unidimensional if it describes a unique sensory characteristic in an unambiguous fashion, i.e. if it represents a single perceptual construct [Rumsey, 2002].

Unidimensionality (in variation)

A group of (multidimensional) stimuli are said to be unidimensional in variation if it can be empirically demonstrated that they differ only in terms of their intensities with respect to one particular, well-defined perceptual effect.

VPA

Verbal Protocol Analysis. A research tool commonly used to analyse and quantify the presence, meanings and relationships of certain words and concepts within a given text or set of texts so as to be able to make inferences about the messages contained in the text(s), for example (see Section 3.2.1 for more information).

VSP

Virtual Surround Panning. A collection of signal-processing techniques developed by Studer for their D950S digital mixing console (see Section 4.4 for more information).
REFERENCES


References


References


References


References


Martens, W. L. (1999): ‘The impact of decorrelated low-frequency reproduction on auditory spatial imagery: Are two subwoofers better than one?’, *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, April 10-12, pp. 67-77


Mason, R. (2002b): Personal communication with the author.


References


