Modelling perceived spatial attributes of reproduced sound

Martin Dewhirst

Department of Music and Sound Recording
Faculty of Arts and Human Sciences
University of Surrey

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

April 2008

Thesis submitted in the requirement of the degree of
Doctor of Philosophy

© University of Surrey 2008
This thesis has described the development of an artificial listener model capable of predicting a number of different perceived spatial attributes at arbitrary locations in the listening area for reproduced sound. Previous research into modelling the perceived spatial attributes of sound reproduction systems has concentrated primarily on the sweet spot in the centre of the listening area. However, good audio reproduction is ideally required at multiple points in the listening area, for example for a family living room with a home cinema system.

A framework for modelling the perception of reproduced audio was developed, including the capture of the original sound-field, modelling the signals at the ears and the translation of the binaural signals to the perceptual domain. Explicitly modelling the binaural signals meant that the same principal cues as human listeners are used and also allowed existing binaural models to be incorporated into the system.

The three most widely researched different perceived spatial attributes were investigated: directional localisation, source width and listener envelopment. The models for predicting each of the three spatial attributes were validated using the results from formal listening tests. The output of the model was highly correlated with the localisation and envelopment results from the listening tests. Lower correlations were obtained for the predicted source width and for some groups of stimuli for directional localisation, and a number of modifications which may improve the performance of the model were identified.
Contents

Contents iii

List of Figures xi

List of Tables xix

Acknowledgements xxv

1 Introduction 1

1.1 Perceived spatial attributes of sound 2

1.2 Spatial attributes across the listening area 3

1.3 Scope and aims of the thesis 4

1.4 The organisation of the thesis 5

1.5 Contributions to the field 7

1.6 Summary 8

2 Spatial perception 9

2.1 The Filter model and different levels of the perception of spatial sound 10

2.2 Low level binaural measures 14

2.3 The primary spatial attribute: localisation 16

2.3.1 Existing models of localisation 17
6.1.2 Using the modified Supper model to predict the width results of the third listening test 183

6.2 The Mason source width model 188

6.2.1 Integration of the Mason source width model into the framework 191

6.2.2 Using the extended Mason source width model to predict the width results of the third listening test 194

6.3 Discussion 199

6.4 Summary 203

7 Envelopment 205

7.1 The data set for direct envelopment 206

7.2 Investigating the relationship between direct envelopment and directional localisation 208

7.2.1 The data set for the investigation of direct envelopment and directional localisation 209

7.2.2 Predicting the localisation of decomposed stimuli 210

7.2.3 Metrics for the prediction of envelopment based on decomposed stimuli 217

7.3 Predicting direct envelopment with complete stimuli 223

7.4 The data set for indirect envelopment 227

7.5 Predicting indirect envelopment with the output of the modified Supper model 228

7.6 Discussion 230

7.7 Summary 232

8 Conclusion 233

8.1 Summary of Chapter 1: Introduction 233

8.1.1 Conclusions from Chapter 1 234

8.2 Summary of Chapter 2: Spatial perception 234

8.2.1 Conclusions from Chapter 2 236

8.3 Summary of Chapter 3: Framework of the "source to percept" model 237

8.3.1 Conclusions from Chapter 3 238

CONTENTS vii
List of Figures

1.1 The organisation of the thesis and the dependencies between the chapters ............. 6

2.1 The Filter model .................................................. 11
2.2 The Jeffress model ................................................ 16
2.3 The structure used by many binaural models of directional localisation .............. 18
2.4 Combining the IID and ITD angles in Macpherson's model ......................... 19
2.5 Effects of a single lateral reflection .................................. 21
2.6 Examples of scene elements in Rumsey's scene based paradigm ..................... 28

3.1 The processes involved in audio reproduction ........................................... 36
3.2 Block diagram of the first four processes of audio reproduction .................... 39
3.3 Block diagram of the first four processes of audio reproduction for the case where the close microphone technique is used .............................. 39
3.4 Block diagram of the first four processes of audio reproduction for the first method of creating stimuli that was used in this thesis ......................... 41
3.5 Block diagram of the first four processes of audio reproduction for the second method of creating stimuli that was used in this thesis ................. 41
3.6 The diffuse field correction ................................................ 43
3.7 The polar co-ordinate system used in this project ...................................... 45
3.8 The orientation of the listener's head and the angle of sources relative to the listener ........ 45
3.9 FCS loudspeaker layout ................................................ 47
3.10 The directivity patterns of a circular array for recording the signals for an eight channel WFS system .................................................. 51

3.11 Pressure plots for 32-channel WFS ............................................. 51

4.1 The equipment layout used in the first listening test .......................... 57

4.2 The connections of the equipment used in the first listening test .......... 57

4.3 ORTF and Williams-Dü microphone configurations .......................... 59

4.4 Screen shots of the Max/MSP user interface from the first listening test . 61

4.5 Box plots of the localisation results for the centre listening position in the first listening test . 65

4.6 Box plots of the localisation results for the left listening position in the first listening test . 66

4.7 Box plots of the localisation results for the right listening position in the first listening test . 67

4.8 The mean localisation angles from the first listening test plotted against the intended angles . 68

4.9 The equipment layout used in the second and third listening tests .......... 71

4.10 The connections of the equipment used in the second listening test ........ 71

4.11 The impulse responses convolved with the original signals to create the left and right loudspeaker signals for the widened stimuli ............................................. 73

4.12 Screen shot of the Max/MSP user interface for the second listening test ... 75

4.13 Box plots of the localisation results for the centre position in the second listening test .... 79

4.14 Box plots of the source width results for the centre position in the second listening test .... 80

4.15 Plots of the localisation and source width results together for the centre listening position in the second listening test ........................................ 81

4.16 The mean localisation angles from the second listening test plotted against the intended angles 82

4.17 Modifications made to the interface in the second listening test ............ 84

4.18 Box plots of the localisation results for the centre listening position in the third listening test 92

4.19 Box plots of the localisation results for the left listening position in the third listening test ... 93

4.20 Box plots of the localisation results for the right listening position in the third listening test . 94

4.21 Box plots of the source width results for the centre listening position in the third listening test 95

LIST OF FIGURES | xii
4.22 Box plots of the source width results for the left listening position in the third listening test 96
4.23 Box plots of the source width results for the right listening position in the third listening test 97
4.24 Plots of the localisation and source width results together for the centre listening position in the third listening test 98
4.25 Plots of the localisation and source width results together for the left listening position in the third listening test 99
4.26 Plots of the localisation and source width results together for the right listening position in the third listening test 100
4.27 Comparison of the standard deviations of the localisation and source width results for the centre position in the third listening test 101
4.28 The mean localisation angles from the third listening test plotted against the intended angles 102
4.29 Modifications to the calculated values of the intended angles 102

5.1 Intermediate results for stimulus 11 from the third listening test 112
5.2 Averaged over time histograms for stimuli 1 and 6 from the third listening test 112
5.3 The structure of the localisation algorithm 114
5.4 Amplitude responses for the filter bank used in Supper's model 114
5.5 The structures of the algorithms converting the rectified and low-passed binaural signals into IID and ITD histograms 116
5.6 The population of the truth values in the columns of the IID look-up table 118
5.7 IIDs for the 18th critical band from the large and small ear HIRIRs in the Gardner and Martin database 119
5.8 Intermediate stages in the generation of the IID look-up table for the 18th critical band 119
5.9 Final stages in the generation of the IID look-up table for the 18th critical band 119
5.10 Removing the [0° 90°] folding from the IID look-up table 121
5.11 IID look-table with rows having a maximum value less than one 121
5.12 Gaps in the IID look-up table after the large and small ear values have been entered 123

LIST OF FIGURES xiii
5.13 The unexpected behaviour of the extending algorithm and the removal of the gaps in the IID look-up table .................................................. 123
5.14 Histograms output by the localisation model prior to the modifications .................. 124
5.15 Histograms output by the model after removing the rear-to-front folding and ensuring there are no gaps .................................................. 124
5.16 Rotation of bottom half of IID look-up table to reduced spikes at 0° .................. 126
5.17 The effect of rotating the bottom half of the IID look-up table to reduce the spikes at 0° on the output histograms .................................. 126
5.18 Applying vertical normalisation to the IID look-up tables .................................. 128
5.19 The effect on the output histograms of applying vertical normalisation to the IID look-up tables .................................................. 128
5.20 Applying horizontal normalisation to the IID look-up tables ............................. 129
5.21 The effect on the output histograms of applying horizontal normalisation to the IID look-up tables .................................................. 129
5.22 The calculation of the correlograms .......................................................... 132
5.23 Errors in the calculation of the correlograms ............................................. 132
5.24 The effect on the output histograms of changing the method of combining the IID and ITD histograms from addition to multiplication ............................................. 133
5.25 The effect on the output histograms of correcting the calculation of the correlograms .................................................. 133
5.26 The effect of the modifications to the localisation model on the ability to predict the localisation results from the three listening tests .................. 136
5.27 Prediction of the localisation results from the first listening test: omitting the sine tone stimuli and the stimuli with misallocated channels ............................................. 140
5.28 Prediction of the localisation results from the first listening test: anechoic and reflective stimuli .................................................. 140
5.29 Prediction of the localisation results from the second and third listening tests .................................................. 142
5.30 Prediction of the localisation results from the third listening test: the single loudspeaker and panned stimuli together and the artificially widened stimuli ............................................. 142
5.31 Anechoic and echoic listening test localisation results ..................................... 145
5.32 Anechoic and echoic localisation results from the model ..................................... 146

LIST OF FIGURES | xiv
5.33 The dimensions of the room models used in ODEON ........................................... 146
5.34 Localisation azimuths output by the model for signal family H and impulse responses calculated in ODEON .......................................................... 148
5.35 Comparison of the IID and ITD model results for signals in the nine signal families generated using impulse responses calculated from the ODEON room model with surfaces with 50% absorbency .......................................................... 150
5.36 The combined IID and ITD model results for signals in the nine signal families generated using impulse responses calculated from the ODEON room model with surfaces with 50% absorbency 150
5.37 Reverberation times plotted against frequency for the ten rooms modelled in CATT-Acoustics 151
5.38 Calculated and intended angles for the third listening test for reproduction rooms with 10% and 40% absorbent surfaces ........................................... 154
5.39 Calculated and intended angles for the third listening test for reproduction rooms with 70% and 100% absorbent surfaces ........................................... 154
5.40 Comparison of perceived localisation angles for different reproduction systems calculated for a listener facing forward (0°) at the sweet spot ........................................... 159
5.41 Comparison of perceived localisation angles for different reproduction systems calculated for a listening facing 45° to the right at the sweet spot ........................................... 159
5.42 The grid used to sample the listening area ................................................................. 165
5.43 Localisation results across the listening area for TCS (pan) where the intended angle of the source is 20° to the left ......................................................... 167
5.44 Localisation results across the listening area for TCS (pan) where the intended angle of the source is 40° to the right ......................................................... 167
5.45 Localisation results across the listening area for TCS (mics) where the intended angle of the source is 20° to the left ......................................................... 168
5.46 Localisation results across the listening area for TCS (mics) where the intended angle of the source is 40° to the right ......................................................... 168
5.47 Localisation results across the listening area for FCS (pan) where the intended angle of the source is 20° to the left ......................................................... 169
5.48 Localisation results across the listening area for FCS (pan) where the intended angle of the source is 40° to the right ......................................................... 169

LIST OF FIGURES | xv
5.49 Localisation results across the listening area for FCS (mics) where the intended angle of the source is 20° to the left ........................................ 170

5.50 Localisation results across the listening area for FCS (mics) where the intended angle of the source is 40° to the right ........................................ 170

5.51 Localisation results across the listening area for 32-channel WFS where the intended angle of the source is 20° to the left ........................................ 171

5.52 Localisation results across the listening area for 32-channel WFS where the intended angle of the source is 40° to the right ........................................ 171

6.1 Predicting the source width results from the third listening test using the standard deviations of the output from the modified Supper model ........................................ 186

6.2 Predicting the source width results from the third listening test using the standard deviations of the output from the modified Supper model: 95% confidence intervals ........................................ 187

6.3 The structure of the Mason source width model ........................................ 189

6.4 Combining the results from Mason’s source width model across frequency ........................................ 190

6.5 Output of the Mason model for stimulus 9 from the third listening test with the listener modelled at the central listening position ........................................ 191

6.6 Generating the width histograms from the output of Mason’s source width model ........................................ 193

6.7 Intermediate results for stimulus 9 from the third listening test ........................................ 194

6.8 Predicting the source width results from the third listening test using the Mason source width model ........................................ 197

6.9 Predicting the source width results from the third listening test using the Mason source width model: 95% confidence intervals ........................................ 198

6.10 Example of single sound source described using the graphical assessment language ........................................ 201

6.11 Examples of modifications to the graphical assessment language ........................................ 201

7.1 The equipment layout used in the direct envelopment listening test ........................................ 207

7.2 Screen shot of the Conetta’s Max/MSP user interface from his envelopment listening tests ........................................ 207

7.3 Screenshot of the Max/MSP user interface used in Conetta’s direct envelopment localisation experiment ........................................ 210
7.4 The AOT histograms output by the modified Supper model for the ninth stimulus from Conetta's direct envelopment localisation experiment ................................................................. 213

7.5 The AOT histograms output by the modified Supper model for the twelfth stimulus from Conetta's direct envelopment localisation experiment .................................................. 214

7.6 Localising the decomposed sources for the first eight stimuli from Conetta's direct envelopment localisation experiment .............................................................. 215

7.7 Localising the decomposed sources for the last eight stimuli from Conetta's direct envelopment localisation experiment .............................................................. 216

7.8 Calculation of the hull metric from the set of localisation angles output by the modified Supper model acting on a decomposed stimulus ...................................................... 218

7.9 Performance of the c90 and hull metrics as predictors of direct envelopment ........... 219

7.10 Results of linear regression models using the two feature sets from Table 7.1 containing four metrics fitted to the results of Conetta's first direct envelopment listening test experiment .... 222

7.11 The averaged over time histogram for the combined (itd*iid) output of the modified Supper model for stimulus 8 from Conetta's first MUSTEA direct envelopment listening test experiment . 224

7.12 The relationship between the three types of spatial impression identified by Griesinger .... 232

E.1 Inter-subject consistency for the localisation results from the first listening test experiment . 265

E.2 Comparison of the localisation results for each test subject with the stimuli with known locations from the first listening test ............................................................... 265

E.3 Inter-subject consistency for the localisation results from the second listening test experiment 269

E.4 Comparison of the localisation results for each test subject with the stimuli with known locations from the second listening test ............................................................... 270

E.5 Inter-subject consistency for the source width results from the second listening test experiment 270

E.6 Inter-subject consistency for the localisation results from the third listening test experiment . 273

E.7 Comparison of the localisation results for each test subject with the stimuli with known locations from the third listening test ............................................................... 274

E.8 Inter-subject consistency for the localisation results from the third listening test experiment . 275

LIST OF FIGURES | xvii
G.1 Perceived localisation angles for the TCS and FCS reproduction systems calculated for a listener facing forward (0°) at the sweet spot .......................... 286

G.2 Perceived localisation angles for the mono and 32-channel WFS reproduction systems and the original sound-field calculated for a listener facing forward (0°) at the sweet spot ........ 287

II.1 The equipment layout used in Conetta’s indirect envelopment experiment ............................... 295
List of Tables

2.1 The perceptually motivated lower integration limits for the calculation of LEV chosen by Soulodre et al. ............................................. 26

2.2 Proposed definitions of spatial attributes in reproduced sound in Rumsey’s scene-based paradigm 29

4.1 The original signals used as the basis for the stimuli in the three listening test experiments . 54

4.2 Details of the seven loudspeakers used in the first listening test experiment .................. 56

4.3 Details of the stimuli for the first listening test ...................................................... 60

4.4 The three listener positions used in the first listening test experiment ......................... 63

4.5 The number of tests at each listening position for the first listening test experiment .......... 63

4.6 The standard deviations, interquartile ranges and total ranges averaged over the four groups of stimuli from the first listening test ......................................................... 64

4.7 Details of the seven loudspeakers used in the second listening test experiment .......... 70

4.8 Details of the stimuli for the second listening test ................................................... 74

4.9 The three listener positions intended to be used in the second listening test experiment ... 77

4.10 Details of the stimuli for the third listening test experiment ..................................... 86

4.11 The number of tests at each position for the third listening test experiment .................. 87

4.12 The localisation errors for the third listening test experiment .................................... 88

4.13 The mean source widths for the third listening test experiment .................................. 89

4.14 Standard deviations of source widths from the third listening test experiment organised by processing ................................................................. 90
4.15 Standard deviations of source widths from the third listening test experiment organised by original signal .......................................................... 90

4.16 Result of Wilcoxon signed-rank test comparing the localisation results from the second and third listening test experiments that had matching combinations of stimulus, listening position and listener .......................................................... 104

4.17 Result of Wilcoxon signed-rank test comparing the source width results from the second and third listening test experiments that had matching combinations of stimulus, listening position and listener .......................................................... 105

4.18 Result of Wilcoxon signed-rank test comparing the range of the source width results from the second and third listening test experiments with the same combination of listening position and listener .......................................................... 105

4.19 The minimum, maximum and range of source width results from the second and third listening tests .................................................................. 106

5.1 Critical band frequencies for the filter bank in Supper’s model .................................................. 115

5.2 The ($R^2$) values between the results for the three listening tests and the localisation model in the three states .................................................................. 135

5.3 The gradient of the line of best fit that passes through the origin between the results for the three listening tests and the model in the three states .................................................................. 135

5.4 Summary of the steps in the generation of the look-up tables and the modifications that have been made to the Supper model .................................................................. 137

5.5 The results of validating the modified Supper model against the results of the three listening test experiments .................................................................. 138

5.6 Pairs of stimuli from the first listening test that were identical except for the reflectivity of the room model used when simulating the microphone capture of the original signals .................. 145

5.7 Details of the materials used in the room model in Odeon to alter the reverberation time .......... 148

5.8 The absolute values of the angles predicted by the modified Supper model for the nine signal families and the different ODEON room models .................................................................. 149

5.9 Comparison of using the two methods of calculating the binaural signals in an anechoic environment with the modified Supper model to predict the localisation results for the third listening test experiment .................................................................. 153

LIST OF TABLES | xx
5.10 The results of fitting the localisations from the model for differently reverberant reproduction environments to the intended angles for the third listening test .......................... 153

5.11 The differences between the intended angles and the predicted angles calculated at the sweet spot with the listener facing forward (0°) ...................................................... 161

5.12 The differences between the intended angles and the predicted angles calculated at the sweet spot with the listener facing 45° to the right ................................................. 163

5.13 The percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold for a source positioned at (−20°, 3) ....................... 173

5.14 The percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold for a source positioned at (40°, 3) ....................... 174

5.15 The percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold for a range of source angles ........................................... 176

5.16 The percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold for a range of listener orientations ........................................ 177

6.1 The RMSEP and correlation coefficients from using the output of the modified Supper model to predict the source width results from the third listening test experiment ........................................ 185

6.2 The correlation coefficients (R) between the width results of the third listening test and the predicted widths from using the SD method of interpreting the combined IID and ITD histogram output from the modified Supper model ................................................. 185

6.3 The RMSEP from using the SD method of interpreting the combined IID and ITD histogram output from the modified Supper model to predict the width results from the third listening test ................................................. 185

6.4 The source width results from the third listening test and the results predicted by the Mason source width model ................................................................. 195

6.5 The correlation coefficients (R) between the width results of the third listening test and the predicted widths from the Mason model .................................................. 196

6.6 The RMSEP from using the Mason model to predict the width results from the third listening test experiment ................................................................. 196

7.1 The results of performing linear regression with Conetta's metrics and also the c90 metric on the results from Conetta's first MUSTEA direct envelopment listening test experiment ........................................ 220
7.2 The table shows the four metrics calculated by taking the standard deviation from an AOT histogram and the model results from which each AOT histogram was generated .......................... 225

7.3 The results of applying linear regression to the data from Conetta's two MUSTEA direct envelopment listening test experiments ........................................... 226

7.4 The results of applying linear regression to the data from Conetta's MUSTEA indirect envelopment listening test experiment .............................................. 229

E.1 Results for Shapiro-Wilk test at the 5% level on the localisation results from the first listening test ..................................................... 264

E.2 Intra-subject consistency for the localisation results from the first listening test experiment ............................................................. 266

E.3 The effect of the tied observations correction on the calculation of the Kendall rank-order correlation coefficient, $T$, for the second listening test experiment results ........................................... 269

E.4 Results for Shapiro-Wilk test at the 5% level on the localisation results from the third listening test ...................................................... 273

E.5 Results for Shapiro-Wilk test at the 5% level on the source width results from the third listening test ...................................................... 275

E.6 Intra-subject consistency for the localisation results from the second and third listening test experiments ...................................................... 278

E.7 The root-mean-square error (RMSE) values between the localisation results from the test subjects and the location angles of those stimuli played through a single loudspeaker ................................. 278

E.8 Intra-subject consistency for the source width results from the second and third listening test experiments ...................................................... 279

II.1 Details of the stimuli used in Conetta's first direct envelopment experiment ............................................................... 291

II.2 Details of the stimuli used in Conetta's second direct envelopment experiment ............................................................... 293

II.3 Details of the stimuli used in Conetta's direct envelopment localisation experiment ............................................................... 294

II.4 Details of the stimuli used in Conetta's indirect envelopment experiment ............................................................... 297

I.1 The Variance Inflation Factors (VIFs) calculated for each of the metrics in the linear regression models predicting the data from Conetta's MUSTEA direct envelopment listening test experiments ........................................... 300
I.2 The Variance Inflation Factors (VIFs) calculated for each of the metrics in the linear regression models predicting the data from Conetta's MUSTEA indirect envelopment listening test experiment

301
For Jenny, Mum and Dad
Acknowledgements

Many people have contributed in many different ways to the completion of this thesis: Philip Jackson, Slawek Zielinski and Francis Rumsey for their very patient supervision, guidance and encouragement; Rob Conetta for help with the experimental side of things and for always being willing to talk about music when we should be working; Sören Bech and David Meares for always asking the right questions and helping me see the light; Ben Supper for his inspirational work; Russell Mason and Tim Brookes for giving me their time even when I ask stupid questions; Bill Backhouse and Eddie Proud for their technical support; Mum, Dad, Vera, Kev, Kerry, Emily, Penny, Andrew, Luke, Thomas, Cathy, David and Rosemarie for being there for me; Charlie Hulme for the music; Fred and Ginger for the eggs; and finally, Jenny, for putting up with me during the creation of the thesis, providing endless encouragement and reminding me about what is important in life.
Chapter 1

Introduction

There is often a need to evaluate the performance of an audio reproduction system, for instance, when designing audio products or codecs, or ensuring a certain level of quality in a broadcasting network. At the most basic level audio reproduction systems create sound that is intended to be heard by one or more listeners. Therefore, as creating sensations in a listener is the motivation for reproducing audio in the first place, it follows that any measures of quality of audio reproduction systems must be based on the perception of the reproduced sound by the listener. The only way to determine what listeners perceive is to ask them, which has led to the use of listening tests as the principal method of evaluating the quality of reproduced audio. However, listening tests are costly and require large amounts of time, expertise and technical resources in order to obtain meaningful results [9, 64]. This has led to much research into finding alternatives to listening tests, in particular developing models that can predict aspects of listeners’ perception of reproduced audio.

Bech and Zacharov [9] state that predictive models can be divided into two categories. The first category consists of models that aim to predict a particular perceptual attribute, for example, the loudness or localisation of the sound. The second category consists of models that aim to quantify the overall performance of an audio system or an acoustic space, for example, the Basic Audio Quality (BAQ), a measure of the overall quality evaluation defined in the international listening test standards [132]. Neither of these are limited to reproduced audio, even the models in the second category can be used in situations such as analysing concert hall acoustics. Indeed, models from both categories have been developed by both the concert hall acoustics community and the audio community (e.g. [23, 57, 143, 158]).

Audio reproduction systems have for some time achieved a high level of monaural timbral and temporal fidelity, and there are a number of algorithms that summarise multiple measures of monaural fidelity, based on signal characteristics, into a numerical scale of quality. For example, the Perceptual Evaluation of Speech Quality (PESQ) [134] and Perceptual Evaluation of Audio Quality (PEAQ) [160] models, which predict the mean opinion score using perceptual models that are mainly concerned with the audibility of noise and distortion. Both these models have been adopted as recommendations by the International Telecommunications Union (ITU). Like most existing sound-quality algorithms, however, PEAQ and PESQ...
are applicable only to individual channels and do not consider the spatial aspects of the reproduced sound. Yet, the spatial aspects of audio equipment are becoming increasingly important, as illustrated by the rise in home cinema use during recent years. Therefore, there is an increasing need for quality assessment of the spatial fidelity of reproduced sound, and any spatial quality scale is likely to be based on a number of perceived spatial attributes. The research described in this thesis investigates and develops models for predicting a number of perceived spatial attributes, with the intention that these can be used as the components in a model that can be used to predict the overall spatial quality.

1.1 Perceived spatial attributes of sound

This section briefly describes what is meant by the term spatial attribute in this thesis. An example of a basic spatial attribute is the location of a sound source relative to a listener. The source location may be interpreted as either the actual physical location of the sound source, or, alternatively, the perceived location of the sound source. Furthermore, these two concepts (i.e. the actual location and the perceived location) are not necessarily identical. This point is fundamental to audio reproduction systems with more than a single loudspeaker: these often rely on the fact that the reproduced sounds are not perceived to be located at the loudspeaker positions. For instance, in the case where the signals for the two loudspeakers in two channel stereo are identical except for their amplitude, these signals are combined at the ears of the listener, who then perceives a single sound to be located between the two loudspeakers [21]. The distinction between a physical attribute and a perceived attribute is not just limited to source location: reproduced audio also often attempts to create the spatial impression due to a reflective environment, but this is not generally done by recreating the actual physical environment.

Supper [158] distinguished between location as the primary spatial attribute and all other spatial attributes not including location as secondary spatial attributes. This classification is reflected in the literature, where the majority of research into spatial attributes has concentrated on perceived location. The research into localisation has mainly been undertaken by psychophysicists with the aim of having a better understanding of the human auditory system. In contrast, the research into secondary spatial attributes has largely been conducted within the context of concert hall acoustics. The rise in popularity of multi-channel audio reproduction systems, e.g. five channel home cinema systems, is one of the factors leading to an increase in the research into spatial attributes in the context of reproduced audio, which has increasingly included both primary and secondary spatial attributes.

Early research in concert hall acoustics tended to consider that there was only a single secondary spatial attribute, often termed spatial impression [7, 20], which was related to the early lateral reflections. More recently, research has shown that listeners perceive two distinct secondary spatial attributes [113], which have since been termed apparent source width (ASW) and listener envelopment (LEV), and the term spatial impression is now generally used to describe both these secondary spatial attributes. ASW and LEV have been shown to be related to the early (before 80ms) and late (after 80ms) reflections in the acoustic environment.
The width of an individual source and the sense of envelopment due to reflections in the environment are just two of the secondary spatial attributes identified by Rumsey [140] in the context of reproduced sound. Many of the secondary spatial attributes identified by Rumsey are specific (or, at least, more relevant) to reproduced sound rather than concert hall acoustics, e.g. the sense of envelopment due to being surrounded by a group of sources. Although there are more secondary spatial attributes than just width and envelopment (and Rumsey has identified more than one type of source and width), recent research into secondary spatial attributes of reproduced sound has still concentrated on width [98] and envelopment [38, 66].

1.2 Spatial attributes across the listening area

Computer algorithms have been developed that calculate a localisation direction from a binaural pair of signals (i.e. those at a listener’s ears), which is achieved by splitting the binaural signals into frequency bands then calculating time differences (using cross correlation) and level differences between left and right [89, 124, 158]. Some researchers have combined implementations of this type of model with Head Related Transfer Functions (HRTFs) to either examine the behaviour of the binaural cues [76, 129] or predict the perceived directional localisation of virtual sources [128]. Computational models have also been developed for predicting secondary spatial attributes in the context of reproduced sound, including source width [98, 158] and envelopment [38].

A limitation of many previous studies into the perceived spatial attributes of sound reproduction systems is that they have concentrated primarily on the sweet spot in the centre of the listening area, e.g. Pulkki et al. [128, 129]. A few researchers have considered the perception of spatial attributes away from the centre position. Härma et al [69] measured the timbral sound quality over a wide listening area by modelling the sound colouration at different points, treating the listener’s head as being acoustically transparent. While it is an important step to consider points other than the sweet spot for evaluating the sound quality of real systems, this approach does not consider perceived spatial sound quality, or adequately model the signals at the listener’s ears.

Macpherson [89] and Rose et al. [136] both used binaural models to assess the spatial performance of audio reproduction systems when the listener was displaced laterally, i.e. moved in a direction parallel to the line passing through the two ears. Macpherson used only one type of reproduction system (conventional two channel stereo) and one intended source location (midway between the two loudspeakers; the same signals were sent to both loudspeakers). The results of this experiment showed that the perceived source location collapsed into the nearest loudspeaker as the listener moved further away from the centre position, due to both the precedence effect and the increased loudness of the nearest loudspeaker. These results were validated using informal listening tests. Rose et al. also considered one type of reproduction system, although in their case this was transaural stereo, where transfer functions and a pair of loudspeakers are used to recreate the binaural signals which would have been experienced in the original sound-field at the ears of the listener [8, 85]. The aim of their experiment was very specific, namely to establish the size of the sweet spot when using transaural stereo.
Good reproduction is ideally required at all points in the listening area, for example for a family living room with a home cinema system. Hence, to understand the performance of any system, we need to measure perceived spatial attributes throughout the listening area. This capability would allow different types of sound reproduction systems to be compared objectively, and could assist measurement and design of new reproduction systems and panning laws, or of efficient spatial audio coding schemes. Therefore, one of the principal aims of the research described in this thesis is to develop models of perceived spatial attributes which are not confined to a single position in the listening area and to validate these models using formal listening tests.

1.3 Scope and aims of the thesis

The aim of the research described in this thesis was to develop an artificial listener, i.e. to develop a model whose acoustical transducers could be positioned anywhere in the listening environment and can then be used to predict a number of different perceived spatial attributes at that location. In particular, the artificial listener was required to be able to predict perceived spatial attributes for reproduced sound. The artificial listener was validated using formal listening tests to ensure that the model is able to adequately fulfill its intended purpose. The development of the artificial listener also included some investigation into the possible applications and limitations of the model.

As the model was required to predict perceived spatial attributes rather than actual physical spatial attributes, it was decided that the model should use the same principal cues as human listeners. This meant that the primary cues for the model to calculate the perceived spatial attributes were the binaural signals corresponding to the position in the listening area. In certain situations the model also used secondary cues, such as using head movements to differentiate between sources from the front and rear of the listener. The use of binaural models to predict perceived spatial attributes is well documented in the literature, both for primary and secondary spatial attributes, e.g. [98, 128]. Consequently, the decision to use binaural signals as the primary cues for the model allowed the use of existing binaural models as a basis for the development of the artificial listener.

All the binaural models described in the literature have used techniques inspired by the physiology of the human auditory system. As the human auditory system is not fully understood, none of these models is exclusively inspired by physiology, and all the models have had to employ some elements where the principal consideration in the design has been the ease and efficiency of implementation rather than physiological accuracy. Nonetheless, by adopting a physiologically based approach wherever possible, the behaviour of the model should more accurately reflect that of human listeners.

One of the motivations for the development of models that can predict aspects of listeners' perception of reproduced sound is that it is faster, cheaper and requires fewer resources than conducting subjective listening tests. If only binaural models are used, with no modelling of the signals before the stage where the binaural signals are input to the binaural models, then the binaural signals have to be physically recorded using
a dummy head (e.g. the KEMAR dummy head [31]). This means that the comparison of different audio reproduction systems at multiple locations in the listening area still requires considerable technical resources (e.g. the dummy head and enough audio equipment to be able to create each of the different reproduction systems) and time (e.g. to record binaural signals at each of the different locations in the listening area). Therefore, for similar reasons to those that motivated the development of the binaural models in the first place, the decision was made to incorporate the acoustical modelling of the sound from the loudspeakers in the reproduction system to the ears of the artificial listener. This has a precedent in the literature, for example, the work of Macpherson [89] and Pulkki et al. [128].

Hence, the task of developing models that are able to predict perceived spatial attributes of reproduced sound at different locations across the listening area can be divided into two stages: firstly, ensuring that the binaural signals are correctly modelled and, secondly, converting the binaural signals to measures of perceived spatial attributes. The explicit modelling of the binaural signals from the loudspeakers also allows different audio reproduction systems to be easily compared, which is not the case with models which omit the binaural signals and extract features for the prediction of perceived spatial attributes directly from the loudspeaker signals, e.g. George [57].

Audio reproduction systems differ not only in the layout of the loudspeakers, but also in the method of generating the signals fed to the loudspeakers. As many of the techniques for generating the loudspeaker signals involve the use of microphones to record an original sound-field, the framework for the artificial listener model was extended to include not only the acoustical modelling of the sound from the loudspeakers to the listener’s ears in the reproduction environment, but also the acoustical modelling of the sound source to the microphones in the original sound-field.

A further requirement of the artificial listener model was that it should be highly modular. This gave the maximum flexibility when it comes to applications for which the model can be used. For example, although the binaural models were integrated into the model to work in tandem with the acoustical modelling stages in order to allow the comparison of different audio reproduction systems, the binaural models could be used directly with dummy head recordings.

In addition to localisation, the primary spatial attribute, two secondary spatial attributes were also investigated. The fact that source width and envelopment have been the secondary spatial attributes that have so far received the most attention from both the concert hall acoustics and audio reproduction research communities suggests that these are the two spatial attributes that have the largest influence on the overall spatial perception. For this reason, the three spatial attributes which were investigated in this thesis were directional localisation, source width and envelopment.

1.4 The organisation of the thesis

This section briefly summarises the contents of each chapter, and the organisation and dependencies of the different chapters are summarised in Fig. 1.1.
Chapter two reviews previous research conducted into evaluating and modelling the spatial attributes of sound. The research fields of concert hall acoustics, auditory psychophysics and the evaluation of reproduced audio are discussed, together with the similarities, differences and overlap between the three fields.

Chapter three describes a framework within which models of perceptual attributes can be developed. The co-ordinate system and the reproduction systems used in the thesis are also described in this chapter.

Chapter four describes the three listening test experiments that were conducted to generate sets of directional localisation and source width results with which to validate the models of individual perceptual attributes.

Chapter five describes the investigation into the perception of directional localisation. This chapter begins with a description of the integration of the Supper model of localisation [158] into the framework described in Chapter 3, including a description of the alterations made to the Supper model to improve its performance. This is followed by a description of the validation of the directional localisation model using the results from the listening tests. The remainder of the chapter describes three investigations using the modified Supper model. The first investigation explores the effect of reflections on the model: firstly, having a reflective recording environment and, secondly, having a reflective reproduction environment. The second investigation compares the performance of different audio reproduction systems at the central listening position. The final investigation demonstrates how the model can be used to calculate localisation angles over the listening area.

Chapter six describes how the modified Supper model and Mason's source width model [98] were incorporated
into the overall model framework and also each model's ability to predict the width results from the listening
tests described in Chapter 4.

Chapter seven describes the development of metrics derived from the histograms output by the modified
Supper model for the prediction of envelopment. The chapter describes the listening tests conducted by
Conetta [39, 38] to investigate two types of envelopment. The metrics developed from the modified Supper
were then used in conjunction with metrics developed by Conetta [38] in linear regression models to predict
the results from Conetta's listening test experiments.

Chapter eight summarises the main results from the thesis and discusses the main contributions from the
project and also the possible future work and improvements that can be made to the model.

1.5 Contributions to the field

The research undertaken for this thesis has resulted in a number of original contributions have been made
to the fields of binaural models and subjective testing. These are summarised below:

- A novel user interface for eliciting localisation and width perceptions from listeners at different positions
  within the listening area was developed and employed in listening test experiments. The performance
  of the method employed by the user interface was critically assessed.

- The localisation model used in the thesis was formally validated for multiple listening positions using
  wide-band noise and speech and music signals and the results from the listening test experiments.

- The localisation algorithm developed by Supper [158] was modified, leading to improved performance
  when validated against the results from the listening tests.

- The limitations of the model with respect to the effects of reflections (in both the recording and
  reproduction environments) were systematically explored.

- The ability of different reproduction systems to produce localisable sources across the listening area was
  investigated. This included the production of maps of the perceived directional localisations calculated
  by the model across the listening area.

- The performance of two different models for predicting source width was formally investigated using
  the results from the listening test experiments.

- It was demonstrated that, when features derived from the modified directional localisation model were
  used in conjunction with the features selected by Conetta [38] in a regression model, the ability of the
  regression model to predict the results from Conetta's envelopment listening test improved.

- A method of differentiating between front and back sources using small head rotations was implemented
  and was shown to work successfully with the stimuli from Conetta's localisation and envelopment
  listening test experiment.
1.6 Summary

This introduction has described the motivation and the background for the research described in this thesis, together with the scope and aims of the thesis and a summary of the main contributions to the field.

The first section in the chapter is a discussion of the motivation for developing models that can predict perceived spatial attributes. In the following section it was discussed how localisation can be considered the primary spatial attribute. It was also discussed how, of the other, secondary, spatial attributes, width and envelopment appear to be judged the most important in terms of the overall spatial perception, as these attributes have been the subject of the largest amount of research, both in the field of concert hall acoustics and in the field of reproduced audio.

Section 1.2 discusses the need to assess perceived spatial attributes at multiple locations across the listening area. While there has been previous research into predicting spatial attributes away from the sweet spot at the centre of the listening area, these have been confined to varying the listener location in just a single dimension and have also only been concerned with directional localisation. One of the principal aims of this project is to investigate perceived spatial attributes, including some secondary spatial attributes, at multiple locations across the listening area.

Section 1.3 discusses the scopes and aims of the project described in this thesis, namely to develop an artificial listener able to predict different perceived spatial attributes at different locations in the listening area of an audio reproduction system. These can be summarised as:

- Three perceived spatial attributes were required to be investigated: directional localisation, source width and envelopment.
- Binaural signals were required to be explicitly modelled, allowing existing binaural models to be used as a basis for the model.
- The model was required to be able to predict the spatial attributes at different locations in the listening area.
- The model was required to include acoustical modelling, both in the recording and reproduction environments, in order to enable the comparison of the spatial attributes of different audio reproduction systems.
- The model was required to be modular wherever possible in order to maintain flexibility.

The remaining sections in the chapter outlined the structure of the thesis and summarised the original contributions to the field.
Chapter 2

Spatial perception

The purpose of this chapter is to provide an overview of the research that has been conducted into the development of objective measures of the spatial attributes of sound. Hence this chapter will provide a basis for the justification of the methods used and the decisions taken during the course of the research described in this thesis.

Theile [159] distinguishes between three different types of spatial sound stimuli that a listener can perceive. These are precedent (direct) sound, non-precedent (indirect) sound and environmental (non-reflected) sound. Precedent sound is defined as the direct sound from located sound sources and allows the directional localisation and, to some extent, the distance of the sound source to be perceived. Theile defines non-precedent sound as the reflections and reverberation which give the spatial impression of the sound, that is, the perception that the listener is in the same physical space as the sound. Environmental (non-reflected) sound is defined as the ambient non-located sound sources which allow the listener to perceive the enveloping atmosphere of the environment. This could be, for example, the sounds created by the audience during a live performance by an orchestra. The reason Theile distinguishes between these three types of sound stimulus is that each type follows different rules for the way in which they are perceived by the listener. Of the three different types of sound stimuli, the first two (precedent and non-precedent sound) have been the subject of much more research with regard to reproduced audio than environmental sound. This is at least partly due to the fact that the majority of audio recordings contain both direct and indirect sounds. In contrast, environmental sounds in recordings are much less common, particularly in music recordings.

Theile identifies the main spatial perceptual attribute created by direct sound to be localisation, and this is often considered to be the most important purpose of spatial hearing, as it can assist in an individual's survival. For this reason, Supper [158] calls the perception of source location the primary spatial attribute. In practice, the localisation of a sound depends not only on the direct sound but also on the early reflections in the non-precedent sound. The early reflections provide cues to the listener that clarify or obscure the cues from the direct sound for localisation. For example, the reflections from the floor of the listening environment are crucial in providing cues for the evaluation of the distance between the listener and the sound source [139].
Supper identifies all the other spatial properties of sound as secondary spatial attributes. These secondary spatial attributes often rely on the cues resulting from the indirect, reflected sound (Theile’s second type of sound stimuli).

Until relatively recently the research into the localisation of sound sources and the research into other aspects of the spatial impression in the listener have been carried out independently of each other. Historically, sound localisation, particularly directional localisation, has been studied by psychophysicists, whose chief motivation has been to determine the mechanisms of the human ear and the related neural processing. This has led to many of the studies involving listener experiments using stimuli such as impulses [84], noise bursts [154], sinusoids [104, 154, 141] and, occasionally, speech [54]. All these stimuli are highly controllable and all were used in anechoic conditions. As knowledge of the perception of sound localisation has increased, so the simple signals used in the listening tests have had to become more sophisticated, for example the tone bursts with controlled onsets and offsets used by Perrot [121].

In comparison with the early research conducted into sound localisation, much early research into other spatial attributes of the sound such as envelopment and width was conducted with the aim of either analysing and predicting concert hall acoustics [11] or, alternatively, analysing the spatial performance of audio equipment [138]. Recently, however, the topics of spatial localisation and other spatial attributes have become much more closely related in research studies, largely because of the development of algorithms that require the calculation of the Interaural Cross Correlation (IACC) from binaural signals for both localisation and source width [72].

This chapter contains four main sections. The first section describes the Filter model developed by Fog and Pedersen [50] and the difference between subjective and objective perceptual attributes. The second section describes the low-level binaural measures which are commonly used in models for predicting both primary and secondary spatial attributes. Finally, the third and fourth sections discuss the research that has been conducted into the development of objective measures for the primary spatial attribute (localisation) and secondary spatial attributes (spatial impression) respectively. This includes a discussion of how these objective measures relate to the case of reproduced sound.

2.1 The Filter model and different levels of the perception of spatial sound

A number of different physiological and mental processes are involved between an auditory stimulus arriving at the ears of a listener and the listener giving a final assessment of the total auditory experience. The processes between a physical stimulus and a subject’s assessment of the stimulus have been researched for other types of stimuli, including the timbre of sound [123], food [155] and image quality [117]. Fog and Pedersen [50] generalised and simplified these models into the Filter model, shown in Fig. 2.1, which was subsequently adopted by Martin and Bech for the evaluation of auditory stimuli [95]. This model divides into five sections: the physical stimulus, the perceived stimulus, the hedonic rating, the mapping from the...
The Filter model, showing the translation of a physical stimulus to a preference (from Martin and Bech [95]).

The first section consists of a description of the physical stimulus, which will vary with the type of stimulus and also the type of assessment being modelled. For instance, in the case of assessing the performance of sampled images, the physical stimulus of the Filter model consists of physical parameters specifying the image and the sampling and interpolating processes [117]. In the case of assessing the performance of different audio reproduction systems, the physical stimulus section of the Filter model can consist of a standard set of measurements, including frequency response and distortion characteristics [95]. One of the tasks of the researcher using the Filter model is selecting suitable physical parameters for the first section. There may be a number of different sets of physical parameters which could be used, and these may have implications on the design of the rest of the model. For instance, as mentioned earlier, Martin et al. [95] describe using a standard set of measurements on an audio reproduction system when modelling the assessment of its performance. An alternative approach to the same problem is to capture the signals arriving at a listener's ears and derive physical measures from these signals. A consequence of this approach is that subsequent sections of the model can be explicitly modelled on the human auditory system, and this is the approach used in this thesis.

The first filter in the model is the mapping from the physical stimulus to the perceived stimulus. This mapping depends on the sensitivity of the physiological mechanism used by the subject to perceive the physical stimulus. In the case of assessing reproduced audio, this mapping depends on the sensitivity of the ears. This means, for example, that listeners will not perceive changes in source location below the minimum audible angle (around 1° in front of the listener [104]) and will also not perceive stimuli which are below the minimum audible pressure [83, 125]. The mapping from the physical stimulus to the perceived stimulus in the first filter depends not only on the thresholds of the senses, but also on sensory selectivity. This is where one or more aspect of the physical stimulus masks another aspect. This is illustrated by the masking thresholds exploited in the perceptual coding used in the MPEG audio codec [27], where the parts of the audio signal which will not be perceived are removed in order to lower the bit-rate.

After the first filter comes the third section in the Filter model, consisting of a description of the perceived
stimulus. As in the description of the physical stimulus in the first section of the Filter model, this will vary with the type of physical stimulus and also the type of modality being being modelled (i.e. the nature of the description of the perceived stimulus will be different when assessing food compared with assessing an audio reproduction system). Martin et al. [95] characterise these descriptions of the perceptual stimulus as being objective if the subjects are sufficiently well trained. The first filter, i.e. the relationship between the physical stimulus and the perceived stimulus, is modelled as a linear regression by Martin and Bech [95] and Bech and Zacharov [9]. They consider the description of the physical stimulus to consist of $N$ different physical attributes, each of which has a value $\Phi_n$ (where $n = 1, 2, \ldots, N$) on its own individual scale. Similarly, the description of the perceptual stimulus is considered to consist of $M$ different perceptual attributes, each of which has a value $\Psi_m$ (where $m = 1, 2, \ldots, M$) on its own individual scale. Then linear regression can be used to attempt to map the two sets of variables:

$$
\Psi_m = \sum_{n=1}^{N} a_{m,n} \Phi_n \quad \text{for } m = 1, 2, \ldots, M,
$$

where $a_{m,n}$ are the regression coefficients.

Following the description of the perceived stimulus is the second filter, mapping the perceived stimulus to the hedonic rating. This mapping depends on the subject’s background expectations, interests and emotions. Whereas the first filter was based on physiological processes, the second filter is based on more cognitive processes, which depend on factors unique to each subject, such as their emotions and the other aspects described above. For example, if a stimulus in a listening test experiment reminds the subject of an enjoyable experience (e.g. a holiday), then the subject is more likely to give a higher preference for this stimulus. Note that because the subject’s experience, expectations and emotions change over time, then so too does the mapping contained in the second filter.

The final section in the filter model consists of the hedonic or affective rating of the stimulus, i.e. how much the subject shows a liking or preference for the stimulus. Note that this rating will depend on the task given to the subject, i.e. what the subjects are required to assess and the same physical stimulus may have different hedonic ratings depending on the task. For example, a music recording may be rated highly if the subject’s task is to give their preference for the use of the recording at a wedding, yet the same recording may be rated much lower if the task is to give their preference for its use at a funeral.

In a similar manner to the first filter, the second filter is also modelled as a linear regression by Bech et al. [95, 9]. If $\Upsilon$ is the hedonic rating of the stimulus, then

$$
\Upsilon = \sum_{m=1}^{M} b_m \Psi_m \quad \text{for } m = 1, 2, \ldots, M,
$$

where $\Psi_m$ is the value of the $m$th perceptual attribute and the values $b_m$ and the regression coefficients. If linear regression models are used for both the filters then substituting Equation 2.1 into Equation 2.2 gives

$$
\Upsilon = \sum_{m=1}^{M} \sum_{n=1}^{N} b_m a_{m,n} \Phi_n,
$$

so the hedonic rating is a linear combination of the physical parameters of the physical stimulus.
Note that there is a difference between performing linear regression twice (firstly from the physical stimulus to the perceptual stimulus, and secondly from the perceptual stimulus to the hedonic rating) and performing a single linear regression from the parameters of the physical stimulus directly to the hedonic rating. On the one hand, using linear regression twice may lead to a more robust model, as the models for the individual perceptual attributes will have been verified. On the other hand, if the set of perceptual attributes is incomplete, then using a single linear regression from the parameters of the physical stimulus to the hedonic rating may include data from the physical parameters for the missing perceptual attributes, and so the final model may give better predictions of the hedonic rating.

The two filters in the model (or the single mapping from the physical stimulus to the hedonic rating) do not necessarily have to be implemented as linear regression models. For instance, the model developed by Choi et al. for predicting the perceived quality in multichannel audio codecs [35] uses a neural network framework for the mapping of the physical measurements to the mean opinion scores of the quality. However, their initial model whose results are reported uses a single layer and pure linear activation function in their neural network, which is identical to linear regression. It is also interesting to note that Choi et al. do not use the model with two filters described above and in Fig. 2.1. Their initial model uses a single linear regression from seven measurements of the audio stimuli and the codecs to the quality mean opinion scores. Furthermore, two of these measurements (the ILDDist and IACCDist) are very influenced by psychoacoustics and the physiology of the human auditory system. This means that it is unclear whether these two measurements would belong to the first section in the Filter model (the physical stimulus, see Fig. 2.1) or the third section (the perceptual stimulus). This is due to the fact that some of the mapping from the physical to the perceptual stimulus (in the first filter in Fig. 2.1) has already been designed into the two measurements. These types of measures are discussed in more detail in Section 2.2.

Bech and Zacharov [9] discuss the use of the Filter model, and within the model they locate Basic Audio Quality (BAQ) [131] as a combination of perceptual attributes. However, they do not consider the BAQ to be an affective or hedonic rating (if this were the case then the combining of the perceptual attributes into BAQ would be identified with the second filter in the Filter model, see Fig. 2.1), but rather as an objective measure. The fact that BAQ is considered by Bech et al. to be a combination of perceptual attributes implies that some cognitive processing is used by listeners when assessing BAQ. However, the fact that this combining is not identified with the second filter in the Filter model implies that Bech et al. do not consider this cognitive processing to be biased by the background expectations, emotions, etc. of the listener.

In practice, the development of a model to predict BAQ will involve calibration and/or validation using listening tests with multiple subjects. In this case, it could be argued that if there are any subjective effects due to background expectations, emotions, etc. from the listeners, then these will become less significant as more listeners are used. This is interesting because it implies that there is some consensus amongst different listeners about how to combine different perceptual attributes to give BAQ. This is supported by the work done by George et al. [57] on developing a model for predicting BAQ for multichannel audio recordings.

Conetta [38] also used listening tests with multiple subjects to gather results to calibrate his model of envelopment. Conetta found that a number of different perceptual attributes affected the sense of envelopment,
including the range of angles from which the sound was perceived to arrive and the rate of auditory events (e.g. the word rate in speech). As is the case for BAQ, these different perceptual attributes must be combined using cognitive processes. Again, as in George’s BAQ listening tests, there was found to be consensus between the different listeners as to how the different lower level perceptual attributes were combined.

The work of both George and Conetta shows that objective perceptual attributes can involve a degree of cognitive processing in addition to the physiological processes, i.e. the first of the two filters in the Filter model can employ both cognitive and physiological processes. They also show that some objective perceptual attributes, such as BAQ and envelopment, appear to result from cognitive processing of other, more physiologically based perceptual metrics. From this is can be seen that there are different levels of objective perceptual attributes according to the degree of cognitive processing. Indeed, as more cognitive processing is applied to a perceived stimulus then it becomes increasingly likely that the subject’s background expectations, experience and emotions will affect the perceived attribute. This suggests that, while the Filter model shown in Fig. 2.1 may be sufficient for modelling many subjective tasks, for more complicated tasks the model may require modification. Instead of considering just the objective perceptual stimulus and the hedonic rating, it may be more useful to consider a continuum of perceptual attributes ranging from the purely objective attributes based solely on physiological processing to the hedonic attributes which are wholly influenced by the expectations, emotions, etc. of the subject. For each perceptual attribute, the amount of cognitive processing and the possible influence of expectations and emotions, etc. will determine the position of the attribute in this range.

2.2 Low level binaural measures

This section describes some of the psychoacoustically inspired physical measures which can be derived from binaural signals. These physical measures have been used by researchers in a number of ways to predict various perceptual spatial attributes, including directional localisation, source width and envelopment. The first of these is the interaural cross-correlation function (IACCF), which is a measure of the similarity between the left- and right-ear signals in a binaural pair. The IACCF is calculated by calculating the correlation between the two binaural signals [72, 142]:

$$IACCF_{T_1,T_2}(\tau) = \frac{\int_{T_1}^{T_2} p_L(t) p_R(t + \tau) dt}{\left[ \int_{T_1}^{T_2} p_L^2(t) dt \int_{T_1}^{T_2} p_R^2(t) dt \right]^{1/2}}$$

where $p_L(t)$ and $p_R(t)$ are the left- and right-ear pressure signals at time $t$, $\tau$ is the delay between the left- and right-ear signals and $T_1$ and $T_2$ define the time interval over which the cross-correlation is calculated. The time taken for a sound wave arriving perpendicular to one side of the head to travel from the closest ear to the furthest ear is around 1ms, so $\tau$ is generally varied over the range -1ms to +1ms.

As the direction from which the sound arrives varies, so too will the difference between the arrival times of the sound at the two ears, known as the Interaural Time Difference (ITD). This means that the maximum
similarity between the left- and right-ear signals will occur when the delay \( \tau \) is equal to the ITD. Therefore, in order to obtain a single number that is a measure of the similarity of the two binaural signals, the *Interaural Cross Correlation* (IACC) is calculated as the maximum value of the IACCF as \( \tau \) ranges from -1ms to +1ms:

\[
IACC_{T1,T2} = \max_{|\tau| \leq 1ms} |IACCF_{T1,T2}(\tau)|, \tag{2.5}
\]

where \( \tau \) is the delay between the left- and right-ear signals and \( T1 \) and \( T2 \) define the time interval over which the cross-correlation is calculated.

Independent researchers [19, 32] have agreed that a high interaural cross-correlation coefficient (close to 1) has a subjective effect of being a single sound source, and as the interaural cross-correlation coefficient decreases towards zero, the perception is that that the sound source becomes wider and more diffuse, i.e. harder to localise. This has led Mason to develop a model for the prediction of the perceived source width based on the IACC [100]. The IACC has also been used to predict listener envelopment [38, 72, 143], which is discussed in detail later in the chapter.

The IACC is also used in the calculation of *Interaural Time Differences* (ITDs). In his paper from 1948, Jeffress [78] proposed a neurological model involving neurons which register when the hair cells in each ear are triggered at the same time. In addition to this, other, similar neurons register when there is a given delay between the hair cells being triggered in the two ears. Having a large number of these neurons, each responding to a different delay between the two ears, will allow the ITD to be determined by observing which of these coincidence-registering neurons is being triggered most often. Jeffress' model was formalised mathematically by Sayers and Cherry [142] so the ITD is calculated as the delay corresponding to the maximum IACCF value:

\[
ITD = \arg \max_{|\tau| \leq 1ms} |IACCF_{T1,T2}(\tau)|, \tag{2.6}
\]

where \( \tau \) is the range of delays over which the ITD is calculated and \( T1 \) and \( T2 \) define the time interval over which the cross-correlation is calculated, typically -1ms and +1ms. A computational model of Jeffress' neurological model is shown in Fig. 2.2, where each element of the output correlogram is an IACCF value corresponding to a different delay, \( \tau \), and the ITD is determined by finding the delay corresponding to the peak in the correlogram. Note that the normalisation of the IACCF due to the denominator (see Equation 2.4) does not affect the calculation of the ITD given in Equation 2.6, and consequently is frequently omitted when the IACCF is only being used to determine the ITD.

The final physical measure in this section are *Interaural Intensity Differences* (IIDs), which are the differences in the energy levels between the two ears. The head of the listener occludes the sound, attenuating the signal at the furthest ear. Also, a sound source to the side of the head will be closer to one ear than the other, leading to the signal at the furthest ear being attenuated more due to its longer path from the source. Together, these mean that the IIDs are used as cues in the directional localisation of sound sources.
IIDs can be calculated by taking the log ratio of the energy levels of the signals at the left and right ears [135]:

$$IID = 10 \log_{10} \frac{\int_0^T (p_R(t))^2 dt}{\int_0^T (p_L(t))^2 dt}$$

(2.7)

where $p_L$ and $p_R$ are the signals at the left and right ears respectively and $T$ is the integration time.

Although IIDs and ITDs have been used in measures of spatial impression [20, 65, 97], the main use of IIDs and ITDs has been in models of binaural directional localisation, as discussed in the next section.

### 2.3 The primary spatial attribute: localisation

Gerzon [59] has summarised the psychoacoustics of sound localisation. Below 700Hz, the listener’s directional localisation is based primarily on the Interaural Time Differences (ITDs) between the signals arriving at the two ears. Gerzon states that this approximately corresponds to the velocity vector of the sound-field that would be at the location of the centre of the head if the head were not there (this is comparable with the principles behind velocity microphones [120]). Gerzon terms the use of ITDs for sound localisation as Makita localisation after Makita [91]. Between 700Hz and 5kHz, it is the direction of the energy field of the sound-field around the listener that is used to provide localisation cues. This can be determined from the Interaural Intensity Differences (IIDs) between the signals arriving at the two ears. Above 5kHz the primary cues for directional localisation appear to depend on the colouration on the sound (i.e. the spectral changes of the sound) provided by the reflections of the pinnae (the external parts of the ears).

The use of the ITD, IID and spectral cues for modelling human sound localisation is generally accepted. The ITDs and IIDs are the most important cues in horizontal localisation and the spectral cues are mainly used.
for determining the elevation [90]. However, many models of sound localisation have been concerned only with localisation in the horizontal plane and not with the elevation of sources. Hence, most existing models of directional localisation use only the ITD and IID cues and do not consider the cues due to the filtering effect of the pinnae.

One aspect of using spectral cues for localisation is that there are actually two signals containing the spectral cues, one for each ear. This has led to research into how these two sets of spectral cues are combined [75, 80]. Indeed, research has also been conducted into how the spectral cues from both ears interact and are combined with the ITD and IID cues in directional localisation [90, 114]. Musicant and Butler propose that the spectral cues help in horizontal localisation as well as in determining the elevation, particularly around the side of the head, where changes in the directional location of the source cause only small changes in the ITDs and IIDs [114]. One problem of using spectral cues for localisation is that they depend partly on prior knowledge of the spectrum of the sound so that the filtering effects of the pinna, etc. can be determined. Wightman and Kistler [166] showed that listeners performed much worse when relying on spectral cues if the spectra of the stimuli were altered during the course of the experiment. This presents a problem when implementing a binaural localisation model using spectral cues where the original spectrum of each stimulus is unknown.

2.3.1 Existing models of localisation

A number of computer models have been developed that take binaural signals as their input and from these calculate the directional localisation of the sound source [89, 124, 158]. Closely related to these are a number of binaural computer models with different stages at the input so that binaural signals can be generated from virtual sources using HRTFs, such as those developed by Pulkki et al. [129] and Huopaniemi et al. [76]. These last two models do not calculate an actual angle of localisation, but use similar processing to calculate the IIDs and ITDs which can then be analysed. Pulkki later added the calculation of azimuths from the IID and ITD cues to his model [128].

All of the binaural models cited so far work on similar principles, illustrated in Fig. 2.3. Firstly, the left- and right-ear signals are each separated into a number of frequency bands, each band with a range of approximately a third of an octave. These critical bands correspond to the position excited by frequencies on the basilar membrane [49]. The left- and right-ear signals for each critical band are then half-wave rectified, followed by low-pass filtering. Together, these two processes model the behaviour of the hair cells in the ear. The purpose of the low pass filter is to simulate the behaviour of the hair cell neurons: once a neuron has fired, it then has a recovery period before it can fire again. Consequently, the neural phase-locking, where the neurons are fired at the same point in the signal's phase, breaks down when the frequency increases [63]. The cut-off frequency for the low-pass filter is typically between 800Hz and 1200Hz (for example, see Colburn [37] and Stern and Shear [153]). Supper's model uses a second-order Butterworth filter with a cut-off frequency of 1100Hz [158]. Once the signals have been rectified and filtered, the IIDs and ITDs are calculated, with the ITDs calculated by finding the delay that maximises the IACCF, as shown in Equation 2.6. Horizontal azimuths are then calculated from the ITDs and IIDs for each critical band.
Although the binaural models cited so far have in common much of the processing described in the previous paragraph and in Fig. 2.3, these models also have notable differences. In Pocock's model [124], the tuning curves of fifty points on the Basilar membrane are modelled, each point corresponding to a different frequency [49], instead of using a filter bank to calculate the signals in different critical bands. Pocock also calculated the IIDs from the signals before the half-wave rectification and low-pass filtering (the modelling of the hair-cells) in order to avoid the non-linearities introduced by the rectification. This is similar to Macpherson's model [89], which uses basilar membrane modelling and hair-cell modelling for the signals used to calculate the ITDs, and uses just a filter bank to separate out the critical bands for the signals used to calculate the IIDs (with no hair-cell modelling for these signals). Supper's model [158] uses the critical band filtering shown in Fig. 2.3, but the calculation of the IIDs is slightly modified, in that the ITDs for each critical band are calculated first, and then the windows used to calculate the IIDs are adjusted by the ITDs so that similar portions of the left- and right-ear signals are used.

As stated before, the initial model developed by Pulkki et al. [129] and the model developed by Huopaniemi [76] calculate the IID and ITD cues, but do not convert either of these sets of cues into azimuths. Pocock [124], Macpherson [89] and Pulkki and Hirvonen [128] all use existing databases of HRTFs to build lookup tables which used to convert the IID and ITD cues for each frequency band into angles. Supper's model [158] also uses lookup tables, but, whereas the other models return a single angle for each IID or ITD cue, Supper's model returns histograms of fuzzy logic truth values. In addition to giving the most likely angle of...
Figure 2.4: Illustration of the method of combining the angles calculated for each cue (IID or ITD) and frequency band in Macpherson’s model [89]. For each cue and frequency band, a component histogram is created by plotting a cosine-shaped hump centered on the calculated angle and with a base width of 10°. The weighted component histograms are then summed to give an overall histogram, as illustrated by the histogram at the bottom of the figure. Only eight component histograms have been shown in the plot for clarity.

localisation for a given cue, this can also be interpreted to give a measure of confidence for the calculated angle.

The models developed by Macpherson and Supper combine the angles for each critical band and each type of cue (IID and ITD) to give a single localisation angle. Macpherson achieves this by creating a histogram for each cue (IID and ITD) in each frequency band every time the processing is triggered. Each of these component histograms has a cosine-shaped hump with a base 10° wide centered on the calculated angle. The component histograms are then weighted and summed to give an overall histogram. Macpherson computes the centroid of this histogram to give a single azimuth. This is illustrated in Fig. 2.4. The use of histograms also allows Macpherson to define a diffuseness parameter as the standard deviation of the combined histogram.

Recall that the model developed by Supper calculates histograms of fuzzy logic truth values for each cue and frequency band. These histograms are combined by first weighting each histogram by the loudness of the signals in each critical band. The histograms in each frequency band are then further weighted according to the duplex theory of sound localisation [130], which states that the ITDs are used for localisation at low frequencies and IIDs are used for localisation at high frequencies. A combined histogram is then calculated by summing all the weighted component histograms, which can then be interpreted in a similar manner to the histograms output by Macpherson’s model.

Pulkki et al. [129] estimate virtual source quality based on these principles, in which binaural signals are
extracted from the virtual sound-field and these are then processed to obtain measures of the quality of the spatial image. Pulkki et al. have obtained the binaural signals by applying HRTFs to each of the virtual sources, then summing these to create a pair of binaural signals for the entire virtual reproduced sound-field. Pulkki and Ilirvonen [128] used an expanded version of the same model, including the calculation of azimuths from the IID and ITD cues, to investigate the performance of two different loudspeaker layouts. The first loudspeaker layout was Five Channel Stereo (FCS, see Section 3.5.1), which was used with signals generated with first-order Ambisonics [62], a modelled spaced microphone array and pairwise amplitude panning using Pulkki's vector base amplitude panning (VBAP) method [127]. The second layout investigated was an array of eight loudspeakers at equally spaced angles (0°, ±45°, ±90°, ±135° and 180°, where 0° is in front of the listener). This was used with signals generated using first- and second-order Ambisonics [92] and pairwise amplitude panning using the VBAP method. Pulkki and Hirvonen also validated their model by conducting directional localisation listening tests using the two reproduction systems and compared the model results with the listening test results. Narrow band noise was used in both the simulations using the model and in the listening tests. It was found that the model was able to explain some of the prominent features from the listening test results. However, the angles resulting from the IID cues were found not to have a clear relationship with the listening test data. There were also systematic deviations between the model and listening test results, with the angles calculated from the ITD cues closer to the median plane than the angles from the listening tests for sources positioned to the side of the head.

Blanco-Martin et al. [17] have started to apply these localisation models to Wave Field Synthesis. They have generated binaural signals in two different ways, each method either using or simulating a linear array of ten loudspeakers in an anechoic chamber. The first method used was to generate the loudspeaker feed signals using virtual microphones and then generate the binaural signals using the Head Related Impulse Responses (HRIRs) from Gardner and Martin's KEMAR database [53]. The first step of the second method is identical to the first method, where the loudspeaker feeds are calculated using virtual microphones. These loudspeaker feeds are then used to drive a real loudspeaker array in an anechoic chamber and a dummy head (manufactured by Head Acoustics) was used to record the resulting soundfield. Two methods of localising the binaural signals were used, one using the Short Time Fourier Transform (STFT) and the other based on an auditory model, as described above (specifically that of Pulkki et al. [129]). Blanco-Martin et al. found that Pulkki's model produced much more accurate results.

2.3.2 The precedence effect and echo suppression

The term "precedence effect" was first used by Wallach et al. [162], and it refers to the changes in perception caused by two similar sounds occurring less than 100ms apart. When the interval between the two sounds is less than 1ms then only one sound is perceived as a single "phantom" source due to summing localisation [18]. As the delay between the first and second sounds increases beyond 1ms, only one sound is heard, coming from the location of the first sound source. As the delay increases further beyond 5ms, there is a perceived difference change in the tone of the sound and an increase in the spatial impression. Two distinct sounds are perceived when the delay increases beyond 50ms. The precedence effect is usually used to refer
Figure 2.5: The data from Barron’s experiments investigating the effects of a single lateral reflection. Figure adapted from Supper [158].

to only the second of these cases, where the total sound is heard only from the location of the first sound.

The listening tests conducted by Wallach et al. [162] investigated the precedence effect by presenting either clicks or music over headphones. It was found that the precedence effect broke down when the delay times were greater than 5ms and 40ms for the clicks and music stimuli respectively. The aim of the listening tests conducted by Haas [68] was not to investigate the spatial effects of the precedence effect, but instead to investigate the masking of a single “reflection”, where the second sound is perceived to be quieter than it really is for delay times between 2ms and 50ms. Although the two listening tests had different aims, the results from Haas and Wallach et al. both agree on a threshold of 50ms for the perception of distinct echoes. Haas used loudspeakers to present the stimuli, which consisted of speech rather than clicks. The investigation by Barron [5], which also used loudspeakers and used orchestral music as the stimuli, showed that a single lateral reflection could result in an increased sense of spatial impression in addition to the image shift identified by Wallach et al. Fig. 2.5 summarises Barron’s results.

A number of binaural directional localisation models have been developed that include some processing to include the precedence effect. This has been done in one of two ways. The first method modifies Jeffress’ model for calculating the cross-correlation between the left- and right-ear signals. This is the approach used by Lindemann [87, 88], which uses inhibitors which attenuate the left- and right-ear delay lines (see Fig. 2.2 according to the level of the cross-correlation and the other ear signal for each time delay. The general idea is that the first wavefront will correspond to high signal levels in both ears and a large cross-correlation for the delay corresponding to the ITD. This will cause the delay lines to be attenuated from the point corresponding to this delay onwards, which will lead to the effects of the early reflections on the localisation model being reduced. Gaik [52] modified Lindemann’s model further by also including weights on the delay lines in each

CHAPTER 2. SPATIAL PERCEPTION | 21
critical band so that the combinations of IID and ITD cues which occurred for real sound-sources in the free-field were emphasised. A consequence of this is that the portions of the binaural signals corresponding to the first wavefronts, which contain the least ambiguous binaural cues, are emphasised.

The second approach to including the precedent effect into binaural directional localisation models is to explicitly detect the onset of each sound and use this information to suppress the effects of the early reflections. This is the approach taken by Griesinger in his theoretical system [66] and also by the models developed by Macpherson [89] and Supper [158]. Macpherson's model calculates the firing rate of the neurons corresponding to hair-cells at different points on the basilar membrane, and determines the onsets of the sounds by the significant peaks in the neural rate signal. As in Lindemann's model, this will occur when the left- and right-ear signals are highly correlated for a given delay and the two signal levels are high. The onset detector in Supper's model uses the same initial processing as his localisation model, with the use of a filter bank to separate the binaural signals into critical bands. This is followed by full-wave rectification (half-wave rectification is used in Supper's localisation algorithm) and the extraction of the envelope for each signal. Finally, fuzzy logic techniques are used to create the onset signals. Both Supper's model and Macpherson's model were successfully verified by using stimuli for which the human response was known or could be assumed, including in reflective listening environments.

2.4 Secondary spatial attributes and spatial impression

The secondary spatial attributes consist of all the spatial properties of sound other than source localisation, which were all originally grouped under the umbrella term spatial impression in the literature. However, after Marshall [93] showed the importance of early reflections in concert hall acoustics, the term spatial impression was often thought to relate to the changes in spatial perception due to the early reflections. Indeed, Barron [5] defined the term spatial impression as being the subjective attribute associated with early lateral reflections. In 1989, Morimoto and Mackawa [113] conducted a series of experiments which showed that listeners perceived two distinct aspects of spatial impression. These two aspects were found to correlate to two different IACC measurements: one measurement over the entire room impulse response and the other measurement over only the portion of the impulse response containing the late reflections. This was clarified by Morimoto and Iika in 1993 [112], where apparent source width (ASW) was defined as “the width of a sound image fused temporally and spatially with a direct sound image” and envelopment was defined as “the fullness of sound images around a listener”. This division of the spatial impression into two areas, the ASW corresponding to the early reflections (the first 80ms) and the listener envelopment (LEV) corresponding to the late reflections (after the first 80ms), was confirmed by the subjective experiments conducted by Bradley and Soulodre in 1995 [25].

The discussion of secondary spatial attributes and spatial impression is divided into three sections. The first two sections discuss measures relating to the early and late reflections in concert hall acoustics respectively. Unlike a concert hall, audio reproduction systems cannot be characterised by a single impulse response. Instead, the spatial attributes of an audio reproduction system are strongly dependent on the recorded

CHAPTER 2. SPATIAL PERCEPTION | 22
signals, so the third section discusses the research into measures of spatial impression which are not based on room impulse responses.

2.4.1 Early reflections

Lateral reflections consist of those reflections approaching the listener at ±90° to the direction in which the listener is facing. The importance of early lateral reflections in concert halls to the impression produced in the listener was first proposed by Marshall in 1967 [93]. In 1981, Barron and Marshall [7] developed the lateral energy fraction, \( L_f \), as a measure of spatial impression at a point in an environment:

\[
L_f = \frac{\sum_{5\text{ms}}^{80\text{ms}} r \cos \phi}{\sum_{0\text{ms}}^{80\text{ms}} r},
\]

where \( r \) is the energy of each reflection and \( \phi \) is the angle of incidence of each reflection (\( \phi = 90° \) corresponding to the direction in which the listener is facing). The lower limit of the sum in the numerator ensures that the direct sound and the lateral reflections in the first 5ms were excluded (these very early lateral reflections were found not to contribute to the spatial impression). The lateral fraction is essentially the ratio of the sum of the energy from lateral reflections in the first 80ms to the sum of the energy of both the direct sound and all the reflections (regardless of direction) in the first 80ms. Hidaka et al. give the equation for the \( L_f \) in a slightly different form [72], changing the summations to integrals and showing the energy as the square of the sound pressure:

\[
L_f = \frac{\int_{5\text{ms}}^{80\text{ms}} p^2(t) \cos \phi \, dt}{\int_{0\text{ms}}^{80\text{ms}} p^2(t) \, dt},
\]

where \( p(t) \) is the sound pressure, \( t \) is time and \( \phi \) is the angle of incidence for each reflection. Bradley [23] developed a modified version of the lateral energy fraction that allowed conventional microphone techniques to be used for its measurement:

\[
LF_{80} = \frac{\int_{5\text{ms}}^{80\text{ms}} p^2_8(t) \, dt}{\int_{0\text{ms}}^{80\text{ms}} p^2(t) \, dt},
\]

where \( p_8(t) \) is the signal at time \( t \) captured by a figure-of-eight microphone orientated with its null axis in line with the sound source and \( p(t) \) is signal captured at the same point using an omni-directional microphone. Note that, \( p_8(t) = p(t) \cos \phi \), so \( p^2_8(t) = p^2(t) \cos^2 \phi \) and the \( LF_{80} \) does not have exactly the same angular response as the \( L_f \). Bradley acknowledges this, and actually used four variations of \( LF_{80} \) in his measurements. These are equation 2.10 given above, the same equation with the lower limit of the numerator set to 0ms, the same equation with a true cosine angular response using a method suggested by Kleiner [86] where the numerator has a lower limit of 0ms, and the same method with a lower limit of 5ms. Bradley found that
changing the lower limit of the numerator from 0ms to 5ms did not make much difference in practice, as the first 5ms consists mainly of the direct sound which is attenuated by the angular response of the figure-of-eight microphone.

A large amount of lateral reflections can cause the signals at the ears to become decorrelated, which has led to researchers associating a high degree of spatial impression with low IACC values [143, 72]. Keet [82] measured the cross-correlation between the first 50ms of the signals from two cardioid microphones positioned so their directivity patterns were 90° apart. This measurement, which approximates IACC0.50 (see Equation 2.5), was found to have an inverse linear relationship with ASW. Schroeder et al. [143] found that IACC values calculated for the first 50 to 140ms were useful for predicting listener preference. Barron [6] determined a relationship between IACC0.80 and LF80, which was investigated further by Bradley [23] for a number of different concert halls. Bradley compared the two different measures for six octaves, centred on the frequencies 125-, 250-, 500-, 1000-, 2000- and 4000Hz, with the conclusion that the IACC0.80 and LF80 are significantly related in the lower four octaves, but not in the highest two. The IACC0.80 measure was refined further by Hidaka et al. [72] who defined the IACCE3 to be the average of the IACC0.80 measures calculated for three octave frequency bands centred on 500Hz, 1kHz and 2kHz. The IACCE3 measure, calculated with in an unoccupied concert hall, was found to be an important objective parameter for predicting the acoustic quality of concert halls during symphonic concerts.

2.4.2 Late reflections

Bradley and Soulodre [25] found that a measure of the relative level of the late arriving lateral sound energy (LCG80) was highly correlated to values of listener envelopment, LEV, obtained in listening tests.

\[
LCG_{80} = 10 \log \left[ \frac{\int_{0ms}^{80ms} p(t) \cos^2 \phi \, dt}{\int_{0ms}^{\infty} p_A(t) \, dt} \right],
\]

where \(p(t)\) is the impulse response of the room, \(p_A(t)\) is the response of the same source measured at a distance of 10m in anechoic conditions and \(\phi\) is the angle of incidence of each reflection (where \(\phi = 90^\circ\) is the direction in which the listener is facing). Note that using a figure-of-eight microphone with the null axis in line with the sound source allows \(LCG_{80}\) to be calculated as

\[
LCG_{80} = 10 \log \left[ \frac{\int_{0ms}^{80ms} p_8(t) \, dt}{\int_{0ms}^{\infty} p_A(t) \, dt} \right],
\]

where \(p_8(t)\) is the signal captured using the figure-of-eight microphone and the other elements in the equation are the same as in equation (2.11). The units of \(LCG_{80}\) are dB.

In their 2002 paper, Soulodre et al. [149] modified the definition of \(LCG_{80}\) so that it could be used with reproduced audio. This was done by removing the denominator, as there is no longer an absolute reference...
level for the source signal. The equation was also generalised for arbitrary lower limits on the integration, giving

\[ LG^\infty_x = 10 \log \left[ \int_{t1}^{t2} p^2(t) \, dt \right] \],

where the elements in the equation are defined as for equation 2.11. Indeed, in a subsequent paper, Soulodre et al. [151] found that a lower limit of 105ms on the integration gave a better correlation with LEV scores obtained from subjective listening tests, i.e. using the measure \( LG^\infty_{105} \).

In their 1995 paper, Bradley and Soulodre [24] defined the relative sound level for concert hall acoustics to be

\[ G^\infty_{10} = 10 \log \left[ \frac{\int_{t1}^{t2} p^2(t) \, dt}{\int_{0}^{\infty} p_A^2(t) \, dt} \right] \],

where \( p(t) \) is the impulse response of the room, \( p_A(t) \) is the response of the same source measured at a distance of 10m in anechoic conditions and \( T_1 \) and \( T_2 \) are the limits of integration. This was modified in the same way as \( LG^\infty_x \) for use with reproduced audio, i.e. by omitting the denominator, giving the following definition for reproduced audio:

\[ G^\infty_{10} = 10 \log \int_{t1}^{t2} p^2(t) \, dt, \]

where the elements in the equation are defined as for equation 2.14. In their 2003 paper, Soulodre et al. [150] also generalised the lateral energy fraction (see equation 2.10) for arbitrary limits, allowing the lateral energy of the late reflections to be calculated:

\[ LF^\infty_x = \frac{\int_{t1}^{\infty} p_8^2(t) \, dt}{\int_{t1}^{\infty} p^2(t) \, dt} \],

where \( p_8(t) \) is the signal at time \( t \) captured by a figure-of-eight microphone orientated with its null axis in line with the sound source and \( p(t) \) is signal captured at the same point using an omni-directional microphone.

By substituting equations 2.15 and 2.16 into equation 2.13, \( LG^\infty_x \) can be expressed as

\[ LG^\infty_x = G^\infty_x + 10 \log [LF^\infty_x] = G^\infty_x + S^\infty_x, \]

where \( S^\infty_x \) is defined as

\[ S^\infty_x = 10 \log [LF^\infty_x] \].

Thus, Soulodre et al. showed that the level components and spatial components of LEV could be considered separately.

The research of Jesteadt et al. [79] and Moore and Glasberg [109] show that forward masking is frequency dependent, with longer forward masking at lower frequencies than at higher frequencies. This implies that
The perceptually motivated lower integration limits for the calculation of LEV chosen by Soulodre et al. [151].

<table>
<thead>
<tr>
<th>Octave band (Hz)</th>
<th>Integration limit $x$ (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>125</td>
<td>160</td>
</tr>
<tr>
<td>250</td>
<td>160</td>
</tr>
<tr>
<td>500</td>
<td>160</td>
</tr>
<tr>
<td>1000</td>
<td>75</td>
</tr>
<tr>
<td>2000</td>
<td>55</td>
</tr>
<tr>
<td>4000</td>
<td>45</td>
</tr>
<tr>
<td>8000</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 2.1: The perceptually motivated lower integration limits for the calculation of LEV chosen by Soulodre et al. [151].

The lower integration limits of the objective measures used to predict LEV will be larger for lower frequency bands than for smaller for higher frequency bands. Taking this into account, Soulodre et al. [150] chose different lower limits for the integrations based on their experimental results, the results of Jesteadt et al. and Moore and Glasberg, with the intention that the resulting measures will be robust. These integration limits are shown in Table 2.1. As equation 2.17 showed that the level and spatial components of LEV could be considered, Soulodre et al. developed the following objective measure for LEV:

$$GS_{perc} = 0.5 \cdot G_{perc} + S_{perc},$$

where $G_{perc}$ is equal to the average value of $G_x$ across the seven octave bands and the lower limits of integration, $x$, for each octave band are given in Table 2.1. Similarly, $S_{perc}$ is equal to the average value of $S_x$ across the seven octave bands with the perceptually motivated integration limits in Table 2.1. The constant of 0.5 before $G_{perc}$ was determined using the results of a subjective listening test. The $GS_{perc}$ measure was found to have an average correlation of 0.979 with the LEV scores from the listening test experiments conducted by Soulodre et al. Soulodre, however, does give the caveat that this objective measure requires access to the impulse response used to create the recording and so cannot be used to directly analyse existing recordings [148].

### 2.4.3 Reproduced audio and measures not based on impulse responses

The measures used to predict spatial impression discussed so far have all been based on room impulse responses, a consequence of having been developed principally for measuring concert hall acoustics. This is true even when the measures have been adapted for use with reproduced audio (e.g. Soulodre et al. [150]), where it was still necessary to have access to the impulse responses used in the creation of the signals). While using impulse responses has the advantage of being simple to implement, impulse responses are uncharacteristic of the majority of real program material. Also, when used by themselves, impulse responses do not create the spatial perceptions in listeners as well as when convolved with anechoic program material. Some of the objective measures developed for concert halls have been applied directly to reproduced audio,
such as by Bareham [4] and Martin et al. [96]. In both these cases test signals rather than typical program material were used: Bareham used a swept sinusoid to make his measurements and Martin et al. used white noise. However, the measures described in Sections 2.4.1 and 2.4.2 are not easily adapted for use with arbitrary signals, such as typical program material. This is important, as it has been shown that the properties of musical instruments or ensembles [102], the rate of speech [68] and the tempo of music [2] all affect the spatial perception of the listener. This means that the spatial perception of reproduced audio is very dependent on the signal content. Consequently a measure based on one simple test signal, such as an impulse response, is extremely unlikely to be able to provide an accurate indication of spatial impression.

Griesinger [65] developed the diffuse-field transfer function (DFT) as a measure of envelopment in listening rooms (i.e. rooms used with an audio reproduction system). The general principle of the DFT is to calculate the fluctuations in the interaural time differences (ITDs). Input signals of filtered noise are used and the ITDs are calculated using a zero-crossing method. The ITD fluctuations are calculated as the difference between the running average of the ITDs and the average ITD of the entire test signal. The DFT value is calculated by applying a 31Hz to 17Hz band-pass filter. Griesinger states that the DFT calculation, using the filtered noise test signals, is intended to correspond to the gaps in recorded music, which is where the reverberation is heard and thus has a large influence on the perception of envelopment.

The interaural cross-correlation fluctuation function (IACCFF) developed by Mason [97] to measure spatial impression in reproduced sound also analyses the fluctuations in ITDs. Whereas Greisinger calculated the ITDs using zero-crossings for the DFT, interaural cross-correlation is used to determine the ITDs for the IACCFF. The biggest difference between the DFT and IACCFF is that the IACCFF is intended to take arbitrary input signals. However, in his thesis, Mason does manually separate the binaural signals into the segments related to the source (i.e. when the direct sound and early reflections are present) and the segments related to the environment (i.e. when only the late reflections are present).

Both Griesinger and Mason acknowledge that different periods of the binaural signals contribute in different ways to the perceived spatial impression. This issue has already been raised in the discussion of the precedence effect and the measures developed for ASW and LEV in concert hall acoustics. In the case where impulse responses are used as the input signals, all of the signal after the initial onset consists of reflections. This allows the first 80ms of the impulse response to be attributed to the early reflections alone and the signal after the first 80ms to be attributed only to the late reflections. The situation is considerably complicated when arbitrary signals are used. Unlike impulses, these signals generally have a much greater duration, so the onset of the direct sound does not coincide with its offset. This means that when signals more typical of programme material are heard in a reflective environment, first the direct sound alone is heard, followed by a combination of the direct sound and the early reflections, and then the direct sound and the late reflections. It is only after the end if the direct sound (after its offset) that only the reflections are heard. These reflections may, of course, be interrupted by another onset of the direct sound. It is these gaps, where only the reflections are present, that are of interest to Griesinger with the DFT and Mason with the IACCFF (although Mason also calculates the IACCFF for the portions of the sound where the direct sound is present separately).
Griesinger formalises this in his 1999 paper, where he defines three types of spatial impression: continuous spatial impression (CSI), early spatial impression (ESI), and background spatial impression (BSI). CSI results from the lateral reflections interacting with a continuous source, ESI results from the early reflections, and BSI results from when the late reflections are heard in the gaps of the original signal. Of these three classes of spatial impression, Griesinger considers only CSI and BSI to be enveloping. The importance of detecting auditory onsets was discussed in the section on localisation and the precedence effect. In order to be able to detect the periods of binaural signals containing only reflections which can be used to calculate measures of BSI, Griesinger proposes that the offsets of the direct sounds are detected in addition to the onsets. Griesinger has not yet implemented this proposed model.

This section has so far shown that, in the literature, moving beyond the use of impulse responses as the signals input to objective measures is intrinsically linked with the desire to create measures of spatial impression which can be applied to reproduced audio. The research of Griesinger and Mason described so far in this section has been inspired to a greater or lesser extent by the objective measurements developed for use with concert hall acoustics. Indeed, the spatial perceptions which the DFT and IACCFF and Griesinger's proposed model for spatial analysis are all intended to predict are closely aligned with the spatial impressions commonly described in concert hall acoustics, i.e. ASW and LEV. Rumsey’s scene based paradigm [140] moves beyond this, recognising that the spatial perception of reproduced audio consists not only of the effects of reflections (in either the recording environment or the reproduction environment), but also of the spatial attributes of the components within the reproduced audio scene and the spatial attributes introduced by the reproduction system itself. The scene based paradigm recognises the elements of the reproduced audio scene and also relevant hierarchical groups of these elements. Typical elements within the reproduced scene include the individual sources and the environment. Some sources may be perceived together as a group rather than individually, e.g. the string section in an orchestra, which the scene based paradigm treats as an ensemble. Fig. 2.6 illustrates typical scene elements.

Table 2.2 contains the spatial attributes in reproduced sound for two dimensions identified by Rumsey. Rumsey differentiates between dimensional quantities, such as width, depth or distance and the immersion attributes of reproduced sound, which are semi-abstract multi-dimensional impressions. Note that environmental envelopment is similar to LEV in the concert hall acoustics literature. Rumsey also notes that individual source width and source focus may be closely related, and these also seem to be closely related to ASW in the concert hall acoustics literature.

The spatial attributes identified by Rumsey and classed as miscellaneous include higher-level attributes (i.e. which include a degree of cognitive processing; see Section 2.1) such as scene left-right skew and scene width homogeneity. In fact, all of the spatial attributes that Rumsey has classed as miscellaneous involve source localisation with some additional cognitive processing. For instance, the skew attributes and the scene width homogeneity attribute involve localising the sources in the scene and comparing the resulting localisations with a reference scene. Similarly, the stability attributes involve assessing how the localisations of sources varies over time, and the source focus attribute involves localising sources and assessing the difficulty of the localisations.
<table>
<thead>
<tr>
<th>Attribute class</th>
<th>Attribute</th>
<th>Construct definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional localisation</td>
<td>Individual source localisation</td>
<td>Azimuth of individual source(s) within a scene</td>
</tr>
<tr>
<td></td>
<td>Ensemble localisation</td>
<td>Azimuth of a group of sources</td>
</tr>
<tr>
<td>Width</td>
<td>Individual source width</td>
<td>Width of individual source(s) within a scene</td>
</tr>
<tr>
<td></td>
<td>Ensemble width</td>
<td>Overall width of a defined group of sources (may be all the sources in the scene if required)</td>
</tr>
<tr>
<td></td>
<td>Environment width</td>
<td>Broadness of (reflective) environment within which individual sources are located</td>
</tr>
<tr>
<td></td>
<td>Scene width</td>
<td>Composite or global width of entire scene</td>
</tr>
<tr>
<td>Distance and depth</td>
<td>Individual source distance</td>
<td>Distance from listener to perceived location of a source</td>
</tr>
<tr>
<td></td>
<td>Ensemble distance</td>
<td>Distance from listener to perceived midpoint of an ensemble</td>
</tr>
<tr>
<td></td>
<td>Individual source depth</td>
<td>Depth of individual source within a scene</td>
</tr>
<tr>
<td></td>
<td>Ensemble depth</td>
<td>Depth of a group of sources</td>
</tr>
<tr>
<td></td>
<td>Environment depth</td>
<td>Depth of (reflective) environment within which sources are located</td>
</tr>
<tr>
<td></td>
<td>Scene depth</td>
<td>Composite or global depth of entire scene, including environment</td>
</tr>
<tr>
<td>Immersion</td>
<td>Individual source envelopment</td>
<td>Sense of being enveloped by a single sound source</td>
</tr>
<tr>
<td></td>
<td>Ensemble source envelopment</td>
<td>Sense of being enveloped by a group of sound sources</td>
</tr>
<tr>
<td></td>
<td>Environmental envelopment</td>
<td>Sense of being enveloped by reverberant or environmental (background stream) sound</td>
</tr>
<tr>
<td></td>
<td>Presence</td>
<td>Sense of being inside an (enclosed) space or scene</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>Scene left-right skew</td>
<td>Degree to which a spatial audio scene is skewed to the left or right from a stated reference position</td>
</tr>
<tr>
<td></td>
<td>Scene front-back skew</td>
<td>Degree to which a spatial audio scene is skewed to the front or back from a stated reference position</td>
</tr>
<tr>
<td></td>
<td>Source stability</td>
<td>Degree to which individual sources remain stable in space with respect to time (assuming nominally stationary source)</td>
</tr>
<tr>
<td></td>
<td>Scene stability</td>
<td>Degree to which the entire scene remains stable in space with respect to time</td>
</tr>
<tr>
<td></td>
<td>Source focus</td>
<td>Degree to which individual sources can be precisely located in space (this may be closely related to individual source width)</td>
</tr>
<tr>
<td></td>
<td>Scene width homogeneity</td>
<td>Evenness of distribution of scene elements compared with a reference scene</td>
</tr>
</tbody>
</table>

Table 2.2: Proposed definitions of spatial attributes in reproduced sound, from Rumsey's scene-based paradigm [140].
Although source localisation, which Supper [158] identifies as the primary spatial attribute, is implicit in all of Rumsey's miscellaneous spatial attributes, he does not discuss it at the same length as the other attributes listed in Table 2.2. One reason for this is that localisation has already been well documented in the literature (see Section 2.3). Another reason is that there have been studies that have found a low correlation between localisation accuracy and listeners' preference for a reproduced sound recording [13, 170]. Rumsey's paper is concerned with the development of techniques for the evaluation and comparison of the spatial sound quality for different audio reproduction systems, which explains why greater emphasis is given to the spatial attributes which are likely to have a large effect on the spatial quality. Zacharov et al. [170] note that there is some confusion over the localisation perceived by the listener: mono sound reproduction systems (i.e. with a single loudspeaker) produced results that were perceived as being well-localised by the listeners. However, the sound sources were localised in the direction of the single loudspeaker, which is to be expected. This shows that there is a difference between localising a sound source well (in what may be the wrong position, for instance at a loudspeaker location) and accurately localising a sound source in a particular position.

Mason has developed a binaural model which can be used to assess the spatial attributes in Rumsey's scene based paradigm [98]. This model calculates both directional localisation and source width. The directional localisation is calculated using the processing stages shown in Fig. 2.3 and combining the IID and ITD azimuths. The source width is calculated using the same IACC values used in the calculation of the ITDs. The IACCs calculated for each critical band are then combined into a single width measurement, taking into account the dependence of the perceived width on the loudness and frequency of the stimuli [101]. Mason has stated that the intention is to extend the model to develop measurements that predict the perceived effect of all the spatial attributes or reproduced sound identified by Rumsey.

Finally, Conetta has developed regression models to calculate the envelopment of reproduced audio [38].
Whereas Rumsey had divided perceived envelopment into three areas (individual source envelopment, ensemble envelopment and environmental envelopment), Conetta divides envelopment into direct envelopment and indirect envelopment. Indirect envelopment corresponds to Rumsey's environmental envelopment, while direct envelopment consists of the envelopment caused by anechoic source signals being played through multiple loudspeakers. Direct envelopment includes Rumsey's ensemble envelopment (when anechoic sources are positioned at different positions around the loudspeaker array) and also Rumsey's individual source envelopment (typically when the same anechoic source is simultaneously played through multiple widely spaced loudspeakers).

Conetta performed a series of multiple stimulus listening test experiments (described in more detail in Chapter 7) investigating direct envelopment. The stimuli for these experiments consisted mainly of anechoic sources constant power panned around an equally spaced circular array of eight loudspeakers. The data from these listening tests were then used to calibrate a regression model. A similar listening test experiment was conducted to investigate indirect envelopment. In this experiment the stimuli consisted of anechoic sources convolved with the impulse responses for a very reverberant hall. Conetta found that the same set of five metrics worked well in the regression models for both direct and indirect envelopment. Again, these metrics are described in detail in Chapter 7. In common with many of the measures for spatial impression, two of Conetta's metrics involve the calculation of IACCs. He also includes a metric based on the total energy at the listening position, which accounts for the effect of loudness on envelopment, and also the mathematical entropy [144] of the signal at the left ear of the listener, which takes into account the effect of the rate of music or speech on spatial impression [2, 68]. Conetta found his set of metrics gave a good fit when used in regression models against the listening test data: his model for direct envelopment had a correlation of 0.95 and a root-mean-square error of prediction (RMSEP) of 6.9% and his model for indirect envelopment had a correlation of 0.89 and a RMSEP value of 11.54%.

2.5 Summary

This chapter has given an overview of the literature on the perception of the spatial aspects of sound. The chapter comprises four main sections: a discussion of the Filter model and the different levels of the perception of spatial sound; the low-level binaural measures which are commonly used in models of spatial perception; a discussion of binaural models for directional localisation; and a discussion of measures of spatial impression.

Section 2.1 described the Filter model developed by Fog and Pederson. In the context of spatial sound, this model separates out what listeners can objectively perceive and what are the listeners' subjective preferences. The first of the two filters in the model is the mapping of objective measures of the physical stimulus to the perceived stimulus. This mapping depends on the thresholds and masking effects of the human auditory system. The second filter in the model is the mapping of the objective perceived stimulus to the listeners' hedonic rating (i.e. whether they like the perceived stimulus or not). This mapping is subjective and depends on the background expectations and emotions of the listener. Bech and Zacharov locate Basic Audio Quality
(BAQ) as an objective attribute, but also as a combination of lower-level perceptual attributes, which suggests that Bech et al. consider some cognitive higher-level processing to be objective and not to be biased by the background expectations, emotions, etc. of the listener. This, in turn, suggests that the boundary between the objective perceived stimulus and the subjective hedonic rating is less clear cut than is suggested by the Filter model.

Section 2.2 discusses the low-level binaural measures which are commonly used in both directional localisation models and also measures of spatial impression. All of the measures described in this section depend on the differences between the left- and right-ear signals which arise due to the spatial properties of a sound-field. The interaural cross-correlation (IACC) and interaural time and intensity differences (ITDs and IIDs) are defined, together with a discussion of how these measures are calculated and how they relate to the human auditory system.

Section 2.3 is a discussion of directional localisation, which Supper termed the primary spatial attribute. Three principal cues are used in directional localisation: IIDs, ITDs and the spectral cues due to the reflections and shadowing of the outer ear. The ITDs are the primary localisation cues below 700Hz, the IIDs are the primary localisation cues between 700Hz and 5kHz, and the spectral cues are the primary localisation cues above 5kHz. Most of the existing binaural localisation models use a similar architecture: the binaural signals are first separated into critical bands and these signals are then half-wave rectified and low pass filtered. The IID cues for each critical band are calculated and ITDs are derived from the cross-correlation of the rectified and filtered left- and right-ear signals in each critical band. The IID and ITD cues are then converted to angles using a database of IID and ITD values for known angles. These angles are then combined, firstly, across the critical bands, and, secondly, across the different types of cue (IID and ITD) to give a single angle of localisation. Early reflections can affect source localisation, known as the precedence effect. Two approaches to allow for the precedence effect have been implemented in localisation models. The first approach introduces inhibitors to the delay lines used to calculate the cross-correlation used to calculate the ITDs, while the second approach explicitly detects the onsets of incoming sounds in order to use the cues from only the direct sound when determining the azimuth.

Section 2.4 is a discussion of spatial impression, which Supper termed the secondary spatial attributes, i.e. all the spatial attributes that are not localisation. Two different spatial perceptual attributes have been recognised in concert hall acoustics. The first of these occurs when the early lateral reflections fuse with the direct sound image, causing the image of the sound source to become wider. This perceptual attribute is termed apparent source width (ASW). The second perceptual attribute is listener envelopment (LEV) and is associated with the late lateral reflections. A time of 80ms has often been used as the boundary between the early and late reflections. In addition to measures of ASW and LEV based on the ratio of the lateral energy (due to the reflections) to the total energy, the IACC has also been used to predict spatial impression. This is because a large amount of lateral reflections cause the signals at the ears to become decorrelated, giving an inverse correlation between IACC and spatial impression.

All the measures of spatial impression developed for concert hall acoustics use impulse responses as their input, yet a number of researchers have shown that the nature of the source signal affects the spatial percep-
tion of sound. This has led to the development of measures which can take more continuous signals as their input, including Griesinger's diffuse-field transfer function (DFT) and Mason's interaural cross-correlation fluctuation function (IACCFF). In order to adapt the measures developed for concert hall acoustics to be used in a meaningful way with arbitrary signals, Griesinger has proposed detecting the offsets of direct sounds in order to be able to isolate the periods of the binaural signals where only the reflections are present. This has not yet been implemented.

Rumsey proposed a scene-based paradigm for evaluating the spatial attributes of reproduced sound. In addition to considering spatial attributes closely related to the ASW and LEV attributes used in concert hall acoustics, this paradigm considers spatial attributes which are specific to reproduced audio. Mason has developed a binaural model for calculating localisation and width which can be used within the scene-based paradigm. Similarly, Conetta has used some of the ideas of envelopment from the scene-based paradigm to develop envelopment models for reproduced sound.

Measures for the prediction of three different spatial attributes have been developed in the context of room acoustics: directional localisation, apparent source width (ASW) and listener envelopment (LEV). Directional localisation models have also been used to assess the spatial performance of audio reproduction systems. In comparison, the research into assessing the secondary spatial attributes of reproduced audio is still very young. Measures for the prediction of ASW and LEV in concert hall acoustics tend to use room impulse responses as their input, and so cannot be directly used with reproduced audio, where the spatial impression is highly dependent on the content of the audio signals. While Rumsey's scene-based paradigm identifies additional spatial attributes relevant to reproduced audio, the models that have since been developed have still concentrated on predicting directional localisation, source width and envelopment for reproduced audio. This illustrates the relative importance which researchers continue to assign to these three spatial attributes.
Chapter 3

Framework of the "source to percept" model

Most opinions of the aim of sound reproduction fall somewhere between two views [140]. The first is that the aim is to reproduce as accurately as possible the impression in the listener that they are at the live event being recorded. The second view is that the aim is to entertain the listener and that the experience of listening to the reproduced sound should be considered distinct to that of listening to a natural sound field. The listener is central to both viewpoints. The aim in both cases is to create an impression in the listener: in the first view the aim is that the impression should be as close to that of a natural sound-field as possible, whereas the aim in the second view is merely that the impression should be entertaining.

If the first view is taken and the aim is to create the illusion of some original sound-field, then if the reproduced sound-field is identical to the original sound-field it follows that a listener in the reproduced sound field will have exactly the same stimuli and hence the same impression as if they were in the original sound-field. However, it is not necessary to recreate the original sound-field to recreate the impression of the original sound-field in the listener. If the reproduced sound-field differs from the original sound-field but the signals at the ears of a listener in the original sound-field, then the listener will still have the same cues and again have the same impression as being in the original sound-field. Furthermore, only certain features of the binaural signals contain the cues used by listeners to perceive the spatial aspects of a sound-field. This means it is possible for the binaural signals for a listener in the reproduced sound-field to be different from the binaural signals for a listener in the original sound-field, yet because these features are identical in both sets of binaural signals then the listeners will have identical perceptions of the two sound-fields. A similar argument was used in the design of compression algorithms based on perceptual models, such as the MPEG 3 codec [26, 27, 67, 81, 165]. Consequently, the perception of the listener is central to comparing the spatial performance of different audio reproduction systems.

This chapter describes a framework for modelling the perception of reproduced audio. This is followed by a
more detailed discussion of the acoustic modelling employed in modelling the creation of the audio signals and calculating the binaural signals at the ears of a listener, including details of their implementation using Matlab. The remainder of the chapter contains a description of the coordinate system used in the thesis and a description of the standard reproduction systems which were investigated in the course of the thesis.

3.1 Modelling reproduced audio

Fig. 3.1 shows the different stages involved in audio reproduction. The first stage, I, is obtaining or creating the sounds on which the reproduced audio is to be based. The second stage, II, consists of capturing the original sound-field. This is done by using an array of one or more microphones to record the sound-field. This is effectively sampling the sound-field at a finite number of discrete points in the space in which the sound-field occurs. The third stage, III, consists of processing the recorded signals from the previous stage and also encoding these signals into a form in which they can be transmitted to an audio reproduction system. This stage frequently involves converting the analogue signals from the microphones into digital audio signals and can also include audio processing such as mixing, compression, panning, etc. The fourth stage, IV, is the transmission of the audio data from the previous stage to the reproduction system. This can be done in a number of ways, including broadcasting the audio data for radio or television reception or storing the audio data on media such as magnetic tape, CD or DVD.

The fifth and sixth stages (V and VI in the block diagram) comprise the audio reproduction system: the first
of these is the decoding the transmitted data into signal which can be used to feed the loudspeakers, which is then followed by the loudspeakers converting these signals into sound, creating the reproduced sound-field. The remaining three stages cover the effects that the reproduced sound-field has on the listener. The seventh stage, VII, determines the sound heard by the listener, i.e. the pressure signals at the ears of the listener. The next stage, VIII, is the process by which the listener translates the signals at their ears into perceptions of the reproduced audio. The final stage, IX, is the process by which the listener communicates to other people their perceptions of different spatial attributes. Note that not all of the processes shown in Fig. 3.1 are involved in every instance of reproduced audio. For example, although the majority of reproduced audio has its origins in sound-fields, typically containing speech or music, some electronic music is created entirely at the processing / encoding stage (III in the diagram).

Fig. 3.1 outlines a framework that can be used to investigate perceived spatial attributes. One of the aims of this investigation is the development of reliable measures of perceived spatial attributes. Bech and Zacharov [9] state that this can be done in one of two ways: by conducting listening tests or by using predictive models. Indeed, formal listening tests require that the different stages in the framework are performed in a controlled, objective manner, not least stage IX, where the listener is required to communicate what they perceive. Of course, the translation to the perceptual domain (VIII) is by its nature subjective, but the intention is to obtain a consensus in the results concerning the perceived attributes by ensuring consistency amongst the other stages in the framework and also by repeating the listening test using a number of different listeners.

A number of spatial models which take binaural signals as their input have already been discussed in Chapter 2. These include models for the prediction of directional localisation [76, 107, 124, 129, 158] and source width [98]. Two of these models, Supper's localisation model [158] and Mason's source width model [98], were integrated into the framework to investigate different spatial attributes of reproduced sound. The integration of Supper's localisation model into the framework is discussed in detail in Chapter 5. Similarly, a discussion of using Mason's width model in the context of the framework is contained in Chapter 6. The next two sections contain a discussion of the creation of the signals for reproduced audio (stages I and II), followed by a discussion of the generation of binaural signals from the reproduction system (stages VI and VII).

### 3.2 Creating signals for reproduced audio

Of the nine stages shown in Fig. 3.1, the evaluation of perceptual attributes of reproduced audio will necessarily concentrate on stages V to IX, the reproduction system and the effects on the listener. However, understanding the transmission stage and the creation of the audio signals in the first three stages is necessary to put the remaining stages in the framework into context. In particular, the processes in stages I to III will affect the types of perceptual attributes that will be stimulated in the listener and will also determine the possible changes in these attributes. In some previous studies of the perception of reproduced audio, particularly those with listening tests, only stages V to IV have been considered explicitly. In these listening test experiments the stimuli were chosen from a selection of existing programme material, and stages I to III
are not considered explicitly. The rationale for this is that it is more important that the stimuli are typical of real programme material than it is to have complete control over the processes used to create the audio signals. However, as the designer of the listening test generally selects the stimuli from a range of programme material with the intention that the stimuli should excite a range of levels of the perceptual attribute being investigated, the method of creating the programme material (i.e. stages I to III) is still being considered implicitly.

Fig. 3.2 shows a block diagram of the first four processes in audio reproduction, as described in Section 3.1 and Fig. 3.1. The use of close microphone techniques [22] can be considered a special case of the processes for stages I and II shown in Fig. 3.2. This is where a microphone is placed close to the sound source in order that the aspects of the sound-field relating to the recording environment are minimised, often to facilitate further processing with the recorded signal further down the signal chain. This is shown in Fig. 3.3. Stages I and II in Fig. 3.2 can be modelled as a linear invariant system, where the pressure at each microphone is the sum of the pressure due to each of the sound sources. The relationship between each sound source and each microphone is modelled by a transfer function. Each transfer function has an impulse response as its equivalent in the time domain. These are affected by a number of factors, including the location and orientation of the source, the directivity pattern of the source, the acoustic characteristics of the environment containing the sources and microphones (the capture environment), the location and orientation of the microphone and the directivity and frequency response of the microphone. In this thesis only sources with omnidirectional directivity patterns and microphones with flat frequency responses were considered.

The transfer functions for the case of anechoic capture environments was modelled directly in Matlab: the signal at the microphone, $S_m$, was calculated as:

$$S_m(t) = \frac{A_m(\theta_{s,m})}{d_{s,m}} S_s(t - d_{s,m}/c)$$  (3.1)

where $S_s(t)$ is the source signal at time $t$, $\theta_{s,m}$ is the angle of the source from the microphone, $A_m(\theta)$ is the attenuation due to the directivity pattern of the microphone at angle $\theta$, $d_{s,m}$ is the distance from the position of the source to the position of the microphone and $c$ is the speed of sound. This takes account of the attenuation due to the propagation of the sound through free space, the directivity pattern of the microphone and the delay due to the time taken for the sound to travel from the source to the microphone. The derivation of Equation 3.1 is discussed in Section 1.12 of Pierce's acoustics book [122]. In the case of reflective capture environments, the impulse responses were calculated using either the CATT-Acoustics [42, 41] or ODEON [36] acoustics modelling software. Note that anechoic environments are a subset of the acoustic environments that can be modelled in both these software packages. However, it was faster and more convenient to model the anechoic capture environments directly in Matlab, as described above, rather than generate the impulse responses using either CATT-Acoustics or ODEON.

In order to study the effects of varying the factors which affect the transfer functions of the sources to the microphones in stages I and II (e.g. the positions of the sources and microphones and the acoustical properties of the capture environment) it is necessary to have a repeatable sound source. This was achieved by modelling a loudspeaker as the sound source and using to play back either existing anechoic recordings

CHAPTER 3. FRAMEWORK OF THE "SOURCE TO PERCEPT" MODEL | 38
Figure 3.2: *Block diagram of the first four processes of audio reproduction (see Fig. 3.1). The processing stage (III) may involve explicitly changing the spatial characteristics of the audio signals in addition to encoding the signals into a format suitable for transmission.*

Figure 3.3: *Block diagram of the first four processes of audio reproduction (see Fig. 3.1) for the case where the close microphone technique is used. Here all the spatial characteristics of the audio signals such as location are only introduced by the processing in stage III.*
or test signals created using Matlab. This is shown in Fig. 3.4. Note that all of stages I and II in this figure were computationally modelled in this project, using either Matlab, ODEON or CATT-Acoustics, as described above. In the work described in this thesis, no spatial additional processing was applied to the recorded microphone signals in stage III, i.e. the recorded signals were fed directly to the loudspeakers without further processing. This was the first method used to create audio signals in this thesis.

The use of close microphone techniques (see Fig. 3.3 and [22]) is common practice in the recording industry and include some spatial processing in stage III (e.g. panning the signal between different loudspeaker channels). In order to study the spatial effects of this kind of processing, a second method was also used to create the audio signals used in this thesis. This is shown in Fig. 3.5, where either existing anechoic recordings or test signals created in Matlab are used as the input to stage III, which includes some spatial processing. Allowing these types of processing means that the audio signals used in the thesis were not limited to only those created by modelling microphone techniques, and so allowed the investigation of a greater range of audio signals and their associated spatial cues.

### 3.3 The reproduction system and the listener

Stages V to IX comprise the reproduction system and the effect on the listener. Griesinger states [65] that the problem of calculating an indication of spatial quality can be divided into two parts: first determining the pressure signals at the ears of the listener (the binaural signals) and secondly, translating these signals into a measure of the perceived spatial quality. However, this does ignore the presence of other cues which can affect the listener's perception of the reproduced audio, such as the listener's awareness of their head movements and cues from other senses, such as visual cues, which can change the context of the binaural signals. Despite this, the binaural signals remain the dominant cue for the perception of reproduced audio, and so this thesis will concentrate on the cues present in the binaural signals when modelling the perception of reproduced audio. The task of determining the binaural signals can be further decomposed. First the transmitted signals have to be decoded into signals which can be fed to the loudspeakers, and then it is necessary to determine the transfer functions from each loudspeaker in the reproduction system to each of the listener's ears. Stage V in the framework, decoding the transmitted signals, varies between different reproduction systems. This stage is simplest when the transmitted signals consist of the analogue signals to be fed directly to the loudspeakers, although increasingly the transmitted signals are digital, requiring conversion to analogue signals. It is not always the case that the transmitted signals are to be fed directly to the loudspeakers, as audio data can be transmitted using formats such as sum-and-difference signals [21, 45, 61] or B-format ambisonic signals [60], which require that the loudspeaker feed signals be decoded from the transmitted signals. Similarly, audio is frequently distributed in formats involving the use of compression algorithms incorporating perceptual models, such as MPEG audio layer 3 [67], which again require that the signals fed to the loudspeakers be decoded from the transmitted data. The next two sections discuss the problem of determining the binaural signals and the translation to the perceptual domain and extracting the measures.
Anechoic recordings
Test signals
(pink noise, sine tones)

Loudspeaker
Acoustical environment

Microphone array
Processing
Transmission

I II III IV

Figure 3.4: Block diagram of the first four processes of audio reproduction (see Fig. 3.1) for the first method of creating stimuli that was used in this thesis. Stages I and II are modelled using software and no spatial processing is applied in stage III to the simulated recorded signals.

Figure 3.5: Block diagram of the first four processes of audio reproduction (see Fig. 3.1) for the second method of creating stimuli that was used in this thesis. Existing anechoic recordings and test signals are input to stage III, which includes any spatial processing to the signals.
3.3.1 Binaural signals and Head-Related Transfer Functions (HRTFs)

The signals at the ears of the listener provide the dominant cues used by the listener in the perception of the spatial qualities of sound. Indeed, many of the cues depend on the differences between the signals at the listener's left and right ears, such as Interaural Time Differences (ITDs), Interaural Intensity Differences (IIDs) and the Interaural Cross Correlation (IACC). These differences between the signals at the two ears are caused by the diffraction and reflection of the head, pinnae and torso, and this can be modelled by a linear invariant system, where the output is called the Head-Related Transfer Function (HRTF) [18]. In practice the HRTF of a subject is determined by measuring the signals received at both ears to obtain its time-domain equivalent, the Head-Related Impulse Response (HRIR) [107].

One aspect of HRTFs is that the functions vary between individual listeners due to the unique size and shape of each listener's head, pinnae and torso. Dummy heads with microphones in the ears have been used to make binaural recordings. There are a number of such dummy heads that are commercially available, for example the Knowles Electronics Manikin for Acoustic Research (KEMAR) dummy head. Although the HRTFs obtained by recording using a dummy head only match the dummy head itself, they do enable standard sets of HRTFs to be created, for example by Gardner and Martin [53, 55]. Experiments performed by Möller et al. [106] and Minnaar et al. [105] show that binaural recordings preserve the localisation cues, and that listeners were able to localise most accurately with binaural recordings made in their own ears. The next most accurate localisations were with binaural recordings made in other people's ears, and the worst localisations were with recordings made using commercially available dummy heads. It was also found that there were significant differences in the ability of listeners to localise sound sources between recordings made using different dummy heads. One method of calculating HRTFs is to approximate the listener's head with a sphere [1, 40, 108]. This has the advantage that large databases of measured HRTFs are not required, although the HRTFs calculated using this method are only approximations. Algazi et al. [1] describe optimising the radius of the spherical head model for individual listeners to improve the performance of the model.

The HRTFs required to model the binaural signals in a listener created by an audio reproduction system depend on a number of different factors. These include the location and orientation of the listener and loudspeakers, both relative to each other and also relative to any reflective surfaces in the reproduction environment. As stated previously, the HRTFs depend on the size and shape of the listener's head, pinnae and torso. The HRTFs also depend on the directivity and frequency response of the loudspeakers and the acoustic characteristics of the reproduction environment, which affect the reflections of the original signal reaching the listener's ears. Once all the factors affecting the HRTFs were defined, two methods of calculating the HRTFs were used.

The first method used to calculate the HRTFs was to interpolate the Gardner-Martin HRTF database [53, 55], which was used to model anechoic reproduction environments. First the angles from each ear to the loudspeaker are calculated. A HRIR is then interpolated at this angle from the Gardner-Martin HRTF database. This interpolation is performed in the frequency domain in order to ensure that the amplitude response and phase response are interpolated smoothly for all frequencies. This is done by performing a Fast
Fourier Transform (FFT) on the seventy-two HRIRs in the Gardner-Martin database which correspond to an elevation of 0° (corresponding to azimuths every 5°). The phase and magnitude of each point in the IIRTFs are then calculated. For each different frequency, the magnitudes are interpolated across the seventy-two IIRTFs using cubic splines. The same procedure is applied to the phases of the HRTFs, interpolating across the seventy-two HRTFs using cubic splines. The inverse FFT is then applied to the interpolated IIRIR to give the interpolated HRIR.

The angles used in the interpolation of the Gardner-Martin IIRTF database were not the angles from the centre of the head to the source, but instead the angles from each ear to the source. This method was used by Supper [158] in the calculation of the IID and ITD values from the Gardner-Martin IIRTF database and ensures that the interpolated HRIRs are not only limited to having sources only 1.4m away from the listener. The angles from the left and right ears to the source are calculated as

\[
\theta_L = \arctan \left( \frac{r \sin \theta + d}{\cos \theta} \right), \quad \theta_R = \arctan \left( \frac{r \sin \theta - d}{\cos \theta} \right),
\]

where \( \theta_L \) is the angle from the left ear to the source, \( \theta_R \) is the angle from the right ear to the source, \( r \) is the distance from the centre of the dummy head to the source (1.4m in the Gardner-Martin IIRTF database) and \( d \) is the distance of each ear from the centre of the head (0.076m for a KEMAR dummy head). This is illustrated in Fig. 3.6. These are the same formulae as used by Supper [158] and use the approximation of a transparent head (in Fig. 3.6 it can be seen that the dotted line from the left ear to the source passes through the dummy head).

The fact that the Gardner-Martin database was created using sources all at a distance of 1.4m from the centre of the dummy head also means that if the source is not at a distance of 1.4m from the listener’s head then the delay and attenuation of the binaural signals due to the sound propagation need to corrected when
using interpolated HRIRs. The distances from each ear to the loudspeaker are calculated. The distance from the left ear to the source in the Gardner-Martin database for the angle $\theta_L$ is calculated. Similarly, the distance from the right ear to the source in the Gardner-Martin database for the angle $\theta_R$ is calculated. These two distances will not be exactly 1.4m due to the radius of the dummy head. This gives a pair of distances for each ear: one being the distance of the sources form the ear in the Gardner-Martin database and the other being the distance of the loudspeaker to the ear in the reproduction environment. For each ear, the delay in the calculated binaural signal is calculated as the time taken for the sound to travel the difference between this pair of distances. Similarly, the attenuation due to the propagation of the sound through space is calculated from this pair of distances. The signal at each ear, $S_e(t)$, is therefore given by

$$S_e(t) = \frac{d_{GM}(\theta, \theta, \theta, \psi)}{d_{e, s}} \text{HRIR}(\theta, \psi) \ast S_s(t + (d_{GM}(\theta, \theta, \theta, \psi) - d_{e, s})/c)$$

(3.3)

where $\text{HRIR}(\theta)$ is the interpolated impulse response for the angle $\theta$, $d_{e, s}$ is the distance of the loudspeaker from the ear, $d_{GM}(\theta)$ is the distance of the ear to the source in the Gardner-Martin database for the angle $\theta$, $S_s(t)$ is the loudspeaker signal at time $t$, $c$ is the speed of sound and $\ast$ represents convolution. Note that, as $\text{HRIR}(\theta, \psi)$ was calculated in the frequency domain in order to ensure the smooth interpolation of the amplitude and phase responses, Equation 3.3 is actually implemented in Matlab as

$$S_e(t) = \frac{d_{GM}(\theta, \theta, \theta, \psi)}{d_{e, s}} \mathcal{F}^{-1}\left(\text{HRTF}(\theta, \psi) \ast \mathcal{F}\left(S_s(t + (d_{GM}(\theta, \theta, \theta, \psi) - d_{e, s})/c)\right)\right)$$

(3.4)

where $\text{HRTF}(\theta)$ is the interpolated transfer function for the angle $\theta$, $\mathcal{F}$ is the discrete Fourier transform, $\mathcal{F}^{-1}$ is the inverse discrete Fourier transform and the other symbols are the same as for Equation 3.3. The discrete Fourier transform and its inverse were implemented as FFTs in Matlab.

The second method of calculating binaural signals was to calculate the HRIRs using either the CATT-Acoustics [42, 41] or ODEON [36] acoustics modelling software. This allowed reflective reproduction environments to be modelled. Again, note that anechoic environments are a subset of the acoustic environments that can be modelled in both these software packages. This is similar to the modelling of the capture environment and, again, it was faster and more convenient to model the anechoic capture environments directly in Matlab, as described above, rather than generate the impulse responses using either CATT-Acoustics or ODEON.

### 3.4 The co-ordinate system used in the thesis

A polar co-ordinate system is used to describe the locations of listeners and loudspeakers in the reproduction environment. Only the horizontal plane is considered in this project, so each point in the this plane is uniquely defined by a co-ordinate pair $(\theta, R)$, where $\theta$ is the azimuth, measured in a clockwise direction, and $R$ is the distance from the origin, as shown in Fig. 3.7). The origin of the co-ordinate system is located at the centre of the listening area and is referred to as the sweet spot.
Figure 3.7: The polar co-ordinate system used in this project. The y-axis is the direction towards the front of the listening area.

Figure 3.8: (a) The orientation of the listener's head. (b) Diagram showing $\omega_0$, the listener angle when the listener is facing towards $0^\circ$, in relation to the source and listener positioned at $(\theta_S, R_S)$ and $(\theta_L, R_L)$ respectively.
Consider a sound source and a listener both positioned in the listening area. The azimuth of the source relative to the listener is referred to as the listener angle. For each source location the listener angle will vary as the position of the listener varies. This contrasts with the angle from the sweet spot to the source location, referred to as the sweet spot location, which is constant for any given source location. If the sound source and the listener are located at \((\theta_s, R_s)\) and \((\theta_l, R_l)\) respectively, then the listener angle \(\omega\) is calculated as

\[
\omega_0 = \begin{cases} 
\arctan \left( \frac{R_s \sin \theta_s - R_l \sin \theta_l}{R_s \cos \theta_s - R_l \cos \theta_l} \right), & \text{if } R_s \cos \theta_s \geq R_l \cos \theta_l \\
\arctan \left( \frac{R_s \sin \theta_s - R_l \sin \theta_l}{R_s \cos \theta_s - R_l \cos \theta_l} \right) + \pi, & \text{if } R_s \cos \theta_s < R_l \cos \theta_l
\end{cases}
\]

(3.5)

\[
\omega = \omega_0 - \alpha
\]

(3.6)

where \(\alpha\) is the angle towards which the listener is facing, as shown in Fig. 3.8(a) and \(\omega_0\) is the listener angle when the listener is facing in the 0° direction, as shown in Fig. 3.8(b).

3.5 The reproduction systems used in the thesis

This section contains descriptions of the three standard audio reproduction systems which were investigated in the research described in this thesis. All of the reproduction systems were implemented in Matlab using the method described in Section 3.3. This section is divided into two: the first subsection discusses Two Channel Stereo and Five Channel Stereo and the second discusses Wave Field Synthesis.

3.5.1 Two Channel Stereo (TCS) and Five Channel Stereo (FCS)

Historically, the most common reproduction system to deliberately include spatial attributes in the reproduced sound is Two Channel Stereo (TCS). The loudspeaker layout of TCS consists of two loudspeakers arranged at angles of 30° and -30° from the listener, as shown by the two loudspeakers labelled L (left) and R (right) in Fig. 3.9. The two loudspeakers and the listener together form an equilateral triangle, which is accepted as the optimum configuration for TCS [46].

The loudspeaker layout and channel configuration of Five Channel Stereo (FCS) is described in the ITU-R BS.775-1 standard [133] and is illustrated in Fig. 3.9. It is also known as 3-2 stereo and 5.1 surround. This reproduction system consists of five channels with full bandwidth and a sixth channel of limited bandwidth for Low Frequency Effects (LFE). This LFE channel is not used in all FCS recordings, and there is not a complete consensus in the literature as to how very low frequencies are localised by listeners [94, 103], especially in small rooms when room modes have to be taken into consideration [118, 163]. The LFE channel was omitted in the research described in this thesis for these two reasons.

The five full bandwidth channels in FCS can be divided into the front channels and the rear channels. The front channels correspond to loudspeakers positioned at -30°, 0° and 30° (labelled L (left), C (centre) and R (right) in Fig. 3.9). The rear two channels correspond to loudspeakers at -110° and 110° (labelled LS (left...
Figure 3.9: The loudspeaker layout used for FCS in this thesis. This conforms to the ITU-R BS.775-1 standard [133]. The layout has left-right symmetry, so only the angles for the loudspeakers on the right are shown on the plot. The loudspeaker layout for TCS consisted of only the left and right loudspeakers (marked L and R in the plot).

surround) and RS (right surround) in Fig. 3.9). The distinction between the front and rear loudspeakers is important, as the design of this loudspeaker configuration was intended primarily front-based sound scenes [139]. Indeed, the fact that FCS will often be used to accompany the images from a television or monitor is reflected in the title of the ITU-R BS.775-1 standard: "Multichannel Stereophonic Sound System With and Without Accompanying Picture" [133]. For these types of sound scenes, the front three channels are used to provide the main sound image, while the rear two channels are intended to supply ambience, the reflected energy corresponding to the room impression and special effects, e.g. explosions. What is important to note here is that the FCS configuration was not designed to provide accurate localisation for all 360° angles.

Despite this, a number of methods of panning sources for more than two channels have been developed and applied to the FCS configuration. The most common method of panning sources to any position with FCS is to use an amplitude panning law developed for two channels, such as the constant power panning law, and apply this to adjacent pairs of loudspeakers [73]. Pulkki [127] developed vector based amplitude panning (VBAP) which can be used with arbitrary loudspeaker positions, as long as the loudspeakers are equidistant from the listener. If only the central listening position is considered then this includes FCS. In two dimensions, VBAP works 'pairwise', i.e. using the two loudspeakers either side of the intended angle, as described above. Pulkki also extended the VBAP panning method to three dimensions, using amplitude panning between three loudspeakers to position the source. Panning laws based on Ambisonic principles have been applied to FCS [44, 62, 111]. Ambisonics uses a Fourier-Bessel decomposition of the sound-field, that is, expressing the sound-field as a series of spherical harmonics [43] and was first developed by Gerzon [58, 60]. Gerzon’s motivation was that the first and second order spherical harmonic decompositions of the sound-field would produce the correct low frequency “velocity” cues and high frequency energy cues...
according to Makita's work [91]. Two studies [96, 164] both concluded that the pairwise constant power panning law performed the best in a comparison of different multichannel panning laws.

### 3.5.2 Wave Field Synthesis

Wave Field Synthesis (WFS) is a method of reproducing sound-fields using arrays of loudspeakers. It was first developed in the late 1980s [15, 16] and is based on Huygen's principle. This states that a wave front may be considered as a secondary source distribution, which implies that the sound wave propagating from a wave front could be generated either by the original source of the wave, or by a distribution of secondary sources situated on the wave front. The idea of Huygens' Principle is expressed mathematically with the Kirchhoff-Helmholtz Integral [15, 43]:

\[
p(\vec{r}) = \int \int_{\partial A} \left[ \nabla p_{0} \cdot \vec{n} - \frac{\vec{R}}{R} \cdot \vec{n}(1 + jkR) \frac{p_{0}}{R} \right] \frac{e^{-jkR}}{4\pi R} dS_{0}
\]

where \( A \) is a given volume, \( p(\vec{r}) \) is the pressure at a point \( \vec{r} \in A \), \( \partial A \) is the boundary of \( A \), \( p_{0} \) is the acoustic pressure on the boundary \( \partial A \), \( \vec{r}_{0} \) is a point on the boundary \( \partial A \), \( dS_{0} \) is an area on the boundary \( \partial A \) centred on \( \vec{r}_{0} \), \( \vec{n}(\vec{r}_{0}) \) is the unit vector perpendicular to the boundary \( \partial A \), \( k \) is the wave number, \( \vec{R} = \vec{r}_{0} - \vec{r} \), and \( R = |\vec{R}| \).

The Kirchhoff-Helmholtz Integral states that the pressure inside a closed volume is uniquely defined by the pressure and pressure gradient on the boundary of the area. The basic idea of WFS is to use microphones to record the pressure and pressure gradients of a sound field on the boundary of a volume, and then to use loudspeakers situated on the boundary of a similar volume to replay the recorded signals in order to reproduce the sound-field inside the original volume. In practice, the loudspeakers used for the reproduction of the sound-field do not have to lie on a similar surface as the microphones, but can instead be on the surface of a volume within the original volume, and the signals captured by the microphones can be extrapolated to the signals fed to the loudspeakers by using a processor simulating the wavefront propagation. The stationary phase approximation [116, 161] is used to derive equations for two-dimensional realizations of WFS systems using only two-dimensional arrays of microphones and loudspeakers.

WFS was originally conceived as a means of acoustic control for concert halls [15, 16] rather than as a surround sound reproduction system in its own right. The proposed use of WFS for acoustic control in concert halls was to have an array of microphones close to the stage in order to capture the waves propagating away from the stage. By extrapolating the signals and using an array of loudspeakers it is possible to amplify the sounds emanating from the stage while still preserving the shape of the wavefronts, so the origin of the amplified sounds will still appear to be from the stage, rather than the loudspeakers being used as secondary sources. In addition to WFS being used to reinforce the direct sound-field, it can also be used to simulate reflections from walls by extrapolating the signals from the microphone array using a processor to simulate the wavefront from a virtual source situated behind the reflective wall [16].
Spatial aliasing

In WFS systems transducers are required on the boundary of the enclosed area both to record the pressure and pressure gradients and also to reproduce the sound-field. Ideally, these transducers would be a continuous distribution, but in practice, discrete distributions of transducers have to be used, i.e., arrays of microphones and loudspeakers. As a consequence, the spatial sound-field is being sampled at discrete points, and so spatial aliasing can occur [43, 116]. The frequency of the spatial sampling, $f_{\text{sample}}$, is related to the distance between the discrete transducers, $\Delta_{\text{sample}}$, by $\Delta_{\text{sample}} f_{\text{sample}} = c$, where $c$ is the speed of sound. Using this and the Sampling Theorem [145], which states that aliasing occurs above the frequency equal to half the sampling frequency, gives the aliasing frequency (the frequency above which spatial aliasing occurs) as

$$f_{\text{aliasing}} = \frac{f_{\text{sample}}}{2} = \frac{c}{2 \Delta_{\text{sample}}}$$

(3.8)

Start [152] has demonstrated that directional localisation is not severely impaired by spatial aliasing when the aliasing frequency is greater than 1.5 kHz.

Microphone arrays for WFS

Hulsebos et al. [74] have researched into the effects of using different types of microphone arrays for two-dimensional WFS systems. The different shapes of microphone array that were considered were linear arrays, cross arrays (effectively two linear arrays bisecting each other at right angles) and circular arrays. In addition to this, different microphone directivity patterns were considered (monopole, dipole and hypercardioid). Hulsebos et al. conclude that the circular array is preferable for two reasons. The first is that single linear arrays are unable to reproduce the wave fronts of plane waves whose direction of propagation has no component perpendicular to the linear array (i.e., when the array is in end-fire configuration). The second reason is that sound-fields reproduced from signals recorded using a linear array exhibit diffraction near the ends of the array.

One aspect of the Kirchoff-Helmholtz equation is that the space inside the loudspeaker array (where the recreation of the sound-field is attempted) should be free of sources. However, Verheijen [161] showed that it is possible to approximate sound sources that are enclosed by the loudspeaker array by recording enclosed sound sources and then feeding the signals backward to the loudspeakers (i.e., reversing the signals in the time domain). This does produce the desired pattern of wave-fronts, but the waves propagate toward the enclosed sound source rather than away from it. Experiments have shown that although this produces the desired effect to some degree on a listener, the contradictory spatial localisation cues supplied to the listener mean that it is not wholly successful. The time reversal method is very close to a similar method for sound focusing proposed by Yon et al. [168, 169], which includes the effect of indirect sound resulting from reflections.

Nicol and Emerit [116] discuss how the signals to the monopole and dipole loudspeakers are not independent, and that the sound-field can be approximated reasonably accurately using just monopole loudspeakers and microphones. This is achieved by using spatial weighting for the loudspeaker feeds, so that not all of the monopole loudspeakers are activated. In the case of a full WFS system using both monopole and dipole
loudspeakers in a circular array to reproduce a plane wave, the sound waves emitted by the monopole and dipole loudspeakers furthest away from the source of the plane wave are out of phase with each other and so cancel each other. Using a circular array of outward facing cardioid microphones (e.g. Daniel et al. [43]) gives the required spatial weighting, as the microphones furthest away from a sound source will have their directivity patterns facing away from the sound source and hence the signal recorded by these microphones will be attenuated. This is illustrated in Fig. 3.10.

Implementing WFS in Matlab

A 32-channel WFS system with a circular loudspeaker array was modelled in Matlab. The loudspeaker signals were calculated by modelling a circular array of outwards facing cardioid microphones, as described above, using the method of modelling the sound-field (stages I and II) in the capture environment described in Section 3.2. The calculation of the binaural signals at listener positions in the reproduction environment (stages VI and VII) was implemented as described in Section 3.3.1. Before calculating binaural signals to use with perceptual models, the Matlab implementation of the WFS model was checked by calculating plots of the instantaneous pressure in the reproduced sound-fields corresponding to original sound-fields consisting of only a single frequency. The pressures at different points in the reproduced sound-field were calculated using the same method as used for the capture of the original sound-field (described in Section 3.2), effectively using an omnidirectional microphone to calculate the pressure at each point.

The Matlab implementation of the 32-channel WFS system was informally verified by recreating all of the WFS plots of reproduced sound-fields in the paper by Daniel et al. [43]. The left hand plot of Fig. 3.11 shows the instantaneous pressure of the sound-field created by the WFS system when attempting to reproduce a plane wave with frequency 600Hz at an angle of 20.1°. It can be seen that WFS has recreated the curvature of the plane wave for most of the listening area. The plotted pressures and both the error contours of this plot closely match the corresponding plot in the Daniel et al. paper. In addition to showing that the WFS system was modelled correctly in Matlab, this also shows that the method of simulating the capture of sound-fields by modelling microphones (described in Section 3.2) has been correctly implemented in Matlab. The right hand plot in Fig. 3.11 shows the reproduced sound-field created by the 32-channel WFS system corresponding to the original sound-field created by an omnidirectional source at (20.1°,2m) emitting a 600Hz sine wave. As in the left hand plot, it can be seen that the curvature of the original sound-field has been recreated for most of the listening area. This shows that the 32-channel WFS system is able to reproduce the curved wave fronts of sounds produced by sources at finite distances.
1.5 Summary

Figure 3.10: The directivity patterns of a circular array for recording the signals for an eight channel WFS system.

Figure 3.11: Plots of the instantaneous pressure of the reproduced sound-fields created using 32-channel WFS. The left hand plot shows the reproduced sound-field for a sinusoidal plane wave with an angle of incidence of 20° (shown by the red line) and a frequency of 600Hz. The right hand plot shows the reproduced sound-field for an original sound-field for an omnidirectional source located at (20.1°, 2m) emitting a sine wave with a frequency of 600Hz. The red line on the right hand plot shows the direction of the source. The blue contours show the areas where the error between the original sound-field and the reproduced sound-field is 20% and the yellow contours show the areas where the error is 59%. The purple circle in the middle of the reproduced sound-field illustrates a listener’s head, although the diffraction effect of having a listener within the sound-field is not modelled.
3.6 Summary

This chapter has described the framework for modelling reproduced audio which was used for the research described in this thesis. The Matlab implementation of simulating the capture of the original sound-field was described. Similarly, there was a discussion of the calculation of binaural signals in the reproduced sound-field. The coordinate system used in the thesis was then described, followed by a discussion of the standard reproduction systems which are investigated in the thesis.
Chapter 4

Listening test experiments

This chapter contains details of the three listening tests that were undertaken as part of the investigation into perceived spatial attributes and the development of an artificial listener. The listening tests were undertaken in order to provide data against which the models of directional localisation and source width could be validated.

At the times when the listening tests were undertaken two existing models of perceived spatial attributes had already been partly integrated into the model framework: the directional localisation model developed by Supper [158] and the source width model developed by Mason [100]. As both these models have already been validated and published, the data from the listening tests was intended to be used only for the validation of the artificial listener model. However, as the integration of both the Supper and Mason models into the “source to percept” framework proved more complicated than initial expectations, the data obtained from the listening tests was also used to inform the development of both the “source to percept” model and the subsequent investigation of spatial attributes.

In the case of the integration of the Supper model, the data from the first listening test was initially used to calibrate and develop a heuristic method for integrating the results of the Supper model into the model framework. However, it became apparent that the heuristic method of integrating the Supper model developed using the first listening test results was not able to predict the directional localisation results obtained from the second and third listening tests. Adapting the heuristic model to be able to predict the results of all three listening tests became increasingly hard to justify, as the heuristic model became increasingly complicated. This also started to raise concerns that further development of the heuristic model to predict the results from all three listening tests would result in the model being over-fitted to the listening test data, meaning that the model would prove not to be robust when used to predict directional localisations from signals and reproduction systems other than those included in the stimuli for the three listening tests. Consequently the decision was made to modify the Supper model itself rather than just modifying the method used to integrate the results from the original Supper model.
### Table 4.1: Description of the original signals used as the basis for the stimuli in the three listening test experiments.

<table>
<thead>
<tr>
<th>Name</th>
<th>Duration (seconds)</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>200Hz sine tone</td>
<td>10</td>
<td>Faded in over 5 seconds then faded out over 5 seconds.</td>
<td>Generated in Matlab</td>
</tr>
<tr>
<td>600Hz sine tone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2kHz sine tone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pink noise</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cello</td>
<td>4.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical guitar</td>
<td>2.5</td>
<td>Musical phrase.</td>
<td><em>Music from Archimedes</em> Bang and Olufsen [3]</td>
</tr>
<tr>
<td>Trumpet</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>3</td>
<td>Faded in and out.</td>
<td></td>
</tr>
<tr>
<td>Male speech</td>
<td>1.5</td>
<td>The words “One, two”.</td>
<td></td>
</tr>
<tr>
<td>Female speech</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

 Rather than being used to calibrate the Supper model, the listening test results were instead used to identify the stimuli for which the original Supper model failed to predict localisation azimuth. The intermediate results at the different processing stages of the Supper model were inspected, from which were determined the reasons for the Supper model’s failure to predict the localisation azimuths. Although the results from the listening tests were used to identify the areas of the original Supper model which were not performing well, the modifications made to these areas of the Supper model were not calibrated by the listening test results.

### 4.1 The original signals used in the listening test experiments

Ten different audio signals were used as the basis for all the stimuli used in the three listening test experiments. These ten signals were subjected to additional processing to create the listening test stimuli. This additional processing is described in the sections corresponding to each listening test. The details of these ten audio signals are contained in Table 4.1. Rumsey [140] differentiates between those listening test experiments where the aim is to investigate auditory perception and those listening test experiments where the aim is to evaluate an audio product, i.e. to evaluate an element in an audio reproduction system. One of the reasons that Rumsey differentiates between these two types of listening test is that the purpose of the listening test affects the decision as to the stimuli to be used.

In studies to investigate auditory perception the stimuli are often chosen to be tones or noise. These simple stimuli have the advantage that it is relatively easy to control all the aspects of these signals and they are also easily reproduced by other researchers wishing to verify or expand upon results in the literature obtained using these signals. However, even in this context these signals can have disadvantages. Such signals can
sometimes be too simple and can fail to create the desired perception in the listener (although this can be
in itself a valid and interesting result for an investigation into auditory perception, just not necessarily the
intended result). This is the case when sine tones are used in localisation experiments [70], and, indeed, is
also shown in the outcome of the first experiment described in this chapter.

In comparison, listening test experiments which are intended to evaluate all or part of an audio reproduction
system have different criteria for the stimuli. Here the aim is to evaluate the performance of the reproduction
system in the context of how the system will typically be used. Hence the stimuli should reflect this and
exhibit similar characteristics as typical program material, or indeed, consist of actual program material.
Rumsey describes stimuli that satisfy these conditions as being *ecologically valid*. These stimuli are typically
much more complicated than tones and noise. The different aspects of this type of stimulus are consequently
much more difficult to control. This can lead to a stimulus exciting multiple perceptual attributes in the
listener, which is not always desirable. On the other hand, some perceptual attributes rely on the presence
of multiple cues in the sound, and also the context of these cues. For listening test experiments that aim to
investigate these perceptual attributes the stimuli have to be ecologically valid in order to create the desired
perceptual attributes in the listeners. Note that there is not a clear distinction between these two types of
stimuli: some program material has very similar characteristics to tones or noise, and conversely stimuli can
be very artificially contrived and yet have very similar characteristics to typical program material.

The aims of the listening experiments described in this chapter do not easily fall into either of the two
categories described by Rumsey. The original aim of the listening tests was to validate the models of
directional localisation and source width. Localisation and source width are both perceptual attributes,
which suggests the listening tests fall into the first category of listening tests, whose aim is to investigate
auditory perception. However, the context of these models is that they are intended to predict perceptual
attributes or reproduced sound. Indeed, the aim of developing and the investigating the behaviour of these
models of individual perceptual attributes is that eventually they will become components of a larger model
which predicts the spatial quality of reproduced sound. As the listening tests in this chapter fall into both
of the categories described by Rumsey, so the ten signals used as the basis of all the stimuli in the listening
tests can be divided into two analogous groups. The first group consists of three sine tones and also pink
noise. The second group consists of four musical signals and two speech signals, all taken from the *Music
from Archimedes* [3]. This range of signals was chosen so that the effects of both simple signals and more
ecologically valid signals could be investigated.
4.2 The first listening test

This section describes the first listening test that was undertaken to validate the computational model of directional localisation. In addition to describing the design and methodology of the listening test, this section also details the results of the listening test and the issues that arose during the course of the experiment.

4.2.1 The equipment used

Table 4.2 contains details of the seven loudspeakers used in the first listening test experiment, which were arranged as shown in Fig. 4.1. The signals used in the listening test were controlled from the Max/MSP environment on an Apple Mac computer. These were output through a Digi 001 hardware interface, then through an ADAT optical cable to a Yamaha 02R digital mixing desk. The connections are shown in Fig. 4.2. The five loudspeakers in the standard FCS configuration were Genelec 1032 active monitor loudspeakers, while the extra two were passive Tannoy loudspeakers. Consequently, whereas the five Genelec loudspeakers were fed with the signals directly from the mixing desk, there was an additional amplification stage using Quad QD4240 power amplifiers between the mixing desk and the two passive Tannoy loudspeakers.

<table>
<thead>
<tr>
<th>Name</th>
<th>Loudspeaker</th>
<th>Active</th>
<th>Distance (metres)</th>
<th>Angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>Genelec 1032</td>
<td>✓</td>
<td>2.2</td>
<td>-30</td>
</tr>
<tr>
<td>Right</td>
<td>Genelec 1032</td>
<td>✓</td>
<td>2.2</td>
<td>30</td>
</tr>
<tr>
<td>Centre</td>
<td>Genelec 1032</td>
<td>✓</td>
<td>2.2</td>
<td>0</td>
</tr>
<tr>
<td>Rear Left</td>
<td>Genelec 1032</td>
<td>✓</td>
<td>2.2</td>
<td>-110</td>
</tr>
<tr>
<td>Rear Right</td>
<td>Genelec 1032</td>
<td>✓</td>
<td>2.2</td>
<td>110</td>
</tr>
<tr>
<td>Extra Left</td>
<td>Tannoy</td>
<td>×</td>
<td>3</td>
<td>-60</td>
</tr>
<tr>
<td>Extra Right</td>
<td>Tannoy</td>
<td>×</td>
<td>2</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 4.2: Details of the seven loudspeakers used in the first listening test experiment. The first five loudspeakers were arranged in the standard FCS configuration. All the distances are measured from the centre of the listening area.

An acoustically transparent curtain was suspended from the ceiling to prevent the subjects from being influenced by seeing the positions of the five loudspeakers to the front of the listener. A scale was placed in the front hemisphere inside the curtain to give the listeners a reference for the different angles at which they might perceive the sound to be coming from. The scale ranged from -50 at -90° from the centre of the listening area to 50 at 90°. The scale is in the range -50 to 50 rather than -90 to 90 to minimise the possibility of the listening test subjects confusing the scale with degrees, as the angle corresponding to a given number on the scale varies as the listener position varies. For instance, when the listener is at the sweet spot then the number 0 (zero) on the scale corresponds to 0°, straight ahead. If the listener position is to the right of the sweet spot then zero on the scale corresponds to an angle to the left.
Figure 4.1: The equipment layout used in the first listening test. The five loudspeakers of the standard FCS layout are shown in black on the dashed circle, which has a radius of 2.2m. Two extra loudspeakers are shown in blue, located at (−60°, 3m) and (15°, 2m). The three green circles show the three listening positions that were considered in the experiment, located at (0°, 0m), (−135°, 0.5m) and (90°, 0.5m). The wavy line represents the acoustically transparent curtain and the half-octagon inside shows the scale used to assist the listeners in the localisation. The black frame of the figure shows the walls of the listening room.

Figure 4.2: The connections of the equipment used in the first listening test.
4.2.2 The design of the stimuli

Forty different signals were used as the stimuli. These forty signals can be divided into four groups of ten according to the signals that were used as the original source. These four groups consist of sine waves, pink noise, solo musical instruments and speech.

The sine wave signals were faded in over 10 seconds and then faded out over 10 seconds in order to eliminate the effects of onsets and offsets on the listener’s localisation. Each sine wave signal had a frequency of either 200Hz, 600Hz or 2000Hz. The decision to use sine waves as one set of stimuli was made because the model was first developed in the frequency domain and was largely influenced by Wave Field Synthesis and High Order Ambisonic reproduction systems. Sine waves were used as stimuli in the listening test to investigate how the model would localise tones of single frequencies. The decision to include sine tones was made knowing that sine tones are notoriously difficult for people to localise according to the literature. The pink noise used in the listening test was generated using Matlab. All the musical and speech signals were sourced from the Music from Archimedes audio CD produced by Bang and Olufsen [3], which consists of anechoic recordings.

Each of these four groups of signals can be further subdivided into three groups according to how the original signals were processed. In the first group, the original source signals are replayed directly over a single loudspeaker with no further processing. These are referred to as the single loudspeaker stimuli. This group of stimuli was included to provide a reference. When a single loudspeaker is used then the intended direction of localisation will match the actual direction of localisation; this is not necessarily true when a stimulus has been created using some form of signal processing and multiple loudspeakers.

The second group consists of the signals created by using Matlab to model the original signal being played through an omnidirectional loudspeaker in an anechoic environment and recorded using a standard microphone technique. Three different microphone configurations were used, corresponding to mono, TCS and FCS reproduction systems. For mono, a single omnidirectional microphone was placed in the centre of the listening area. For TCS, two microphones with cardioid directivity were arranged in the ORTF configuration [157]. For FCS stereo, an array of cardioid microphones was used [167]. The microphone configurations for TCS and FCS stereo are shown in Fig. 4.3. In all of the above cases, the signal captured by each microphone was used to drive the corresponding loudspeaker in the standard FCS loudspeaker setup in the listening room. These are referred to as the anechoic recording stimuli. This group of stimuli was included because each of its stimuli uses multiple loudspeakers (with the exception of mono), the signal generation is still grounded in the physical world and the microphone techniques used to generate the signals are also used in real world recordings. Finally, the framework of using virtual microphone techniques allows the signals for different reproduction systems (i.e. different loudspeaker configurations) to be generated in a similar fashion, thus facilitating comparison between different reproduction systems in arbitrary configurations.

CHAPTER 4. LISTENING TEST EXPERIMENTS | 58
The third group of signals used in the listening test is similar to the second group of signals, except that an acoustically reflective environment was modelled in Matlab by using impulse responses generated by CATT-Acoustics. These are referred to as the reflective recording stimuli. The echoic environment that was modelled had the same dimensions as the listening room used in the listening tests, but had different absorption and reflection coefficients for the surfaces in the room, resulting in a more reverberant environment. The third group of stimuli was included for the same reasons as the second group of stimuli, and also because it allowed the effects of room acoustics to be investigated. Details of all forty stimuli are given in Table 4.3. All the stimuli were loudness equalised using an NTI Acoustifyzer AL1 sound level meter to 78db SPL (A-weighted, slow), the level recommended in ITU-R BS 1116 [131]. The levels of the stimuli were then adjusted to give equal values of perceived loudness as calculated by Moore et al.’s loudness model [110]. Finally, the stimuli were checked by ear to ensure that the loudness equalisation was successful and the levels of any stimuli which were judged to have different perceived loudness to the others were adjusted.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Original signal</th>
<th>Reproduction system</th>
<th>Reflective</th>
<th>Distance (metres)</th>
<th>Angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200Hz sine tone</td>
<td>Single</td>
<td>x</td>
<td>2.2</td>
<td>-18 -30 -40</td>
</tr>
<tr>
<td>2</td>
<td>600Hz sine tone</td>
<td>Single</td>
<td>x</td>
<td>3</td>
<td>-50 -60 -64</td>
</tr>
<tr>
<td>3</td>
<td>2kHz sine tone</td>
<td>Single</td>
<td>x</td>
<td>2</td>
<td>21 15 1</td>
</tr>
<tr>
<td>4</td>
<td>600Hz sine tone</td>
<td>Mono</td>
<td>✓</td>
<td>2.2</td>
<td>33 30 17</td>
</tr>
<tr>
<td>5</td>
<td>200Hz sine tone</td>
<td>TCS</td>
<td>✓</td>
<td>2.2</td>
<td>-18 -30 -40</td>
</tr>
<tr>
<td>6</td>
<td>600Hz sine tone</td>
<td>FCS</td>
<td>✓</td>
<td>2.2</td>
<td>33 30 17</td>
</tr>
<tr>
<td>7</td>
<td>2kHz sine tone</td>
<td>TCS</td>
<td>✓</td>
<td>3</td>
<td>58 60 54</td>
</tr>
<tr>
<td>8</td>
<td>200Hz sine tone</td>
<td>TCS</td>
<td>✓</td>
<td>2.2</td>
<td>33 30 17</td>
</tr>
<tr>
<td>9</td>
<td>600Hz sine tone</td>
<td>FCS</td>
<td>✓</td>
<td>2.2</td>
<td>-18 -30 -40</td>
</tr>
<tr>
<td>10</td>
<td>2kHz sine tone</td>
<td>TCS</td>
<td>✓</td>
<td>3</td>
<td>58 60 54</td>
</tr>
</tbody>
</table>

Table 4.3: (continued on next page)
Table 4.3: This table contains details of the forty stimuli used in the first listening test experiment. The details of the original signals used in the creation of the stimuli can be found in Table 4.1. The Reproduction System column shows whether the stimulus was played through a single loudspeaker (shown as “Single” in the table) or else played through a conventional reproduction system (see Section 3.5). For those stimuli played through a single loudspeaker, the distances and angles given in the table correspond to the position of the loudspeaker relative to the centre of the listening area. For the stimuli that have been created by simulating recording with conventional microphone techniques, the distances and angles in the table correspond to the position at which the source was modelled relative to the centre of the modelled microphone array. (continued from previous page)
4.2.3 The design of the user interface

The user interface was written in Max/MSP. The reasons that Max/MSP was chosen as the platform on which to implement the user interface were, firstly, its ease of controlling multichannel digital audio and, secondly, that it is the most common platform in the Institute of Sound Recording research group at Surrey. Two screen shots of the user interface are shown in Fig. 4.4. There are four elements in the user interface, as shown in the left half of Fig. 4.4: a box at the top of the interface which shows a plan of the listening area during the test, a box containing the number that the subject has selected, a pink button and a green button.

At the beginning of the test the box at the top of the user interface is blank and the subject presses the pink button to begin the test (see the left half of Fig. 4.4). Once the pink button has been pressed the first recording is played and a plan view of the listener's head, the scale and the acoustically transparent curtain appears in the display box at the top of the interface. The number 888 also appears below the graphics box next to the caption “Number on scale”. The number 888 is used as the initial value so that the cases where the test subject does not enter a value can be easily identified and removed from the results: the values entered by the subjects using either the mouse or the keyboard are limited to the range -100 to 100, which includes all angles around the subject. Once the subject has listened to the recording they can either enter the number on the scale using the keyboard or click inside the display box. This will cause an arrow to appear in the display box pointing away from the picture of the listener's head and the number 888 to be replaced by the number on the scale (see the right half of Fig. 4.4). The position of the picture of the

Figure 4.4: Screen shot of the Max/MSP user interface at the beginning of the listening test (above left) and during the test (above right).
The listener's head will change according to the listener's position in the listening room, i.e. when the listener is in the centre listening position the picture of the head will appear in the centre of the display box (as shown in the right half of Fig. 4.4), in the right listening position the head will appear to the right and in the left (rear) position the head will appear to the left and also lower in the display box. The direction of this arrow can be changed by dragging the mouse with the left button held down inside the display box. The number displayed below the graphics box and the arrow shown inside the graphics box are linked. The stimulus is played again by pressing the green button at the bottom of the user interface. Once the listener is satisfied with their response then pressing the red button will save it, play the next stimulus, remove the arrow from the graphics box and reset the number to 888. The pink and green buttons allow the subject to control the rate of progress through the listening test. This process is repeated until all the stimuli have been played and all the listener's responses have been recorded, at which point the display box will become blank and the message on the pink button will read "Start next listening test". The space bar and the letter "s" on the keyboard can be used as alternatives for the pink and green buttons respectively. This allows the subject to record their responses using either just the keyboard, or just the mouse, or a combination of the two. Once the test has finished then the results are saved as a text file whose name is automatically created from the time at which the test was started.

4.2.4 The listening test procedure

Before the test commenced each subject was given a printed sheet describing the intention of the test, the spatial attribute they should be describing and instructions on how to use the user interface. In addition to this the test supervisor gave a demonstration of how to use the user interface. The listener was also instructed to try to keep their head facing forward and to minimise their head movements when evaluating the direction of the sound localisation. Each session with a subject lasted about twenty minutes, the time varying slightly as the tests progressed at a rate determined by the subject.

Three different listener positions were considered for the listening tests. These are described in Table 4.4. Each different subject participated in up to three different listening tests, sitting in a different listening position for each test. Ideally each of the different listeners would have participated in three sessions, which would result in data for each of the three listening positions for each subject. However, the limited availability of the resources (particularly the listening room itself) meant that not all subjects were able to participate in three separate sessions. The numbers on the scale were transformed into angles in the analysis of the listening test results.

1If the subject changes the direction of the arrow by dragging the mouse then the number displayed in the user interface will change to the number on the scale that corresponds to the arrow direction. Conversely, if the subject types a number via the keyboard then the direction of the arrow in the graphics box will change to point at the appropriate point on the picture of the scale.

2By avoiding the task of explicitly naming the files the execution of the listening tests is made simpler for both the subject and the person running the experiments: all that has to be recorded is the time at which each listening test occurred, and this can then be matched to appropriate results file after the conclusion of the experiments. Also, using a time stamp as the name for each results file eliminates the possibility of overwriting any of the results files.
In each listening test session the subject heard eighty stimuli: the first forty were a randomised sequence of the forty different signals, and this was followed by the forty different stimuli again in a different random order. The reason that each sound was presented twice was to determine whether the subject was consistent in their responses. The order in which the stimuli were presented to each subject was randomised prior to the listening tests. Each listener had the stimuli ordered in a different sequence. This was to minimise any bias that may be introduced by having all subjects hear the sounds in the same order due to the relationships between consecutive sounds.

### 4.2.5 Results and analysis

Seventeen listeners participated in the first listening test experiment. Eleven of these performed the test three times, once at each of the three listening positions. The remaining six performed the test only once at a single listening position. Table 4.5 shows the number of listening tests performed at each of the three listening positions.

<table>
<thead>
<tr>
<th>Location relative to sweet spot</th>
<th>Listener orientation (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centre</td>
<td>0</td>
</tr>
<tr>
<td>Left</td>
<td>-135</td>
</tr>
<tr>
<td>Right</td>
<td>90</td>
</tr>
</tbody>
</table>

Table 4.4: The three listener positions used in the first listening test experiment.

A total of 3074 separate perceived angles of localisation were collected from the seventeen listeners in the first listening test. An analysis of the localisation results for the different listeners is contained in Section E.1 in Appendix E, which concluded that none of the listeners needed to be screened from the results. The mean was calculated for each test item, which gave 120 angles (the 40 stimuli for each of the three listener positions) which can then be compared with the angles calculated by the Matlab model. Figures 4.5 to 4.7 show box plots of the results obtained from the first listening test experiment for the three listening positions.

Subject 9 consistently placed the source direction directly behind the head for the sine tone stimuli. Subject 9 also verbally informed the test supervisor that this was deliberate, stating that any signals that were very
hard to localise had been placed behind the head. Every subject who participated in the listening test verbally informed the test supervisor of the difficulty of localising the sine tone stimuli. This can be seen in Figs. 4.5 to 4.7 and is consistent with the literature. The sine tones were particularly susceptible to head movements: small movements causing large differences in the perceived localisation of the sound, which may be partly attributed to the presence of room modes.

The standard deviation was calculated for each test item. These standard deviations were then averaged over the four different groups of stimuli, as shown in Table 4.6. The large value of the average standard deviation calculated for the sine tones confirms that the listeners found these stimuli hard to localise. This agrees with the literature. The average standard deviations calculated for the other three groups are similar to each other and these also concur with the literature. The results of the sine tone stimuli will be omitted from the remainder of the analysis of the listening test results due to the difficulty experienced by the subjects when localising these stimuli.

<table>
<thead>
<tr>
<th>Stimuli group</th>
<th>Standard deviation</th>
<th>Interquartile range</th>
<th>Total range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sine tones</td>
<td>32.6°</td>
<td>25.1°</td>
<td>158.5°</td>
</tr>
<tr>
<td>Pink noise</td>
<td>5.9°</td>
<td>6.9°</td>
<td>23.9°</td>
</tr>
<tr>
<td>Music</td>
<td>4.8°</td>
<td>5.4°</td>
<td>19.0°</td>
</tr>
<tr>
<td>Speech</td>
<td>4.9°</td>
<td>6.2°</td>
<td>20.1°</td>
</tr>
</tbody>
</table>

Table 4.6: The standard deviations, interquartile ranges and total ranges averaged over the four groups of stimuli from the first listening test.

Each of the forty stimuli can be considered to have an actual position. Recall that three types of processing were used to create the stimuli used in the first listening test. The first process is that the original signals were played directly through a single loudspeaker. For these stimuli the actual position is considered to be the location of the loudspeaker through which the stimulus is played. The second type of processing is to model the original signal being emitted as single omni-directional source in an anechoic environment and then modelling a conventional microphone technique to generate loudspeaker signals for either a mono, TCS or FCS reproduction system. The third type of processing is identical to the second type except that a reverberant recording environment is modelled instead of anechoic environment. The actual position of the stimuli created using the second and third processes is assumed to be identical to the position of the source relative to the centre of the microphone array in the modelled recording environment.

Figure 4.8 shows these actual positions plotted against the results from the listening test. The left hand graph shows all the results except for the sine wave stimuli and the misallocated-channel stimuli, as discussed below. This graph shows how the reproduction systems do not always place the source at its intended location. The right hand graph in Fig. 4.8 shows only the single loudspeaker stimuli, which were localised by the test subjects close to the expected locations of the loudspeaker positions.
Figure 4.5: Centre position box plots for the first listening test experiment. The grey line inside each box corresponds to the median of the recorded angles for that test item and the upper and lower edges of each box correspond to the upper and lower quartiles respectively. The upper whisker on each box plot extends to the largest recorded angle (i.e. furthest to the right) less than or equal to the upper quartile plus 1.5 times the interquartile range. Similarly, the lower whisker extends to the smallest recorded angle (i.e. furthest to the left) greater than or equal to the lower quartile minus 1.5 times the interquartile range. Any recorded angles that are either greater than the upper limit of the top whisker or less than the lower limit of the bottom whisker are classed as outliers, which are plotted as grey crosses. The first line below the box plots contains the number of the stimulus corresponding to each box plot, details of which can be found in Table 4.3. The second line shows whether the original signal used in the creation of the stimulus was a sine tone (T), pink noise (N), music (M) or speech (S). The third line below the box plots shows whether the stimulus was played back through mono (M), TCS (T), FCS (F) or a single loudspeaker (S). An R in the fourth line shows that a reflective environment was modelled in the creation of the stimulus. An A shows that either an anechoic environment was modelled or no environment was modelled in the creation of the stimulus. The vertical dotted lines separate the groups of stimuli with the same intended locations, shown by the horizontal dotted black lines. The dashed red line shows the 0° azimuth.
Figure 4.6: Left position box plots for the first listening test experiment. See the caption of Fig. 4.5 for details of how to interpret both the box plots themselves and the table below the box plots. The details of the stimuli can be found in Table 4.3.
Figure 4.7: Right position box plots for the first listening test experiment. See the caption of Fig. 4.5 for details of how to interpret both the box plots themselves and the table below the box plots. The details of the stimuli can be found in Table 4.3.
Figure 4.8: These two graphs show the mean values of the first listening test results plotted against the intended positions. The left hand graph shows all the results except the sine wave stimuli and the stimuli with misallocated channels. The right hand graph shows only the results where the stimuli were from a single loudspeaker (omitting the sine wave stimuli).

4.2.6 Discussion

Misallocation of channels

The patch that was used on the Yamaha O2R digital desk was modified from an existing patch that had been used for the playback of 5.1 surround sound recordings. The original 5.1 patch had been set up with the signals for the front three loudspeakers in the first three channels, the low frequency signal used for the subwoofer in the fourth channel and the signals for the two rear loudspeakers in the fifth and sixth channels. The sound files containing the stimuli for the listening test contained seven channels, with the seven channels allocated to the loudspeakers in the following order: Left, Right, Centre, Rear Left, Rear Right, Extra Left and Extra Right. The patch on the mixing desk was changed to match this format.

The anechoic recording stimuli were mistakenly stored using the format for the original 5.1 surround sound patch, resulting in the Rear Left and Rear Right signals being misallocated to the Rear Right and Extra Left loudspeakers respectively. Only the anechoic recording stimuli that were intended to be replayed over FCS (stimuli 6, 16, 17, 27, 35 and 37) contained these misallocated signals. All the source locations considered in the listening tests were in the front hemisphere, so the levels of the misallocated signals were small compared to the signals for the front three loudspeakers. Consequently the misallocated signals were not readily apparent to the listeners, who did not report this issue to the test supervisor.

The misallocated signals may result in confusing or conflicting cues arriving at the ears of the listener. It is implicit in the design of the computational model that it will mainly be used for ecologically valid signals,
i.e. signals that are likely to be encountered in the real world, and it can be argued that the misallocated channel stimuli are no longer ecologically valid. However, the exact nature of the signals used as stimuli in the listening test is secondary: the main purpose of the listening test was to validate the computational model from the point where signals are played over loudspeakers to the directional localisation perceived by a listener. Therefore the comparison of the listening test results with the model results for the misallocated channel stimuli is still valid for checking the accuracy of the model.

Improvements to the user interface

There were a few issues with the design of the user interface. The major issue was that some subjects accidentally double-clicked on the "Save Angle" button, resulting in that stimulus being omitted from the test. The user interfaces for the following listening test experiments included validation of the listener responses, which ensured that the listener had entered a response before proceeding to the next stimulus.

Non-matching loudspeakers

Ideally all seven loudspeakers used in the listening test would have had identical responses. One subject claimed to be able to differentiate between the responses of the five active Genelec loudspeakers and the two passive Tannoy loudspeakers. Also, at least one subject reported hearing harmonic distortion on the sine wave stimuli coming from the passive loudspeakers. This distortion was intermittent, and when it did appear, it was only really apparent on the sine tone stimuli. Both of these issues were resolved by using a set of identical active loudspeakers in the following listening test experiments.

4.2.7 Summary of the first listening test

This section described the methodology and results of the first listening test, including details of the novel graphical user interface created to elicit the listeners' responses. The aim of this listening test was to provide a set of localisation results against which the directional localisation model can be validated.

The results showed that the sine tone stimuli were consistently more difficult to localise than the music, speech and pink noise stimuli. Consequently the sine tone stimuli were omitted from the further analysis. These results are consistent with the literature.

Each stimulus was created by either routing the original signal to a single loudspeaker or by simulating the capture of a sound-field using a microphone array. The results show that the stimuli replayed directly through a single loudspeaker were localised close to their intended position (i.e. the position of the loudspeaker). This was not the case for the majority of the stimuli, which created by modelling microphone arrays. A consequence of this is that, in order for the model to be successfully validated, the model has to predict the listening test results rather than predict the intended locations of the stimuli.
Although the first listening test was successful in that it provided a set of results against which the localisation model could be validated, a number of issues became apparent during the course of the listening test and the subsequent analysis of the results (described in Section 4.2.6). These issues, together with a need for listening test data to validate the source width model, provided the motivation for further listening tests.

4.3 The second listening test

This section describes the second listening test that was undertaken to validate the computational model of directional localisation. As in the section on the first listening test, this section will describe the design and methodology of the listening test together with the results and the issues that arose during the course of the experiment. The second listening test was actually designed with two aims. The first is to validate the directional localisation model. The second aim is to start investigating the perception of source width.

4.3.1 The equipment used

Table 4.7 contains details of the eight loudspeakers used in the second listening test experiment, which were arranged as shown in Fig. 4.9. The signals used in the listening test were controlled from the Max/MSP environment on an Apple Mac laptop computer. These were output through a Fireface digital audio interface to eight Bang and Olufsen Beolab 3 active loudspeakers. The connections of the equipment are shown in Fig. 4.10. The same acoustically transparent curtain and scale ranging from -50 to 50 were used as in the first listening test.

<table>
<thead>
<tr>
<th>Name</th>
<th>Loudspeaker</th>
<th>Active</th>
<th>Distance (metres)</th>
<th>Angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beolab 3</td>
<td>✓</td>
<td>2.2</td>
<td>-30</td>
</tr>
<tr>
<td>2</td>
<td>Beolab 3</td>
<td>✓</td>
<td>2.2</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>Beolab 3</td>
<td>✓</td>
<td>2.2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Beolab 3</td>
<td>✓</td>
<td>3</td>
<td>-15</td>
</tr>
<tr>
<td>5</td>
<td>Beolab 3</td>
<td>✓</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>Beolab 3</td>
<td>✓</td>
<td>3</td>
<td>-60</td>
</tr>
<tr>
<td>7</td>
<td>Beolab 3</td>
<td>✓</td>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td>8</td>
<td>Beolab 3</td>
<td>✓</td>
<td>3</td>
<td>-45</td>
</tr>
</tbody>
</table>

Table 4.7: Details of the seven loudspeakers used in the second listening test experiment. The first three loudspeakers were arranged in the same positions as the front three loudspeakers in the standard FCS layout. All the distances are measured from the centre point of the listening area.
Figure 4.9: The equipment layout used in the second and third listening tests. The loudspeakers labelled 1 to 3 are at a radius of 2.2m and are identical to the front three loudspeakers in the standard FCS layout. The remaining loudspeakers, labelled 4 to 8, are at a radius of 3m. The details of all the loudspeakers are given in Table 4.7. The three green circles show the three listening positions, located at (0°, 0m), (-90°, 0.75m) and (135°, 0.75m). Note that although all three listening positions were used in the third listening test, only the centre listening position was used in the second listening test. The wavy line represents the acoustically transparent curtain and the half-octagon inside shows the scale used to assist the listeners with the localisation and source width tasks. The black frame of the figure shows the walls of the listening room.

Figure 4.10: Diagram showing the connections of the equipment used in the second listening test.
4.3.2 The design of the stimuli

Twenty-four different signals were used as the stimuli, details of which can be found in Table 4.8. These twenty-four signals can be divided into three groups of eight according to the signals that were used as the original source. These three groups consist of pink noise, solo musical instruments and speech. As in the first listening experiment, the pink noise was generated in Matlab and all the musical and speech signals were sourced from the anechoic recordings on the Music from Archimedes audio CD produced by Bang and Olufsen [3].

Each of these three groups of signals can be subdivided into three more groups according to how the original signals were processed before being played on the loudspeakers in the listening room. In the first group each of the original source signals is replayed directly over a single loudspeaker with no further processing. This group is included to provide a reference, as in the first listening test. The second group consists of signals panned using the constant power panning law between pairs of loudspeakers. This group of stimuli was included to investigate how well the computational model can localise phantom sources, i.e. signals where the source is perceived to be located at a position other than a loudspeaker location. With the constant power panning law the gains applied to the loudspeakers are calculated by

\[ \theta_m = \frac{\theta_{\text{pan}} - \theta_1}{\theta_2 - \theta_1}, \quad g_1 = \cos \theta_m, \quad g_2 = \sin \theta_m, \]

where \( \theta_1 \) and \( \theta_2 \) are the angles the two loudspeakers from the centre point of the listening area, \( \theta_{\text{pan}} \) is the angle to which the original signal is being panned and \( g_1 \) and \( g_2 \) are gains applied to the signals for the first and second loudspeakers respectively. The angle \( \theta_{\text{pan}} \) is measured from the centre point of the listening area and also satisfies the condition \( \theta_1 \leq \theta_{\text{pan}} \leq \theta_2 \).

The third group consists of signals that have been processed to simulate wide sources. As stated above, the second listening test was designed to fulfill two tasks: the first is to continue the validation of the computational model for directional localisation, and the second task is to begin to gather experimental data for the perception of source width. The third group of stimuli is included in the second listening test to facilitate this second task. The method that was used for producing the wide sources was suggested by Blauert [18] and has also been used in source width experiments by Mason.

This method transforms a single mono signal into a pair of signals. The intention of the method is that when this pair of signals is played simultaneously from a stereo pair of loudspeakers then the sound perceived by a listener is similar to that of the original signal, except that the width of the sound source is perceived to have increased. The resulting pair of signals will be referred to as the Left and Right signals to facilitate the discussion.

Both the Left and Right signals are created by having a copy of the original signal followed, after a short delay, by another copy of the original signal at a reduced gain. The difference between the creation of the Left and Right signals is that in the Left signal the delayed copy of the original signal has the same phase as the first copy, whereas in the Right signal the delayed signal is 180° out of phase. In other words, the Left and Right signals are created by convolving the original signal by the pair of impulse responses shown.
Figure 4.11: The impulse responses convolved with the original signals to create the left and right loudspeaker signals for the widened stimuli.

in Fig. 4.11. These Left and Right signal pairs are then assigned to pairs of loudspeakers. For the second and third stimuli groups, the pairs of loudspeakers to which the signals are applied were chosen to be on the same radius from the centre point of the listening area.

All the stimuli were loudness equalised using an NTI Acoustilyzer AL1 sound level meter to 78db SPL (A-weighted, slow), the level recommended in ITU-R BS 1116 [131]. The levels of the stimuli were then adjusted by ear and using the Moore et al.'s loudness model [110] to have equal perceived loudness.
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Original signals</th>
<th>Processing</th>
<th>Loudspeakers used</th>
<th>Distance (metres)</th>
<th>Angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pink noise</td>
<td>Single</td>
<td>1</td>
<td>2.2</td>
<td>-10 -30 -34</td>
</tr>
<tr>
<td>2</td>
<td>Pink noise</td>
<td>Single</td>
<td>4</td>
<td>3</td>
<td>-1 -15 -21</td>
</tr>
<tr>
<td>3</td>
<td>Pink noise</td>
<td>Single</td>
<td>8</td>
<td>3</td>
<td>-33 -45 -45</td>
</tr>
<tr>
<td>4</td>
<td>Pink noise</td>
<td>Panned</td>
<td>5,7</td>
<td>3</td>
<td>53 -44 30</td>
</tr>
<tr>
<td>5</td>
<td>Pink noise</td>
<td>Panned</td>
<td>1,2</td>
<td>2.2</td>
<td>6 -13 -21</td>
</tr>
<tr>
<td>6</td>
<td>Pink noise</td>
<td>Panned</td>
<td>2,3</td>
<td>2.2</td>
<td>32 15 1</td>
</tr>
<tr>
<td>7</td>
<td>Pink noise</td>
<td>Widened</td>
<td>6,8</td>
<td>3</td>
<td>-42 -53 -51</td>
</tr>
<tr>
<td>8</td>
<td>Pink noise</td>
<td>Widened</td>
<td>2,3</td>
<td>2.2</td>
<td>32 15 1</td>
</tr>
<tr>
<td>9</td>
<td>Female speech</td>
<td>Single</td>
<td>3</td>
<td>2.2</td>
<td>19 0 -11</td>
</tr>
<tr>
<td>10</td>
<td>Male speech</td>
<td>Single</td>
<td>7</td>
<td>3</td>
<td>66 60 46</td>
</tr>
<tr>
<td>11</td>
<td>Female speech</td>
<td>Single</td>
<td>5</td>
<td>3</td>
<td>28 15 4</td>
</tr>
<tr>
<td>12</td>
<td>Male speech</td>
<td>Panned</td>
<td>1,2</td>
<td>2.2</td>
<td>35 18 4</td>
</tr>
<tr>
<td>13</td>
<td>Female speech</td>
<td>Panned</td>
<td>4,5</td>
<td>3</td>
<td>8 -6 -14</td>
</tr>
<tr>
<td>14</td>
<td>Male speech</td>
<td>Panned</td>
<td>5,7</td>
<td>3</td>
<td>55 47 33</td>
</tr>
<tr>
<td>15</td>
<td>Male speech</td>
<td>Widened</td>
<td>4,5</td>
<td>3</td>
<td>14 0 -9</td>
</tr>
<tr>
<td>16</td>
<td>Female speech</td>
<td>Widened</td>
<td>5,7</td>
<td>3</td>
<td>47 38 24</td>
</tr>
<tr>
<td>17</td>
<td>Classical guitar</td>
<td>Single</td>
<td>2</td>
<td>2.2</td>
<td>44 30 13</td>
</tr>
<tr>
<td>18</td>
<td>Trumpet</td>
<td>Single</td>
<td>6</td>
<td>3</td>
<td>-51 -60 -57</td>
</tr>
<tr>
<td>19</td>
<td>Cello</td>
<td>Single</td>
<td>5</td>
<td>3</td>
<td>28 15 4</td>
</tr>
<tr>
<td>20</td>
<td>African percussion</td>
<td>Panned</td>
<td>1,2</td>
<td>2.2</td>
<td>19 0 -11</td>
</tr>
<tr>
<td>21</td>
<td>Classical guitar</td>
<td>Panned</td>
<td>5,7</td>
<td>3</td>
<td>53 44 30</td>
</tr>
<tr>
<td>22</td>
<td>Trumpet</td>
<td>Panned</td>
<td>4,8</td>
<td>3</td>
<td>-26 -39 -40</td>
</tr>
<tr>
<td>23</td>
<td>Cello</td>
<td>Widened</td>
<td>4,8</td>
<td>3</td>
<td>-16 -30 -33</td>
</tr>
<tr>
<td>24</td>
<td>African percussion</td>
<td>Widened</td>
<td>1,2</td>
<td>2.2</td>
<td>19 0 -11</td>
</tr>
</tbody>
</table>

Table 4.8: This table contains details of the twenty four stimuli used in the second listening test. The details of the original signals used in the creation of the stimuli can be found in Table 4.1. The Processing column shows whether the stimulus was panned through a pair of loudspeakers, artificially widened or played through a single loudspeaker. The “Loudspeakers used” column refers back to Table 4.7 and shows which of the loudspeakers were used to play each stimulus. The distance for each stimulus is from the sweet spot to the loudspeakers used to play the stimulus. For those stimuli played through two loudspeakers, both the loudspeakers used are on an arc with the same radius from the centre point of the listening area. The intended location of each stimulus is determined by calculating the intended sweet spot azimuth and assuming that the source lies on the circle with the same radius from the sweet spot as the loudspeakers through which the stimulus was played. The intended sweet spot azimuths were calculated (i) as the location of the loudspeaker for those stimuli played through a single loudspeaker (ii) using equation 4.1 for those stimuli panned between a pair of loudspeakers and (iii) as midway between the angles of the pair of loudspeakers used to play the artificially widened stimuli. The last three columns of the table contain the intended listening angles for each of the three listening positions.
4.3.3 The design of the user interface

The user interface for the second listening test was written using a combination of Max/MSP and JavaScript. This combination allowed easy control of multichannel digital audio while allowing the logic of the user interface to be written in a conventional text-based programming language rather than having to create increasingly complex programs using the Max/MSP graphical language. A screen shot of the user interface for the second listening test is shown in Fig. 4.12. There are six elements in the user interface. The first is a large box in the top right corner which shows a plan of the listening room during the test. Below this there are two buttons, one pink and one green. To the left of the large box are the remaining three elements in the interface: these are boxes which show the numbers entered by the listener during the test.

The box in the top right of the interface is blank at the beginning of the test. The caption on the pink button reads “Start next listening test” and the subject presses this button to begin the test. Once the pink button has been pressed the first recording is played and a plan view of the listener’s head, the scale and the acoustically transparent curtain appear in the display box at the top right of the interface. The caption on the pink button also changes to “Save Angle #1”.

When the subject has listened to the stimulus they are required to provide three responses. The first response is the direction in which the stimulus is perceived to be coming from. The other two responses are the directions of the left and right edges of the sound. These last two responses are used to calculate the perceived width of the sound source.

The directional localisation is entered by the listener in one of two ways. The first is to click inside the display box with the mouse. This causes an arrow to appear in the display box pointing from the centre of the picture of the listener’s head to the position of the mouse. Moving the mouse with the left mouse button still held down allows the test subject to point the arrow in the picture in any direction. Alternatively, the user can release the mouse button and just click in another location in the picture to reposition the arrow.

Figure 4.12: Screen shot of the Max/MSP user interface for the second listening test.
number appears in the yellow box to the left of the display box. This number changes for each new position of the arrow and corresponds to the number on the scale to which the arrow is pointing. The second method of entering a direction into the interface is to type the numbers corresponding to the direction on the scale into the keyboard. The typed number will appear in the yellow box to the left of the display box and an arrow will appear in the picture in the display box pointing at the corresponding point on the scale.

Clicking on any of the three number boxes to the left of the display box changes the box that has been clicked to yellow and also changes which of the responses is currently being entered by the user. That is, clicking on the button next to the caption “Direction of sound:” turns this button yellow and allows the user to enter the perceived direction using either the mouse or keyboard as described above. Clicking on the button next to the caption “Left edge of sound:” turns this button yellow and allows the user to enter the perceived direction of the left edge of the sound in the same way as for the perceived direction of the sound. Similarly, clicking on the button next to the caption “Right edge of sound:” allows the user to enter the perceived direction of the right edge of the sound. Changing between the three different responses can also be done using the up and down arrow keys on the keyboard. When there is already one or more arrow on the picture in the display box and a different one of the three responses (direction, left edge or right edge) is selected, then all existing arrows on the picture become stationary and grey. The new response will generate a new black arrow which can then be positioned as described above, using either the keyboard or the mouse.

Any of the three responses (direction, left edge and right edge) can be altered by reselecting the appropriate button to the left of the display box. This will firstly make the arrow in the picture corresponding to the response change from grey to black, then change the other arrows to grey and finally allow the direction of the black arrow to be edited as described before. The stimulus is played again by clicking on the green button at the bottom of the interface. Once the test subject is satisfied with their responses to the stimulus then clicking on the pink button will save the subject’s responses and move on to the next stimulus. The new stimulus is then played, the numbers are removed from the three buttons to the left of the display box and all three arrows are removed from the picture in the display box. The space bar and the letter “s” on the keyboard can be used as alternatives for the green and pink buttons respectively. This allows the subject to use the interface with either just the keyboard, just the mouse, or a combination of the two. The subject can listen to each stimulus as many times as desired by clicking on the green button and the test only moves on to the next stimulus when the test subject clicks on the pink button. Consequently, the test subject controls the rate of progress through the test. The same method of entering three directions (one for the perceived location and one each for the left and right edges of the sound) is then followed for the new stimulus, with the green and pink buttons used to replay the stimulus and move on to the next stimulus respectively. This process is repeated until all twenty-four stimuli have been played and the listener’s responses have been recorded, at which point the display box will become blank and the caption on the pink button will read “Start next listening test”.

The results are saved as a text file upon completion of the test. The name of the file containing the results is generated from the time at which the test began. Each results file consists of twenty-four rows, one for each stimulus presented in the test. Each row contains the name of the stimulus file, the numbers from the scale corresponding to the location, left edge and right edge responses and also the time at which the listener
submitted their responses for that stimuli. The interface only allows the listener to progress to the next stimulus when four conditions are satisfied. The first condition is that there are three responses from the listener corresponding to the location, left edge and right edge. The second condition is that the arrow for the left edge of the sound must be to the left of the location arrow. Similarly, the third condition that the arrow for the right edge of the sound must be to the right of the direction arrow. The fourth condition is that any numbers entered using the keyboard must be in the range -100 to 100. If any of these conditions are not met then a message box appears in the interface informing the user why the interface will not proceed to the next stimulus. The requirement of satisfying these four conditions ensures that the responses from the listeners are reasonable and also prevents omitted results due to accidentally clicking twice on the pink button (this had been a minor issue with the interface used in the first listening test).

4.3.4 The listening test procedure

The preparation of the subjects before each listening test session was similar to the preparation in the first listening test. Before the test commenced each subject was given a printed sheet describing the intention of the test, the spatial attributes they should be describing and instructions on the use of the interface. This was supplemented by a demonstration of the user interface by the test supervisor. Each listener was also instructed to try to keep their head facing forward and to minimise their head movements during the test. Each session with a subject lasted about thirty minutes, although the time taken varied slightly as the rate of progress through the test was determined by the listener. The intention was to use the three listening positions listed in Table 4.9. However, due to technical difficulties and time constraints on the use of the listening room, only a single listening position was used in the test which was at the sweet spot with the listener facing towards 0°.

<table>
<thead>
<tr>
<th>Position name</th>
<th>Location relative to sweet spot</th>
<th>Listener orientation (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angle (degrees)</td>
<td>Distance (metres)</td>
</tr>
<tr>
<td>Centre</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Left</td>
<td>-90</td>
<td>0.75</td>
</tr>
<tr>
<td>Right</td>
<td>135</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 4.9: The three listener positions intended to be used in the second listening test experiment.

In each listening test session the subject heard the twenty-four different stimuli. The stimuli were presented in a different random order for each listener. As in the first listening test, this was to minimise any bias that may be introduced by having all the subjects hear the stimuli in the same order.
4.3.5 Results and analysis

Ten listeners participated in the second listening test. All ten subjects performed the test once at the central listening position. An analysis of the localisation and source width results for the different listeners is contained in Sections E.2 and E.4 in Appendix E, which concluded that none of the listeners needed to be screened from the results. Fig. 4.13 shows box plots of the results obtained from the test. The labels on the x-axis show the number of the stimulus and also that only the central listening position was used. Dotted lines have been added to divide the results into three groups corresponding to the types of signal processing used in the creation of the stimuli. The stimuli for the left hand group of results consisted of the original signals each played over a single loudspeaker. The group of results in the middle of Fig. 4.13 corresponds to signals panned between pairs of loudspeakers. Finally, the right hand group of results corresponds to the stimuli that had been generated using the source widening method described in Section 4.3.2.

As can be seen from this Fig. 4.13, there are very few outliers and the responses from the listening test subjects are more consistent than the results obtained from the first listening test. This can be attributed to two differences between the first and second listening tests. The first difference is the user interface: the interface used in the second test performed a lot more validation of the data. The second difference is that the processing used to generate the stimuli has changed. The microphone techniques that were modelled in the creation of the stimuli for the first listening test do not produce very localisable signals. In contrast, panning laws such as the constant power panning law have been developed specifically to create localisable signals. It is surprising, however, that the stimuli which were artificially widened were not obviously more difficult for the subjects to localise than either the panned stimuli or the stimuli coming from a single loudspeaker.

Fig. 4.14 shows box plots of the width results obtained from the second listening test. A comparison of these box plots with those for the localisation results show that the subjects' width responses were much less consistent. Fig. 4.15 shows the mean of the localisation azimuth, the mean of the left edge and the mean of the right edge for each stimulus. From this figure it can be seen that the width of each stimulus appears to be independent of the directional localisation of the stimulus. Finally, Fig. 4.16 shows that the localisation results were very close to the intended positions of the stimuli.
Figure 4.13: Centre position box plots for the localisation azimuth for the second listening test experiment. The first line below the box plots contains the number of the stimulus corresponding to each box plot. Details of these stimuli can be found in Table 4.8. The second line shows whether the original signal used in the creation of the stimulus was pink noise (N), music (M) or speech (S). The third line below the box plots shows whether the stimulus was panned through a pair of loudspeakers (P), artificially widened (W) or played back through a single loudspeaker (S). See the caption of Fig. 4.5 for details of how to interpret each box plot.
Chapter 4: Listening Test Experiments

Figure 4.14: Centre position box plots for the source width for the second listening test experiment. The first line below the box plots contains the number of the stimulus corresponding to each box plot. Details of these stimuli can be found in Table 4.8. The second line shows the intended azimuth in degrees from the centre point of the listening area. The third line shows whether the original signal used in the creation of the stimulus was pink noise (N), music (M) or speech (S). The fourth line below the box plots shows whether the stimulus was panned through a pair of loudspeakers (P), artificially widened (W) or played back through a single loudspeaker (S). See the caption of Fig. 4.5 for details of how to interpret each box plot. The vertical dotted lines separate the groups of stimuli with the same intended locations.
Figure 4.15: Centre position plots of means of the azimuth of localisation and the left and right edges of each stimulus for the second listening test experiment. The top edge of each box corresponds to the mean of the right edge responses for that stimulus, the bottom edge of each box corresponds to the mean of the left edge responses and the line across the middle of each box corresponds to the mean of the directional localisation responses. The first line below the plots contains the number of the stimulus corresponding to the plot above. The second line shows whether the original signal used in the creation of the stimulus was pink noise (N), music (M) or speech (S). The third line below the plots shows whether the stimulus was panned through a pair of loudspeakers (P), artificially widened (W) or played back through a single loudspeaker (S). The vertical dotted lines separate the groups of stimuli with the same intended locations, shown by the horizontal dotted black lines. The dashed red line shows the 0° azimuth.
4.3.6 Discussion

The principal issue that arose during the preparation of the second listening test was that the Bang and Olufsen Beolab 3 active loudspeakers would automatically switch themselves off if they did not receive a signal over a period of thirty seconds. The Beolab 3 loudspeakers were recently acquired by the Institute of Sound Recording and this listening test was the first time that the loudspeakers had been used in an experimental situation.

The Beolab 3 loudspeakers are generally used together with other Bang and Olufsen equipment. The connections to the loudspeakers are made with 8-pin DIN plugs. When the loudspeakers are connected to Bang and Olufsen equipment the DIN connectors allow the loudspeakers to be powered through the same cables. In addition to the analogue audio signals and the power supply, Bang and Olufsen equipment connected to the loudspeakers also send additional information, including when the loudspeakers should switch themselves on or off. The analogue audio output of the Fireface digital audio interface is fed via quarter-inch sockets. Bang and Olufsen do allow for the fact that some of their customers want to connect their loudspeakers to audio equipment from other manufacturers. To enable this, Bang and Olufsen produce cables that have a phono plug at one end and an 8-pin DIN plug at the other. One of these cables was reverse engineered by the support staff at the Institute of Sound Recording, who then created cables with the same internal wiring, each with a quarter inch plug on one end and a 8-pin DIN plug on the other.

In general domestic usage the source of the audio signals to the Beolab 3 loudspeakers is likely to be a television, radio, CD player or similar device. When in use these will generally have at least a low level audio signal, which will be enough to ensure that the loudspeakers do not switch themselves off. However, the design of the second listening test is such that only two loudspeakers are being used to play back audio...
signals at any given time. The sequence in which the stimuli are presented is generated randomly for each listener taking part in the test. As the rate of progress through the test is controlled by the subject, with the subjects able to listen to each of the stimuli as many times as desired, it is extremely likely that some of the loudspeakers will switch off due to having thirty seconds with no audio signal.

The Beolab 3 loudspeakers also automatically switch on when an audio signal is detected at its input. However, there is a delay of around half a second between one of the loudspeakers detecting an audio signal at its input and the loudspeaker producing any sound. Consequently, if the listener selects a new stimulus and the loudspeakers required to play the stimulus have automatically switched off, then the first half second of the stimulus is silent and the rest of the stimulus begins abruptly. The situation is worse when a stimulus is being played back over a pair of loudspeakers where one loudspeaker has automatically switched off but the other has played an audio signal more recently and so is still active. In this case, when the stimulus is first played it can be heard first through the loudspeaker still switched on and is then joined half a second later by the other loudspeaker. This results in the listener localising the sound in the direction of the loudspeaker that was not switched off, due to the precedence effect. Therefore, in order to obtain any meaningful results from the second listening test, a method of preventing the Beolab 3 loudspeakers from automatically switching off had to be found.

A temporary solution to the problem of the loudspeakers automatically switching off was implemented by sending a 20Hz sine tone through to the loudspeakers. This tone was audible only when listening to one of the loudspeakers at a distance of 15cm or less. In order to prevent Max/MSP and the Fireface digital audio interface from overloading, this 20Hz tone had to be attenuated whenever one of the listening test stimuli was being played through the loudspeakers.

The level of attenuation applied to the 20Hz tone was determined by a control signal created by first summing the stimuli audio signals and then using full wave rectification and a low pass filter, as shown in Fig. 4.17. The coefficients of the low pass filter and the rate at which the attenuator changed the gain were adjusted to minimise the perception of the changes in amplitude of the 20Hz tone, to ensure there was no overloading of the audio signals and that the 20Hz tone was sufficiently loud when the stimuli were not playing to ensure the loudspeakers did not automatically switch off.

It was possible when using the modified interface to make the signals overload when the gain of the 20Hz tone was not changed quickly enough. However, the conditions required to make the 20Hz tone modification fail were rare and were difficult to reproduce. Indeed, only one of the ten listeners who participated in the second listening test reported hearing any distortion in the replaying of the stimuli, and this only occurred once. None of the participants reported hearing the 20Hz tone and none of the loudspeakers switched themselves off due to no audio signal.
4.3.7 Summary of the second listening test

This section described the methodology and results of the second listening test, including the creation of a novel graphical user interface to elicit the listeners' responses. The aim of this listening test was twofold: to expand the set of listening test results against which the directional localisation model can be validated, and also to provide a set of source width listening test results against which the source width model can be validated. The stimuli for the listening test were generated using three methods. The first method was to route the original signal directly to a single loudspeaker, the second method was to pan the original signal between a pair of loudspeakers and the third method used an algorithm to artificially widen the signal.

The results from the listening test show that these were localised much closer to their intended locations and also that the listeners were much more consistent in their responses compared to the results of the first listening test. This suggests that the stimuli used in the second listening test were easier to localise than the stimuli used in the first listening test. The results also show that the algorithm intended to widen the signals did not create a noticeable increase in the perceived widths elicited from the test subjects.

The second listening test successfully expanded the set of directional localisation results against which the model can be validated. Furthermore, these localisation results exhibit different characteristics in terms of accuracy and ease of localisation compared to the results obtained in the first listening test. However, the listening test was less successful with its second aim of providing source width data with which to validate the model. The algorithm used to widen the signals did not result in larger source widths being elicited from the test subjects compared to the panned stimuli. Consequently, only a narrow range of source width results was obtained, which was insufficient to validate the source width model. Together with the loudspeaker issues described in Section 4.3.6, this provided the motivation for the third listening test.

Figure 4.17: Diagram of the modifications made to the interface used in the second listening test to ensure the loudspeakers did not automatically switch themselves off.
4.4 The third listening test

The third listening test was undertaken for two reasons. The first reason was that only the central listening position was used in the second listening test, due to the time taken to fix the problem of the loudspeakers switching themselves off. The second reason for undertaking a third listening test was that the listening test results for the stimuli that had been widened using the method described in Section 4.3.2 were not significantly different to the listening test results for the stimuli that had not been widened. This suggested that a different method of generating wide stimuli needed to be found.

4.4.1 The equipment used

There was only one difference between the equipment used in the second and third listening test experiments. The major issue with the equipment in the previous experiment was the problem of the loudspeakers switching themselves off. After discussing this issue with Bang and Olufsen, a method of disabling the power saving functionality in the loudspeakers was implemented. This was done by modifying the cables to the loudspeakers to allow batteries to be connected and supply a small current between two of the pins in the DIN sockets on the loudspeakers. This solved the problem and the loudspeakers no longer switched themselves off during the course of the experiment. Other than this modification, the equipment used in the second and third listening test experiments were identical.

4.4.2 The design of the stimuli

Twenty seven different signals were used as the stimuli in the third listening test, details of which are contained in Table 4.10. As in the second listening test, the stimuli can be divided into three groups according to the signals that were used as the original source; pink noise, solo musical instruments and speech. As in the first two listening tests, the pink noise was generated in Matlab and all the musical and speech samples were sourced from the anechoic recordings on the Music from Archimedes audio CD [3]. Following the same pattern as the second listening test, each of these three groups can be subdivided into three more groups according to how the signals were processed before being played on the loudspeakers in the listening room. These three processing groups are the same as in the second listening test: sources played over a single loudspeaker with no further processing, sources panned between pairs of loudspeakers and sources that have been artificially widened. The stimuli in the first two groups (using a single loudspeaker or panned between a pair of loudspeakers) are identical to the corresponding stimuli in the second listening test. However, a different method to that used in the second listening test was used to artificially widen the stimuli in the third group.

The method used to widen the sources in the third listening test was adapted from a method developed by Dr Tim Brookes at the Institute of Sound Recording at the University of Surrey [30]. The original source signal is passed through an all-pass filter. The resulting signal has a magnitude spectrum identical to the
Table 4.10: This table contains details of the twenty seven stimuli used in the third listening test experiment. The details of the original signals used in the creation of the stimuli can be found in Table 4.1. The distances and angles contained in the table were calculated as described in the caption to Table 4.8. Note that stimuli numbers 1 to 6, 9 to 14 and 17 to 22 are identical to the corresponding stimuli used in the second listening test experiment. The remaining stimuli are those that were artificially widened. These were newly created for the third listening test experiment.
original source signal, but a different phase spectrum, i.e. the filtered signal is a decorrelated version of the original signal. This pair of signals are then panned to two separate locations using the constant power panning law and a pair of loudspeakers.

All the stimuli were loudness equalised using an NTI Acoustilyzer AL1 sound level meter to 78db SPL (A-weighted, slow), the level recommended in ITU-R BS 1116 [131]. The levels of the stimuli were then adjusted by ear and using the Moore et al.'s loudness model [110] to have equal perceived loudness.

4.4.3 The design of the user interface

The user interface from the previous listening test experiment was reused in the third listening test experiment. The only modification made to the user interface was that each stimulus was continuously looped until the user moved on to the next stimulus. The functionality of the green button was altered so that pressing it once paused the playback of the current stimulus and pressing the button again resumed the playback. As in the previous experiment, the space bar had the same functionality as the green button.

4.4.4 The listening test procedure

The listening test procedure used in the third experiment was identical to that of the previous experiment, except that two additional listening positions were also used. The three listening positions are described in Table 4.9.

4.4.5 Results and analysis

Twelve listeners participated in the third listening test experiment. Nine of these subjects participated in three listening tests, once at each of the three listening positions. Two of the twelve subjects performed the test at only two of the three listening positions and the remaining subject performed the test only once. Table 4.11 shows the number of listening tests performed at each of the three listening positions.

<table>
<thead>
<tr>
<th>Listening position</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>Centre</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 4.11: The number of tests at each listening position for the third listening test experiment.

An analysis of the localisation and source width results for the different listeners is contained in Sections E.3 and E.4 in Appendix E, which concluded that none of the listeners needed to be screened from the results. Figs. 4.18 to 4.20 show box plots of the localisation results obtained from the third listening test experiment. Table 4.12 summarises the differences between the intended and the listening test azimuths for
the different listening positions and types of processing. From the table and figures it is apparent that the listeners' responses more closely matched the intended angles of the stimuli in the centre listening position compared to either the left or right listening positions. This was expected for those stimuli that were panned between pairs of loudspeakers, as the panning law was optimised for the centre listening position. Similarly, the algorithm used to artificially widen the sources uses the same panning law to position both the original signal and the decorrelated signal, and so the widened stimuli were expected to exhibit similar behaviour to the panned stimuli. It can also be seen from Table 4.12 and Figs. 4.18 to 4.20 that listener responses for the stimuli played through a single loudspeaker were the closest to the intended angles. As the intended angle coincides with the position of the single loudspeaker through which each of these stimuli was played, any differences between the listener responses and the intended angles are due to either the limitations of the ability of the listeners to localise sound, or alternatively noise or errors due to the experimental procedure.

The differences between the listener responses for the panned stimuli and the intended angles were larger than those for the single loudspeaker stimuli. This was expected, as panning laws are not perfect and are known introduce some spatial distortion, which is well documented. Note that at the centre listening position, which is the position for which the panning law was optimised, there are comparable differences between the listener responses and the intended angles. The algorithm used to artificially widen some of the stimuli was designed specifically to widen signals, and creating accurately localisable sounds was of secondary importance. As such, it is not surprising that the differences between the listener responses and the intended angles were larger for the widened stimuli compared to either the panned stimuli or the single loudspeaker stimuli.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
</tr>
<tr>
<td>Single</td>
<td>2.62</td>
</tr>
<tr>
<td>Panned</td>
<td>6.97</td>
</tr>
<tr>
<td>Widened</td>
<td>28.58</td>
</tr>
</tbody>
</table>

Table 4.12: Localisation errors. The mean errors (in degrees) between the intended localisation azimuths and the localisation azimuths from the listeners. For each combination of stimulus and listening position the error was calculated as the magnitude of the difference between the median azimuth (calculated from the responses from all the listeners) and the intended azimuth. These errors were then grouped according to the processing type and listening position and the mean error was calculated for each of these groups.

Figs. 4.21 to 4.23 show box plots of the source width results. Table 4.13 summarises the source width results from the listening test for the different listening positions and types of processing. From the table and figures it is clear that in the listening test the stimuli that had been artificially widened were perceived as being wider than the panned stimuli and the single loudspeaker stimuli. The panned stimuli were also perceived to be wider than the single loudspeaker stimuli. This was expected, as pair-wise panning has been noted to have the effect of blurring the position of phantom sound source, which is closely related to an increase in width. More surprising was the fact that the stimuli for all three types of processing were perceived as being wider in the centre listening position than at either the left or right positions.

CHAPTER 4. LISTENING TEST EXPERIMENTS | 88
Table 4.13: Mean source widths. Grand means of the width results from the third listening test. The grand means were calculated by first calculating the mean width for each combination of listener position and stimulus, and then calculating the mean of these values grouped according to the listening position and the process used to generate the stimulus. The units of the mean widths are degrees.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
</tr>
<tr>
<td>Single</td>
<td>8.17</td>
</tr>
<tr>
<td>Panned</td>
<td>10.29</td>
</tr>
<tr>
<td>Widened</td>
<td>20.15</td>
</tr>
</tbody>
</table>

Figs. 4.24 to 4.26 show the mean of the location azimuth, the mean of the left edge and the mean of the right edge for each stimulus. These plots illustrate the relationship between the source width results and the directional localisation results. Of particular note are the plots for stimuli 25 and 27 in Fig. 4.25. These show mean location azimuths that are clearly not centered between the mean left and right edges: in both cases the mean location azimuth is closer to the mean left edge.

Fig. 4.27 shows a comparison of the standard deviations for the localisation azimuth and source width results for the centre listening position. It can be seen from this that the standard deviations of the source widths were generally larger than the standard deviations of the localisation azimuths. This infers that there was more agreement between the listeners for the localisation results than there was for the source width results, which suggests that determining source width is a harder task for listeners than localising sources. The standard deviations at the other two listener positions are similar to those shown in Fig. 4.27.

Tables 4.14 and 4.15 show the standard deviations for the source width results summarised by processing type and the type of original signal respectively. Table 4.14 shows that the source width results for the widened stimuli had a larger spread of values than the other two processing types. Some of the listeners who participated in the third listening test reported to the test supervisor that the widened stimuli sounded artificial and “phsey” rather than wide. This may explain why the source width results for the widened stimuli were less consistent than the results for the other two processing types.
Table 4.14: Standard deviations of source widths organised by processing. The averages of the standard deviations of the source width results from the third listening test. The units of the standard deviations are degrees. For each combination of stimulus and listening position the standard deviation was calculated from the source width results for all the listeners. These standard deviations were then grouped according to the processing type and the listening position and the average standard deviation was calculated for each of these groups.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Left</th>
<th>Centre</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>3.77</td>
<td>7.86</td>
<td>3.88</td>
</tr>
<tr>
<td>Panned</td>
<td>5.20</td>
<td>10.86</td>
<td>5.33</td>
</tr>
<tr>
<td>Widened</td>
<td>14.29</td>
<td>14.45</td>
<td>10.81</td>
</tr>
</tbody>
</table>

Table 4.15: Standard deviations of source widths organised by original signal. The averages of the standard deviations of the source width results from the third listening test. The units of the standard deviations are degrees. For each combination of stimulus and listening position the standard deviation was calculated from the source width results for all the listeners. These standard deviations were then grouped according to the original signal and the listening position and the average standard deviation was calculated for each of these groups.

<table>
<thead>
<tr>
<th>Original signal</th>
<th>Left</th>
<th>Centre</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music</td>
<td>10.17</td>
<td>9.79</td>
<td>7.65</td>
</tr>
<tr>
<td>Noise</td>
<td>7.30</td>
<td>14.87</td>
<td>7.14</td>
</tr>
<tr>
<td>Speech</td>
<td>5.79</td>
<td>7.97</td>
<td>5.23</td>
</tr>
</tbody>
</table>

The left hand plot in Fig. 4.28 shows the localisation results from the listening test plotted against the intended angles (see Table 4.10). This plot includes the $R^2$ value and the root mean error of prediction (RMSEP) calculated from using the intended angles as predictors of the listening test results. Note that there are a number of stimuli where the angles from the listening test do not match the intended angles. The right hand plot in Fig. 4.28 shows the same results, but with the artificially widened stimuli omitted. From this it can be seen that most of the outliers in the left hand plot belong to the set of artificially widened stimuli, and the removal of these stimuli brings the root mean square error of prediction down from 12° to 6°. However, there are still a few outliers around 0° in the right hand plot in Fig. 4.28, all of which were identified as belonging to the subset of stimuli that were created using the constant power panning law and also corresponding to the off-centre listening positions. This may be partly due to the way the intended angles were calculated for the panned stimuli at off-centre listening positions.

For the panned stimuli, the intended angle at the sweet-spot (where the listener was equidistant from both the active loudspeakers for all the panned stimuli in the third listening test) was calculated by rearranging Equation 4.1 to find $\theta_{\text{pan}}$. This phantom source was then assumed to have the same distance from the
sweet spot as the two loudspeakers, $L_Sd$, giving the phantom source a location of $(\theta_{p\text{an}}, L_Sd)$. The intended angles for the two off-centre listener positions were then calculated from the listener position to $(\theta_{p\text{an}}, L_Sd)$. This assumes that the phantom source remains located at $(\theta_{p\text{an}}, L_Sd)$ even when the position of the listener changes.

An alternative method of determining the intended angle of the panned stimuli at off-centre listener positions was also implemented. This method uses the fact that Equation 4.1 is valid when the two active loudspeakers are equidistant from the listener position. Consider the case when the listener position is off-centre and so one loudspeaker is closer than the other to the listener, as shown in the left hand plot of Fig. 4.29. The point $P$ in this plot is the point on the direct path between the listener and loudspeaker 1 such that the distance between the listener and $P$ is equal to the distance between the listener and loudspeaker 2. Consider the case where the loudspeakers are omnidirectional and the listening environment is anechoic. The binaural signals at the listener position are created by the superposition of the binaural signals due to loudspeaker 1 alone and the binaural signals due to loudspeaker 2 alone. If the near-field effects of the HRTFs are negligible, then the same binaural signals can be generated by the superposition of the binaural signals due to loudspeaker 2 alone and the binaural signals due to a source situated at $P$ playing a delayed and attenuated copy of the signal fed to loudspeaker 1. The delay, $t_P$, is calculated as the time taken for the sound to travel from loudspeaker 1 to point P,

$$ t_P = \frac{L_{Sd,1} - L_{Sd,2}}{c}, $$

where $L_{Sd,1}$ is the distance from the listener to loudspeaker 1, $L_{Sd,2}$ is the distance from the listener to loudspeaker 2 and $c$ is the speed of sound. The reduced gain, $g_P$, of the source at $P$ is equal to the attenuation due to the propagation through free space from loudspeaker 1 to point $P$,

$$ g_P = g_1 \left( \frac{L_{Sd,2}}{L_{Sd,1}} \right)^2, $$

where $g_1$ is the original gain of loudspeaker 1 and $L_{Sd,1}$ and $L_{Sd,2}$ are the distances to loudspeakers 1 and 2 respectively. As point $P$ and loudspeaker 2 are equidistant from the listener, Equation 4.1 can now be used with the gain from the source at $P$, $g_P$, the gain from loudspeaker 2, $g_2$ and the angles $\theta_1$ and $\theta_2$ (see the left hand plot in Fig. 4.29) to calculate the angle to which the source has been panned. This method will be referred to as the equidistant method of calculating the intended angles.

The right hand plot in Fig. 4.29 shows the intended angles calculated using the equidistant method plotted against the localisation results from the third listening test experiment. As in the right hand plot in Fig. 4.28, the artificially widened stimuli have been omitted. A comparison of the right hand plots in Figs. 4.28 and 4.29 shows that the equidistant method of calculating the intended angles for the panned stimuli at off-centre listener positions has a closer match to the listening test results. There are still some data points around 0° in the right hand plot of Fig. 4.29 where the intended angles are different to the listening test results. These are probably due to the fact that the equidistant method ignores the effect of the delay, $t_P$, of the signal at the point $P$, which will also influence the perceived location of the phantom source.
CHAPTER 4. LISTENING TEST EXPERIMENTS | 92

Figure 4.18: Centre position box plots for the localisation azimuths for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret both the box plots themselves and the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.19: Left position box plots for the localisation azimuth for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret both the box plots themselves and the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.20: Right position box plots for the localisation azimuth for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret both the box plots themselves and the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.21: Centre position box plots for the source width for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret the box plots and the caption of Fig. 4.14 for details of how to interpret the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.22: Left position box plots for the source width for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret the box plots and the caption of Fig. 4.14 for details of how to interpret the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.23: Right position box plots for the source width for the third listening test experiment. See the caption of Fig. 4.5 for details of how to interpret the box plots and the caption of Fig. 4.14 for details of how to interpret the table below the box plots. The details of the stimuli can be found in Table 4.10.
Figure 4.24: Centre position plots of means of the azimuth of localisation and the left and right edges of each stimulus for the third listening test experiment. See the caption of Fig. 4.15 for details of how to interpret both the plots themselves and the table below the plots. The details of the stimuli can be found in Table 4.10.
Figure 4.26: Left position plots of means of the azimuth of localisation and the left and right edges of each stimulus for the third listening test experiment. See the caption of Fig. 4.15 for details of how to interpret both the plots themselves and the table below the plots. The details of the stimulus can be found in Table 4.10.

CHAPTER 4. LISTENING TEST EXPERIMENTS
Figure 4.26: Right position plots of means of the azimuth of localisation and the left and right edges of each stimulus for the third listening test experiment. See the caption of Fig. 4.15 for details of how to interpret both the plots themselves and the table below the plots. The details of the stimuli can be found in Table 4.10.
Figure 4.27: *Comparison of the standard deviations for the localisation azimuth (blue) and source width (red) results for the centre position in the third listening test.*
Figure 4.28: These two plots show the intended angles plotted against the localisation results for all three listening positions from the third listening test experiment. The left plot includes all the stimuli from the third experiment, while the right hand plot omits the artificially widened stimuli.

Figure 4.29: The left hand plot illustrates how the values of the intended angles were modified for the panned stimuli at off-centre listening positions. The right hand plot shows the modified intended angles plotted against the localisation results for all three listening positions from the third listening test experiment. The artificially widened stimuli have been omitted from the right hand plot.
4.4.6 Comparison of the results from the second and third listening tests

This section contains a discussion of comparisons that were made between the results from the second and third listening test experiments. Two different comparisons were undertaken. The first comparison was to test whether there was a significant difference between the results for the combinations of test subject, listening position and stimulus that were repeated between the two experiments. This was to check whether the listening test experiments produced repeatable results. The second comparison was to test whether there was a significant increase in the range of elicited source widths for the source widening algorithm for the third listening test experiment compared to the second listening test experiment.

Results from the same combination of stimulus, listening position and test subject

As discussed above in Section 4.4.2, the stimuli used in the third listening test can be divided into three groups according to the type of processing applied to the original source signals: stimuli played through a single loudspeaker, stimuli panned through a pair of loudspeakers and stimuli that have been artificially widened. The stimuli in the first two of these groups are identical for both the second and third listening tests. Additionally, nine of the ten listeners who participated in the second listening test also participated in the third listening test where they gave results for the centre listening position. Therefore, the repeatability of the listening test procedure can be assessed by inspecting the results from the second and third listening test experiments corresponding to these nine test subjects, the centre listening position and both the single loudspeaker and the pair-wise panned stimuli groups.

Section E.4.1 in Appendix E includes details of using the Shapiro-Wilk test to determine whether the localisation results for the repeated combinations were normally distributed. This showed that a third of the repeated stimuli had results which were not normally distributed. A consequence of this is that a non-parametric statistical test was required to determine the repeatability of the listening test procedure. Table 4.16 contains the results of performing a Wilcoxon signed-rank test [48, 146] on the localisation results from the second and third listening test experiments for the repeated combinations of stimulus, listener and listening position. The test was performed at the 5% significance level. This shows that the null hypothesis was rejected and so there was a significant difference between the localisation results with the same combinations from the two experiments.

Possible explanations for these differences include (i) the changes made to the experimental procedure between the second and third listening test experiments, (ii) poor experimental technique, or (iii) inconsistent test subjects. Excluding the changes to the artificially widened stimuli, which have been excluded from the Wilcoxon signed-rank test, the experimental procedure for the third listening test was mostly the same as that of the second listening test. Two explicit changes were made to the experimental procedure. The first was the modification to the loudspeaker cables that prevented the loudspeakers automatically switching themselves off (see Section 4.4.1). This should have resulted in an improvement to the quality of the stimuli presented over the loudspeakers, as the temporary solution implemented in the second listening test could be discarded. The temporary solution consisted of presenting a 20Hz tone to the loudspeakers when no stimuli
Table 4.16: Result of Wilcoxon signed-rank test comparing the localisation results from the second and third listening test experiments that had matching combinations of stimulus, listening position and listener. The null hypothesis of the test was that there was no difference between the localisation results for the second and third listening tests. The test was performed at the 5% significance level.

<table>
<thead>
<tr>
<th>Test statistic, T</th>
<th>Null hypothesis accepted at 5% level</th>
<th>Effect size, r</th>
</tr>
</thead>
<tbody>
<tr>
<td>2229</td>
<td>x</td>
<td>-0.27</td>
</tr>
</tbody>
</table>

were played (see Section 4.3.6) and the implementation of this solution occasionally caused audible distortion to the stimuli. Therefore, the removal of this temporary solution should have improved the quality of the stimuli, which could have caused a change in the results. The second change was that the user interface was altered so that the stimuli were looped rather than having to be triggered each time by the listener (see Section 4.4.3). The intention of this change was to make it easier for the subjects in the experiment, which was verified by informal feedback gained from the listeners in the third listening test. The intra-subject consistency was assessed in Section E.4.1 in Appendix E, including a comparison of the localisation results for the stimuli played through a single loudspeaker, i.e. the stimuli with a known source location. This showed that the localisation results were more accurate in the third experiment compared to the second experiment. This supports the proposal that the changes made to the experimental procedure have improved the results of the experiment and so account for the differences between the localisation results exposed by the Wilcoxon signed-rank test whose results are shown in Table 4.16.

Section E.4.1 in Appendix E also includes details of using the Shapiro-Wilk test to determine whether the source width results for the repeated combinations were normally distributed. This showed that five of the eighteen repeated stimuli had results which were not normally distributed. Consequently, a non-parametric statistical test was required to determine the repeatability of the listening test procedure with regard to the source width results. Table 4.17 contains the results of performing a Wilcoxon signed-rank test on the source width results from the second and third listening test experiments for the repeated combinations of stimulus, listener and listening position. The test was performed at the 5% significance level. This shows that the null hypothesis was accepted and so there was not a significant difference between the localisation results with the same combinations from the two experiments. Note that this did not include the artificially widened stimuli, only the stimuli played through a single loudspeaker and the pair-wise panned stimuli.
### Table 4.17: Result of Wilcoxon signed-rank test comparing the source width results from the second and third listening test experiments that had matching combinations of stimulus, listening position and listener. The null hypothesis of the test was that there was no difference between the source width results for the second and third listening tests.

<table>
<thead>
<tr>
<th>Test statistic, $T$</th>
<th>Null hypothesis accepted at 5% level</th>
<th>Effect size, $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3162.5</td>
<td>✓</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

The range of the source width values

The section investigates whether the changes made to the artificially widened stimuli between the second and third listening test experiments resulted in any changes to the range of source width results. For each combination of listening position and listener in the second listening test there is a set of source width results, with each member of this set corresponding to a different stimulus. A value for the size of the range of source widths for each of these sets was calculated as the difference between the largest and smallest source widths obtained from that combination of listening position and listener. Similarly, each combination of listening position and listener in the third listening test corresponds to a value for the size of the range of source widths obtained in the third listening test. Nine of the listeners who participated in the second listening test experiment also participated in the third listening test experiment, where they gave results for the central listening position. Therefore, the range of source widths for each of these nine listeners at the central listening position in the second listening test was compared with the range of source widths for the same combinations of listener and listening position in the third listening test. As the source width results have been shown to be not normally distributed (see Section E.4.2 in Appendix E), a non-parametric test was used for this comparison. Table 4.18 shows the results of performing a Wilcoxon signed-rank test on the ranges of source width between the second and third experiments for the nine listeners at the central listening position. From this it can be seen that the null hypothesis was accepted, so any differences in the ranges of source width values between the two experiments is not statistically significant at the 5% level.

### Table 4.18: Result of Wilcoxon signed-rank test comparing the range of the source width results from the second and third listening test experiments with the same combination of listening position and listener. The null hypothesis of the test was that there was no difference between the range of the source width results for the second and third listening tests.

<table>
<thead>
<tr>
<th>Test statistic, $T$</th>
<th>Null hypothesis accepted at 5% level</th>
<th>Effect size, $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>✓</td>
<td>-0.06</td>
</tr>
</tbody>
</table>
However, Table 4.19 shows the ranges in source widths from the two listening tests, and this does show that the range of elicited widths was greater for the third listening test. This was the intention of the change to the widening algorithm. As Table 4.19 shows a clear increase in the range of width responses, the acceptance of the null hypothesis in the Wilcoxon signed-rank must be due to the large variations in the width responses obtained for each width stimuli. This is supported by the width box plots in Figs. 4.14, 4.21, 4.22 and 4.23 and the comparison of the standard deviations for the localisation and width results at the centre position in the third listening test shown in Fig. 4.27.

<table>
<thead>
<tr>
<th>Listening test</th>
<th>Minimum source width</th>
<th>Maximum source width</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Angle (degrees)</td>
<td>Stimulus</td>
<td>Angle (degrees)</td>
</tr>
<tr>
<td>Second</td>
<td>5.76</td>
<td>12</td>
<td>18.18</td>
</tr>
<tr>
<td>Third</td>
<td>7.20</td>
<td>11</td>
<td>34.82</td>
</tr>
</tbody>
</table>

Table 4.19: The minimum, maximum and range of source width results from the second and third listening tests.

4.4.7 Summary of the third listening test

This section described the methodology and results of the third listening test. The aims of this listening test were the same as those of the second listening test, namely providing directional localisation and source width results against which the model can be validated.

As in the previous listening test, three methods were used in the creation of the stimuli. Two of these, the direct routing to a single loudspeaker and the pair-wise panning, were identical to the methods used in the second listening test: indeed, the stimuli generated using these two methods were identical in both listening tests. The third method, however, involved the use of different algorithm to artificially widen the signals.

Results were obtained for three different listening positions. The localisation results were similar to those from the second listening test, although a much larger set of results was obtained due to the use of more than a single listening position. The localisation results for the widened sources included two notable features: they were not localised at their intended positions when the listeners were off-centre, and the wider range of angles elicited from the test subjects suggests that the widened stimuli were more difficult to localise. Thus, the third listening test was successful in that it continued to expand the set of directional localisation results against which the model can be validated.

The other aim of the listening test was to provide a set of source width results for the validation of the model. One of the reason for performing the third listening test was that only a relatively small range of source widths had been obtained in the second listening test. An initial inspection of the third listening test results suggested there was an increase in the range of source widths compared to the previous listening test. However, this was shown not to be statistically significant at the 5% level.
Statistical tests were also performed on the results of those combinations of stimulus, listener and listening position that were present in both the second and third listening test experiments. From these, the localisation results were shown to be significantly different at the 5% level between the two experiments. Further investigation showed that these differences were confined to the results of only three listeners, who all showed improved accuracy in their localisation results. This can be attributed to the changes to the experimental procedure made between the second and third listening test experiments. The source width results for the repeated combinations of stimulus, listener and listening position between the two experiments were shown not to be significant at the 5% level.

4.5 Summary

This chapter has described the methodology and results of three listening test experiments. The aim of all three of these experiments was to provide a set of localisation results against which the directional localisation model could be validated. The second and third experiments were also designed to provide a set of listening test results which could be used to investigate source width.

Each stimulus in the first listening test experiment was created by either routing the original signal to a single loudspeaker or by simulating the capture of a sound-field using a microphone array. The results of the first experiment showed that the stimuli played directly through a single loudspeaker were localised close to the position of the loudspeaker, while the stimuli created by simulating microphone techniques were often not localised close to the position of the original modelled source. Consequently, to be successfully validated, the model has to predict the listening test results rather than the intended location of each stimulus. In the subsequent listening test experiments the stimuli were created using constant power panning instead of simulating microphone arrays in order to control more accurately the apparent source location in the reproduced sound-field. The results from the first listening test experiment showed that sine tone stimuli were consistently much more difficult to localise than the music, speech and pink noise stimuli. This is consistent with the literature, e.g. Hartmann [70]. Consequently the sine tone stimuli were omitted from the subsequent listening test experiments.

The design of the second and third listening test experiments differed from that of the first experiment in that the test subjects were also required to evaluate the source width of each stimulus. It was originally intended that there would be only two listening test experiments, but a third experiment was required for two reasons. Firstly, technical issues with the loudspeakers meant that there was only time to perform the listening test at a single listening position. The second reason was that the stimuli which had been artificially widened were not judged by the test subjects to be much wider than the other stimuli, giving a small range of source width values in the results of the listening test. Different methods of widening sources were tried in order to remedy this and a method based on a random phase all-pass filter was used in the creation of the widened stimuli for the third listening test experiment. An initial analysis of the results of the second and third listening test experiments showed that the localisation results were much more consistent than the source width results.
The stimuli for both the second and third listening test experiments were divided into three groups. The first group consisted of signals played through a single loudspeaker. The second group consisted of signals panned between a pair of loudspeakers. The third group consisted of signals that had been artificially widened, with a different method used to widen the signals for each of the two experiments, as discussed above. The stimuli in the first two groups (i.e. the single loudspeaker and panned stimuli) were identical in the second and third listening test experiments. This overlap between the stimuli of the two experiments meant that a number of statistical tests could be performed to compare the results of the two experiments.

The difference between the localisation results of the second and third listening tests experiments were found to be significant at the 5% level. However, this was found to be due to the results of three of the nine subjects who participated in both tests. In all three cases the results improved with respect to stimuli when there was a known true angle (i.e. those stimuli played through a single loudspeaker). This may be attributed to the improvements in the design of the listening test experiment. The fact that the localisations obtained from the experiments had a relatively small spread also contributed to the rejection of the null hypothesis. The differences between the source width results of the second and third listening test experiments were found not to be significant at the 5% level. Finally, an increase in the range of width results could be seen when comparing the results of the second experiment to the results of the third, although the differences were found not to be significant at the 5% level.
Chapter 5

Directional localisation

This chapter describes how a binaural model of directional localisation can be combined with the framework described in Chapter 3 to investigate the localisation properties of reproduced audio. The localisation of sound sources is one of the most important spatial attributes of reproduced audio [59, 140]. As such, a number of computational models have been developed that calculate directional localisation using binaural signals [76, 107, 124, 129, 158]. These all work on similar principles. First, the binaural signals are separated into frequency bands. The resulting signals are then half-wave rectified and low-pass filtered. IIDs and ITDs are then calculated for each frequency band, using cross-correlation to calculate the ITDs. Horizontal azimuths are then calculated from the IIDs and ITDs and, finally, these are combined to give the directional localisation of the sound source. The decision was made to use the directional localisation model developed by Supper [158], partly due to the fact that there was access to the model's source code. This allowed the model to be further developed in order to be incorporated into the model framework described in Chapter 3, and also allowed the possibility of modifying the model in order to improve its performance.

The incorporation of the Supper model into the model framework is described in the first section of the chapter, followed by a description of the validation of the model using the localisation results from the three listening test experiments. The remainder of the chapter describes using the model to investigate different aspects of directional localisation with reproduced audio. First is an investigation into the effects of reflective signals on directional localisation. Second is a comparison of the ability of different reproduction systems to reproduce source locations for a listener located at the sweet spot. This is followed by an investigation into how the model can be used to predict localisation across the listening area for different reproduction systems.
5.1 How the Supper model was incorporated into the model framework

This section contains a description of the investigation that was conducted into the details of Supper's model for calculating the directional localisation from a binaural pair of signals [158]. This investigation was initiated because of the difficulties that were experienced when integrating the output of the Supper model into the model framework described in Chapter 3.

Initially a data driven approach was used for the integration of the Supper model into the overall model framework. In this approach the Supper model was treated as a "Black Box", where the internal workings of the model were considered of secondary importance to the data output by the model. The reason for adopting this approach was that the main focus of the research was to be modelling perceived spatial attributes at points across the listening area. Converting a binaural stream into a direction of localisation is just one aspect of the project. This problem has been researched and is well documented in the literature. Consequently, the decision was made to use an existing algorithm for this part of the modelling in the project. However, it was found that the output of the Supper model needed to be subjected to additional processing in order to give good results when compared to results obtained from the subjective listening tests. This additional processing had been developed in a heuristic manner, i.e. certain parts of the results from the Supper model were ignored or suppressed as they did not fit either the expected results or the experimental results, but no more justification was given for why this was the case. It follows from this that the heuristic processing was not very defensible, especially given that the heuristic processing was becoming increasingly complicated in order to produce satisfactory results as more subjective listening test results became available for validation. The other factor that motivated the investigation into the Supper model became apparent when analysing the results of the third set of listening tests: that the software implementation of the heuristic method used to integrate the Supper model crashed when processing the signals used as stimuli in the listening tests. This was diagnosed to be a problem with the heuristic method itself rather than simply a mistake in the software implementation.

5.1.1 Treating the Supper model as a "Black Box": the heuristic interpretation

The output of the Supper model is a time series of histograms. The elapsed time between consecutive histograms in this series is 408µs. The values in each histogram are in the range [0 1]. These values are fuzzy logic "truth values" and represent the degree of membership of a set. In this case, this is the set of binaural stimuli that can produce the Interaural Intensity Difference (IID) and Interaural Time Difference (ITD) cues corresponding to a sound source localised at that angle.

For a given stimulus it is desirable to obtain a single value for the directional localisation. Stationary stimuli will be considered first. This allows the entire time series to be considered when attempting to extract a single value for the directional localisation. Indeed, all the stimuli used in the three listening tests have been stationary. The most obvious method of considering the entire time series is to simply average the
histograms over time (see Fig. 5.1). This will result in a single histogram for each stimulus which shows the average truth value for each angle. The histograms are weighted by the calculated loudness. This means that those histograms corresponding to the quiet sections in the stimulus will contribute much less than the loud sections.

The first method that was used to obtain a single angle was to calculate the median of the histogram. One problem that was discovered with this method was that some of the averaged-over-time (AOT) histograms had more than a single peak. The median value of the histogram would often fall between two of these peaks. In practice this meant that neither the IID nor the ITD cues in the binaural signals were consistent with the median angle. This suggests that reducing the AOT histogram by taking the median value is not a good method for obtaining realistic directional localisations. This was confirmed by creating binaural signals corresponding to sound sources from a known location. In the cases where there were multiple maxima in the AOT histogram, one of the maxima would coincide with the actual location of the sound source. Additionally, this would often be the largest peak in the histogram. This suggested the method of obtaining a single value from the histogram by finding the angle that corresponds to the maximum value (i.e. the mode). This will be referred to as the “peak picking” method.

Fig. 5.2 shows some of the results output by the Supper model for some of the stimuli used in the third listening test. The results in the figure correspond to the 1st and 6th stimuli in the third listening test. The 1st stimulus was a one second burst of pink noise played through a single loudspeaker positioned at an angle of $-30^\circ$ and a distance 2.2m from the centre of the listening area. The 6th stimulus was a one second burst of pink noise panned using the constant power panning law to an angle of $15^\circ$ from the centre of the listening area. The two loudspeakers used to play the panned signal were at angles 0° and 30° and a distance of 2.2m from the centre of the listening area. For both results the listener was positioned at a distance of 0.75m and an angle of $135^\circ$ from the centre of the listening area and facing straight ahead at an angle of 0°. All the results in Fig. 5.2 were generated using the Supper model in its original state.

The maximum value in the histogram in the left hand graph in Fig. 5.2 is close to $-90^\circ$. However, the peak at around $-40^\circ$ seems to correspond much better with both the actual position of the sound source and the angle elicited from the subjects in the third listening test (shown by the vertical line in the graph). This problem of having spurious large peaks around $\pm90^\circ$ in the AOT histograms occurs often, giving rise to large errors. It was found that it was possible to obtain much better results for the stimuli under consideration by considering only the portion of the AOT histogram with angles in the range $\pm80^\circ$.

The second major problem that was encountered when integrating the Supper model into the model framework described in Chapter 3 is illustrated by the right hand graph in Fig. 5.2. In this AOT histogram can be seen a large narrow spike at 0°. The sound source was constant power panned midway between loudspeakers 2.2m away from the centre of the listening area at angles of 0° and 30° to the right. The vertical line in the graph shows the median angle elicited from the subjects in the listening test. This angle is close to the peak in the histogram around 20°. However, the maximum value in the histogram is at the spike at 0°, and consequently this is the angle selected by the peak picking method. Having a spike at 0° appears frequently in the AOT histograms, often when smaller peaks in the histogram are in much closer agreement with the
Figure 5.1: These two graphs show the intermediate results for stimulus 11 from the third listening test. This stimulus consisted of a sample of female speech of the words “One, two”, which lasted 1.5 seconds. The sample was played through a single loudspeaker positioned 3m from the centre of the listening area and at an angle of 15° to the right. The results modelled the listener at the central listening position. The graph on the left shows the time series of histograms output by the Supper model, and the graph on the right shows this data averaged over time (the AOT histogram). For clarity, the graph on the left shows only every 50th histogram calculated. Also note that for clarity the y-axis scales are different for the two plots, with a much larger scale for the right hand plot.

Figure 5.2: These two graphs show the combined azimuth histograms (created by summing both the ITD and IID azimuth histograms) averaged over time. The left and right hand graphs correspond to the 1st and 6th stimuli from the third listening test respectively. Both graphs were created by modelling the listener in the right hand position in the listening test. More details of the stimuli and position of the listener can be found in the text. The vertical line in both graphs shows the median of the directions elicited from the subjects in the listening test.
angles obtained in the listening tests. As with the large peaks at ±90°, these spikes appear to be artifacts of the Supper model. The resolution of the angular scale on all of the histograms is 1°, and the spike at 0° only affects the 0° angle and neither of the ±1° angles. The fact that the spike is so narrow suggests that these spikes are artifacts of the Supper model. This led to the decision to remove the average truth value for 0° and replace it with a value obtained by linear interpolation from the values at ±1°. As with limiting the histogram to ±80°, removing the spikes at 0° from the AOT histograms before using the peak picking method continued to improve the match between the interpreted results of the Supper model and the results of the listening tests.

In summary, it was found that an additional two steps were necessary when interpreting the results of the Supper model to obtain good results. These are, firstly, to limit the AOT histogram to ±80° and, secondly, to ignore the 0° value and recalculate it using linear interpolation. These two steps were derived using a heuristic method, i.e. the only justification for using them is that they improved the results. However, there are disadvantages to both of these steps. The first is that ignoring those parts of the AOT histograms that are close to ±90° means that the method cannot produce good results for stimuli where the sound source is genuinely positioned near ±90°. Similarly, the second drawback is that small errors are likely to be introduced when the sound source is genuinely at 0°. Also, the anomalies in the results from the Supper model may be symptomatic of underlying problems in the algorithms used by the Supper model.

5.1.2 Overview of the algorithm in the Supper model

The Supper model was designed to analyse spatial attributes from binaural stimuli. It takes as its input a pair of binaural signals and outputs a series of azimuth histograms. An overview of the process used to calculate these histograms is shown in Fig. 5.3. The first stage in the process uses a filter bank to separate each of the left and right ear binaural signals into twenty-four critical bands. Details of the frequencies of the critical bands are contained in Table 5.1 and the amplitude responses of the filters are plotted in Fig. 5.4. This generates a binaural signal pair for each of the 24 critical bands, giving a total of 2 × 24 = 48 signals. All of these signals are then rectified and low-pass filtered and then used as the input for three different processes. The first of these processes calculates a series of lateral angle histograms from the IIDs of each binaural signal pair. The second process is similar to the first, but instead of using the IIDs it uses the ITDs of each binaural signal pair to calculate a series of lateral angle histograms. The third process (which uses the 24 filtered binaural signal pairs) calculates the loudness for each critical band.

The 24 histogram series derived from IIDs are then combined with the 24 histogram series derived from the ITDs using a process based on the duplex theory. This theory states that the ITDs contribute more to the lateralisation decision process at low frequencies and IIDs contribute more at higher frequencies. The output of this stage is a time series of histograms for each of the 24 critical bands. The final stage combines these by weighting the results for each critical band, giving as its output a single series of lateral angle histograms.
Figure 5.3: The structure of the process of calculating the source localisation data from the binaural signals.

Figure 5.4: Amplitude responses for the filter bank used in Supper's model. Adapted from Supper [158].
<table>
<thead>
<tr>
<th>Critical Band</th>
<th>Frequency (Hz)</th>
<th>Bandwidth (Hz)</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td>Centre</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>100</td>
<td>60</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>200</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>300</td>
<td>250</td>
</tr>
<tr>
<td>4</td>
<td>300</td>
<td>400</td>
<td>350</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>510</td>
<td>455</td>
</tr>
<tr>
<td>6</td>
<td>510</td>
<td>630</td>
<td>570</td>
</tr>
<tr>
<td>7</td>
<td>630</td>
<td>770</td>
<td>700</td>
</tr>
<tr>
<td>8</td>
<td>770</td>
<td>920</td>
<td>845</td>
</tr>
<tr>
<td>9</td>
<td>920</td>
<td>1080</td>
<td>1000</td>
</tr>
<tr>
<td>10</td>
<td>1080</td>
<td>1270</td>
<td>1175</td>
</tr>
<tr>
<td>11</td>
<td>1270</td>
<td>1480</td>
<td>1375</td>
</tr>
<tr>
<td>12</td>
<td>1480</td>
<td>1720</td>
<td>1600</td>
</tr>
<tr>
<td>13</td>
<td>1720</td>
<td>2000</td>
<td>1860</td>
</tr>
<tr>
<td>14</td>
<td>2000</td>
<td>2320</td>
<td>2160</td>
</tr>
<tr>
<td>15</td>
<td>2320</td>
<td>2700</td>
<td>2510</td>
</tr>
<tr>
<td>16</td>
<td>2700</td>
<td>3150</td>
<td>2925</td>
</tr>
<tr>
<td>17</td>
<td>3150</td>
<td>3700</td>
<td>3425</td>
</tr>
<tr>
<td>18</td>
<td>3700</td>
<td>4400</td>
<td>4050</td>
</tr>
<tr>
<td>19</td>
<td>4400</td>
<td>5300</td>
<td>4850</td>
</tr>
<tr>
<td>20</td>
<td>5300</td>
<td>6400</td>
<td>5850</td>
</tr>
<tr>
<td>21</td>
<td>6400</td>
<td>7700</td>
<td>7050</td>
</tr>
<tr>
<td>22</td>
<td>7700</td>
<td>9500</td>
<td>8600</td>
</tr>
<tr>
<td>23</td>
<td>9500</td>
<td>12000</td>
<td>10750</td>
</tr>
<tr>
<td>24</td>
<td>12000</td>
<td>15500</td>
<td>13750</td>
</tr>
</tbody>
</table>

Table 5.1: The lower and upper frequencies for each critical band in the filter bank in Supper's model. The table also includes the centre frequency, bandwidth and quality factor (Q) of each critical band. Adapted from Supper [158].

CHAPTER 5. DIRECTIONAL LOCALISATION | 115
Figure 5.5: The structure of the process for converting the rectified and low-passed binaural signals for each of the critical bands into the IID histograms. The process for calculating the ITD histograms from the same input data has the same structure.

The structure of the histogram-from-IIDs block is shown in Fig. 5.5. An important point to note here is the use of the IID look-up tables to convert the calculated IIDs to azimuth histograms. These look-up tables have to be generated only once, which reduces the computational load on the model. The structure of the histograms-from-ITDs block is identical, although the generation of the ITDs is necessarily different to that of the ITDs.

5.1.3 The creation of the look-up tables in the Supper model

This section describes the process in the Supper model which is used to generate the IID and ITD look-up tables which are used in the localisation algorithm. The process of creating the IID can be divided into two parts. The first part calculates the IIDs for each angle in the range [0°, 180°] using the Gardner-Martin head related impulse response (HRIR) database. This database was created using a KEMAR dummy head and consists of HRIR for a large pinna and a small pinna. Note that only half of the full 360° range needs to be calculated due to symmetry. The second part populates the truth values in the look-up table using the IID values calculated for both the large and small pinnae. Only the second part of the process will be described in this section, as this was the part that was changed during the course of the investigation into the Supper model. The process of creating the IID look-up tables is illustrated by Figs. 5.6 to 5.9.
Populating the table with IIDs from the large and small ear HRIRs

Each IID look-up table consists of a truth value (i.e. a value in the range [0 1]) for each combination of angle and IID. In the left hand graph in Fig. 5.7 the entries in the table that correspond to the IIDs calculated using the large ear data are assigned a truth value of 1. All the other entries in the table have a truth value of 0. The entries in the table have been shaded to illustrate their values: black represents a truth value of 1 and white represents a value of 0. The right hand graph in Fig. 5.7 shows the table populated in a similar manner for the small ear IID table.

Let Intensity$_L$ and Intensity$_R$ be the signal intensities of the left and right ear signals respectively. The IID values are calculated from the left and right ear impulse responses as:

\[ \text{IID} = \frac{\text{Intensity}_R - \text{Intensity}_L}{\text{Intensity}_R + \text{Intensity}_L} \]  

Consequently, the IID values are in the range \([-1,1]\) and are dimensionless. Only angles in the range \([0°\, 180°]\) are entered in the look-up table (taking advantage of left-right symmetry), so only IID values in the range \([0,1]\) are entered in the table. The look-up table has an IID resolution of 0.02, giving 51 rows, each corresponding to a different IID value.

The two tables shown in Fig. 5.7 are then combined into a single table. Each entry is assigned a value of 1 if the corresponding entry in either of the two tables shown in Fig. 5.7 has a value of 1. Otherwise the entry is assigned a truth value of 0. The left hand graph in Fig. 5.8 shows the resulting table. Plot A in Fig. 5.6 illustrates a single column (i.e. corresponding to a single angle) from this table. Plot A shows the two entries with truth values of 1, one of which will correspond to the IID calculated from the small ear HRIRs, and the other to the IID from the large ear HRIRs.

Filling the enclosed region

This table is modified further by filling in the region between the large and small ear truth values with values of 1. This is done by looping through the columns in the table, each of which corresponds to a given angle. This process is illustrated by plots A and B in Fig. 5.6, and the resulting truth value table is shown in the centre graph of Fig. 5.8.

Applying linear transitions (ramping)

The next step in the generation of the IID look-up tables is to apply a linear transition at the edges of the regions with truth values of 1. Again, this is done by looping through the columns in the table. The length of these transitions is dependent on the distance between the large and small ear IIDs: the larger this distance the longer and shallower are the linear truth value ramps applied above and below the original region. This step is shown in plot C Fig. 5.6, with the resulting table shown in the right hand graph of Fig. 5.8.
Figure 5.6: This shows the how the truth values are populated for each column of the IID look-up table. The top graph (A) shows only two entries with non-zero truth values. Both of these values are one and they correspond to the IIDs calculated from the large and small ear HRIRs. The second graph (B) shows truth values of one assigned to the region between the large and small ear IIDs. The third graph (C) shows the linear ramps applied both above and below the region with non-zero truth values. The fourth graph (D) shows the region of non-zero truth values being extended for large IIDs (the grey area shows the region of the column that has changed from plot C).

Limiting the range to \([0° \ 90°]\) (folding)

After this the IID look-up tables are folded around the 90° angle. This means that any binaural signals corresponding to sound sources in the rear hemisphere are mapped onto the front hemisphere through reflection in the vertical plane passing through the listener's ears. Disambiguating front and back sources typically involves head movements in real listeners. The Supper model accepts arbitrary binaural signals as its input and cannot interact with the generation of these signals in the way that real listeners can move their heads. This means that the Supper model has no way of differentiating between front and back sources, which explains the strategy of mapping all the binaural signals into the front hemisphere. The truth value table resulting from this reflection at 90° is shown in the left hand graph of Fig. 5.9.

Extending the table for large IIDs

The final step in the generation of the IID look-up table is to extend the table for large IIDs. This is only done on those columns in the look-up table with the highest non-zero truth values. In all these columns all the truth values above the highest non-zero truth value are assigned an arbitrary value of 0.2. This process is illustrated in plot D in Fig. 5.6. Let \(IID_{\text{max}, B}\) be the largest IID calculated from the Gardner and Martin large and small ear impulse responses for the \(B\)th critical band. The process described in this section ensures that for each critical band any IIDs greater than \(IID_{\text{max}, B}\) will be assigned similar truth value histograms as those corresponding to \(IID_{\text{max}, B}\). The smaller truth values of 0.2 reflect the fact that there is less certainty about these results. The final IID look-up table is shown in the right hand graph of Fig. 5.9.
Figure 5.7: The two graphs show the IIDs plotted against angles for the 18th critical band. The left hand graph was calculated using the large ear results in the Gardner and Martin HRIR database and the right hand graph was calculated using the small ear results.

Figure 5.8: These graphs show intermediate stages in the generation of the IID look-up table for the 18th critical band. The left hand graph shows large ear and small ear results plotted on the same axes. The centre graph shows the area between the two sets of results having been filled in. The right hand graph shows the addition of ramps above and below the filled in areas.

Figure 5.9: The left hand graph shows IID results above 90° folded over. The right hand graph shows the graph extended for large IID values. The data shown in the right hand graph is part of the completed IID look-up table. Both graphs use the data from the 18th critical band.
5.1.4 Investigating and modifying the creation of the look-up tables

This section describes the investigation into the look-up table generation algorithm. It describes the changes that have been made to this process to improve the integration of the Supper model with the model framework described in Chapter 3. This section also contains the justification for making these changes.

Removing the \([0° \, 90°]\) folding

Perhaps the biggest difference between the model framework developed in this project and the original Supper model is that in the former the positions and orientations of the sound sources and listeners are modelled. This means that head movements can be incorporated into the model together with the resulting changes in the binaural signals. This should be compared to the Supper model, where the model itself has no influence on the creation of the binaural signals.

One result of this is that there is no longer any need to map all of the sound sources onto the front hemisphere. This is because it is now possible to determine whether a sound source is in the front or rear hemisphere by observing the effects of head movements on the calculated IIDs. Once this has been determined, the 0° to 90° portion of the IID look-up table can be used to determine the azimuth of sources in the front hemisphere and 90° to 180° portion of the table can be used for those sources in the rear hemisphere.

The head movement functionality is not covered in this chapter, so only sources in the front hemisphere need to be considered. Consequently the IIDs for the rear hemisphere will no longer be folded onto the front hemisphere, resulting in the IID look-up table shown in Fig. 5.10. This means that each IID will correspond to a truth value histogram with less ambiguous peaks (as those peaks associated with sources in the rear hemisphere will be removed). This results in clearer, more accurate histograms.

Removing gaps

Once the IID look-up table algorithm had been modified so that the IIDs from the rear hemisphere were no longer mapped onto the front hemisphere, another issue arose. This was the presence of rows with a maximum value less than 1 in the IID look-up tables for some of the critical bands. In particular, the rows that are of concern are those whose maximum value is less than one and also which are below the row corresponding to the maximum IID from the large and small ear HRIRs. These will be termed “max-less-than-one” (MLTO) rows to facilitate the discussion. Fig. 5.11 shows rows 18 and 20 are MLTO rows in the IID look-up table for the 19th critical band.

The MLTO rows occur when there is a steep gradient in the curves of IID against localisation angle and the large and small ear IID values coincide for a given angle. They arise because only one element in each column is given a truth value of 1 when the values for the large (and small) ear IIDs are first entered in the table (as described in Section 5.1.3). This can be seen in Fig. 5.12. The removal of the rear-to-front mapping described in Section 5.1.3 decreased the number of elements with a truth value of 1 in the \([0° \, 90°]\) range of
Figure 5.10: The left hand plot shows the HD look-up table for the 18th critical band, which included the folding of the truth values in the rear hemisphere into the front hemisphere. The right hand plot shows the same look-up table when the folding has been omitted.

Figure 5.11: The left hand plot shows the HD look-up table for the 19th critical band just before the table is extended upwards (i.e. before Section 5.1.3). The right hand plot is a magnified section of this table. This shows rows 18 and 20, which do not have maximum values of 1.
the look-up table. One factor in this is that the difference between the large and small ear IIDs for a given angle is generally greater in the range [90° 180°]. Hence, there is an increase in the number of MLTO rows.

The occurrence of MLTO rows became an issue because of the algorithm used to extend the table upwards (Section 5.1.3). This algorithm loops through the rows of the table, starting at the bottom row, until it finds an MLTO row. The row just below this is then used to extend the table upwards. This process uses the assumption that the first MLTO row encountered in this way is just above the maximum IID calculated from the large and small ear HRIRs. When this is not the case, which can occur when the [0° 90°] folding has been removed from the look-up table generation, then this algorithm starts extending the table upwards too early. The resulting table for the 19th critical band IID look-up table is shown in the left hand graph in Fig. 5.13. This table will give erroneous localisation histograms for large IIDs.

Even without the issue of the unexpected behaviour of the extending algorithm, the presence of MLTO rows is undesirable. The IIDs corresponding to the MLTO rows will have less influence on the overall output compared to the adjacent IIDs, leading to systematic errors in the output of the model. For this reason the decision was made to ensure no MLTO rows below the maximum IID from the large and small ear HRIRs, as these systematic errors would still be present if only the extending algorithm was modified. More processing to remove any MLTO rows was added after the stage described in Section 5.1.3 where the large and small ear IIDs have just been entered into the table. This additional processing simply interpolated a line between the IID values for adjacent angles in the table and any elements in the table that were crossed by this line were assigned a truth value of 1. The results of this processing on the area of the table in the right hand plot in Fig. 5.12 can be seen in the right hand plot of Fig. 5.13. This additional processing successfully prevented any unwanted MLTO rows, and so the extending algorithm (Section 5.1.3) once again behaved as expected.

When the only modification of the Supper model had been the removal of the rear-to-front mapping (Section 5.1.4), the unexpected behaviour of the extending algorithm had meant that no useable IID look-up tables had been created. This was remedied by ensuring no MLTO rows below the maximum IID value calculated from the HRIR database. This means that a comparison can be made between results from the original Supper model and results from the model when it has been modified as described in this and the previous section. Fig. 5.14 shows the results from the original Supper model for one of the stimuli from the third listening test. Fig. 5.15 shows the results from the modified model for the same stimulus. From looking at the two sets of graphs there can be seen a clear improvement resulting from the modifications made to the look-up tables. The two largest peaks in the AOT histogram in Fig. 5.14 are around the angles -50° and -30°. Indeed, all the bins in the range [-10° -90°] have values with a similar order of magnitude, so the results are ambiguous. This is particularly visible in the left hand graph of Fig. 5.14. This should be compared with Fig. 5.15 where there is a clear angle of localisation in both graphs. Furthermore, this angle is around 15°, which corresponds quite closely to the panned position of the stimuli at -13°.
Figure 5.12: The left hand plot shows the IID look-up table for the 19th critical band just after the first stage described in Section 5.1.3, where the large and small ear IIIDs have just been entered into the table. The right hand plot is a magnified section of this table showing the gaps in rows 18 and 20.

Figure 5.13: The left hand plot shows the IID look-up table for the 19th critical band, showing the unexpected behaviour of the extending algorithm (see Section 5.1.3). The right hand plot shows the same area of the table shown in Fig. 5.12, but with the modifications to remove the gaps previously in rows 18 and 19.
Figure 5.14: These graphs show the intermediate results for stimulus 5 from the third listening test. The graph on the left shows the time series of histograms output by the Supper model and the graph on the right shows the AOT histogram. These results were calculated using the original IID and ITD look-up tables. The stimulus consisted of a one second burst of pink noise. The sample was played through two loudspeakers 2.2m away from the centre of the listening area and at angles of ±30°. The constant power panning law was used to position the sample at an angle 13° to the left of the central listening position. The listener was modelled facing forwards at the central listening position.

Figure 5.15: These graphs show the intermediate results for stimulus 5 from the third listening test, as in Fig. 5.14. The graph on the left shows the time series of histograms output by the Supper model and the graph on the right shows the AOT histogram. The IID and ITD look-up tables used to generate these results have no rear-to-front folding (Section 5.1.4) and no gaps (Section 5.1.4).
Reducing the spikes at 0°

This still leaves the spikes in the AOT histograms at 0°. These are due to the fact that only the angles [0° 90°] are included in the look-up table. The histograms returned by the Supper model cover the range [-90° 90°] in 1° increments. If the IID is positive then the corresponding row in the IID look-up table is used to populate the [0° 90°] range of the output histogram and the [-90° -1°] part of the histogram is set to all zero truth values. Conversely, if the IID is negative then the corresponding row in the IID look-up table is mapped onto the [-90° 0°] part of the histogram and the [1° 90°] range is set to all zero truth values.

However, the ramped truth values (see Section 5.1.3) have the effect of smoothing the histogram output of the Supper model. These linear ramps are created by looping through the columns of the look-up table (ie. looping over each angle in the range [0° 90°]). As for each critical band the IID look-up table is two dimensional, the overall effect of this is similar (though not identical) to applying the ramping by looping over the rows in the table. The effect of using a look-up table with only the range [0° 90°] is that the effect of this ramping is not present on both sides around 0°.

This problem was overcome in the following way. The truth value table in Sections 5.1.3 to 5.1.3 is modified so that the number of rows is increased from [0 50] to [-50 50]. There is no change in the way the truth values for the large and small ear IIDs are entered into the table and the enclosed region is filled (Sections 5.1.3 to 5.1.3). However, the increase in the number of rows means that the ramps which are added in Section 5.1.3 can now extend below the row corresponding to a zero IID.

The bottom half of the table is then rotated around the origin, as shown in Fig. 5.16, resulting in a look-up table that covers [-90° 90°] with the same number of rows as the original table. This new table is twice as long as the original table, and the truth values added in Section 5.1.3 below the zero IID row are now in the columns corresponding to the angle [-90° 0°]. This means there is no longer any abrupt cut-off in the truth values around 0°, and hence no more spikes at 0°.

Fig. 5.17 illustrates how removing the abrupt cut-off at 0° in the look-up tables reduces the spikes at 0°. These two graphs also show how this can affect the overall results of the model. If a peak-picking method is used to obtain a single angle of localisation from the AOT histogram then the reduction of the 0° spike results in a shift of 15° to the left. The nature of the stimulus means that it is hard to determine the angular localisation of the signal. However, the fact that the pair of loudspeakers through which the signals are played are at 45° to the left and 15° to the right suggests that the look-up tables with the 0° spike reduction are more accurate.

Normalising the values

Consider the IID look-up table for the 18th critical band shown in the left hand graph of Fig. 5.18. This table has been modified from the original look-up table in the Supper model as described in Sections 5.1.4 to 5.1.4. Here the column corresponding to 90° has a lot of non-zero truth values. This means that a large
Figure 5.16: This graph shows the rotation of the bottom half of the IID look-up table about the origin through 180° to ensure there is no abrupt cut-off at 0°, thus reducing the occurrence of spikes at 0°.

Figure 5.17: These graphs show AOT histograms for stimulus 8 from the third listening test. The graph on the left shows the results when the look-up tables had only been modified according to Sections 5.1.4 (no rear-to-front mapping) and 5.1.4 (no gaps). The graph on the right shows the results when the look-up tables had the same modifications and also the modifications described in Section 5.1.4 to reduce the spikes at 0°. The stimulus consisted of a one second burst of pink noise. A decorrelated version of the original sample was created using an all-pass filter. The original signal and the decorrelated version were then panned between two loudspeakers positioned 3m away from the centre of the listening area at angles of 45° to the left and 15° to the right. The listener was modelled facing forwards at the central listening position.
range of calculated IIDs will result in the look-up table returning a truth value histogram with a non-zero value for the 90° bin. Hence angles that have non-zero truth values corresponding to a large range of IIDs (such as 90° in this example) have undue influence when all the intermediate results are combined to give a single direction of localisation. This means that those angles where the relationship between the angles and IIDs is ambiguous are emphasised. This is not the desired performance of the model, where ideally the least ambiguous intermediate results should have the most influence on the overall results.

In order to improve the output of the model three more steps were added to the IID look-up table generation algorithm. The first step is to determine the maximum value of each column in the table. The second step is to normalise each column in the table, i.e. multiply each column by a factor to ensure the column sums to one. The third step is to multiply the normalised column by the maximum value calculated in the first step. The resulting vertically normalised IID look-up table for the 18th critical band is shown in the right graph in Fig. 5.18. Fig. 5.19 illustrates the effect of the vertical normalisation on the look-up tables. The left hand graph shows the results using look-up tables without the vertical normalisation. The maximum peak in this graph is at 90°. This should be compared with the right hand graph, which shows the results from using the vertically normalised look-up tables. Here the maximum peak is around 60°.

Consider the IID look-up table for the 21st critical band shown in the left hand graph of Fig. 5.20. Now consider the two cases corresponding to calculated IIDs of 25 and 48 respectively. The truth value histogram that is returned for the IID of 25 has a total area under the curve of 3.37. In comparison, the area under the curve for the truth value histogram returned for the IID of 48 is 21.33. This means that the truth value histogram for an IID of 48, where the angular localisation is ambiguous, has more influence on the final results than the histogram for an IID of 25, where the certainty of localisation is greater. This demonstrates how having the maximum values in each row to always be 1 has the effect of emphasising the histograms for which there is least certainty and can lead to anomalous results. This is the opposite of the desired performance of the model, where ambiguous intermediate results should have less influence on the overall results than the intermediate results which are more certain. This also explains the large peaks at ±90°, as discussed in Section 5.1.1.

Using a similar method as that used for the vertical normalisation, three more steps were added to the IID look-up table generation algorithm. The first step is to determine the maximum value of each row in the table. The second step is to normalise each row in the table, i.e. multiply each row by a factor to ensure the row sums to one. The third step is to multiply the normalised row by the maximum value calculated in the first step. The resulting horizontally normalised IID look-up table for the 21st critical band is shown in the right hand graph in Fig. 5.20. Fig. 5.21 illustrates the effect of the horizontal normalisation on the look-up tables. The left hand graph shows the results for a stimulus from the third listening test using look-up tables without the normalisation. The maximum peak in this graph is at 0°. This should be compared with the right hand graph which shows the results using normalised look-up tables. Here the maximum peak is around −18°.
Figure 5.18: The left hand plot shows the IID look-up table for the 18th critical band before vertical normalisation has been applied. The right hand plot shows the same table after vertical normalisation has been applied.

Figure 5.19: These graphs show $AOT$ histograms for stimulus 4 from the third listening test. The graph on the left shows the results when the look-up tables had been modified as described in Sections 5.1.4 to 5.1.4. The graph on the right shows the results when the look-up tables had the same modifications and also the vertical normalisation described in Section 5.1.4. The stimulus consisted of a one second burst of pink noise. The sample was played through two loudspeakers 3m away from the centre of the listening area and at angles of $15^\circ$ and $60^\circ$ to the right. The constant power panning law was used to position the sample at an angle $44^\circ$ to the right of the central listening position. The listener was modelled facing forwards at a point 75cm away and at an angle $135^\circ$ to the right of the central listening position.
Figure 5.20: The left hand plot shows the IID look-up table for the 18th critical band when only horizontal normalisation has been applied. The right hand plot shows the same table after both horizontal and vertical normalisation have been applied.

Figure 5.21: These graphs show AOT histograms for stimulus 4 from the third listening test. The graph on the left shows the results when the look-up tables had been modified as described in Sections 5.1.4 to 5.1.4 together with the horizontal normalisation described in Section 5.1.4. The graph on the right shows the results when the look-up tables had these same modifications with the addition of the vertical normalisation also described in Section 5.1.4. A description of the stimulus is included in the caption for Fig. 5.21. The listener was modelled facing forwards at a point 75cm away and at an angle 135° to the right of the central listening position.
5.1.5 Calculating the correlograms

The component correlogram at time \( t \) is calculated as

\[
C_M(t) = \begin{bmatrix}
L_{t+(-M)} \times R_{t-(M)} \\
L_{t+(-M+1)} \times R_{t-(M+1)} \\
L_{t+(-M+2)} \times R_{t-(M+2)} \\
\vdots \\
L_{t+(M-2)} \times R_{t-(M-2)} \\
L_{t+(M-1)} \times R_{t-(M-1)} \\
L_{t+(M)} \times R_{t-(M)}
\end{bmatrix},
\]

where \( L_t \) and \( R_t \) are the left and right binaural streams respectively, \( M \) determines the length of the correlogram. The units of time are chosen to be the time between consecutive samples in the binaural streams. Each correlogram used to calculate an ITD is the sum of a given number, \( N \), of consecutive component correlograms:

\[
J_{M,N}(t) = \sum_{n=0}^{N-1} C_M(t + n) = \begin{bmatrix}
\sum_{n=0}^{N-1} (L_{t+(n-M)} \times R_{t-(n-M)}) \\
\sum_{n=0}^{N-1} (L_{t+(n-M+1)} \times R_{t-(n-M+1)}) \\
\sum_{n=0}^{N-1} (L_{t+(n-M-1)} \times R_{t-(n-M-1)}) \\
\sum_{n=0}^{N-1} (L_{t+(n+M)} \times R_{t-(n+M)})
\end{bmatrix},
\]

In the Supper model \( M = 9 \) (i.e. each component correlogram has \( 2M + 1 = 19 \) bins) and \( N = 14 \). This
should mean that 14 component correlograms are summed for each ITD correlogram. In fact, for the lower
critical bands $N$ is increased to ensure that it is longer than one period of any signal in the critical band.
In addition to this, $1$ in $6$ decimation is built into the model. This means that the indexing into the two
binaural streams increases by $6$ for each consecutive ITD correlogram. This indexing is illustrated in Fig.
5.22. Let $H_k$ be the correlogram used to calculate the $k$th ITD value. If $c$ is the initial offset used as an
index into the two binaural streams, then

$$H_k = J_{9,14}(6k + c)$$  \hspace{1cm} (5.4)

The Supper model makes use of the following property to reduce the number of calculations required:

$$H_k = H_{k-1} - J_{9,6}(6k - 6 + c) + J_{9,6}(6k + 8 + c)$$  \hspace{1cm} (5.5)

The derivation of this equality is shown in Appendix F. This means that only the component correlograms
that contribute to the change from $H_{k-1}$ to $H_k$ are calculated. However, the original implementation of the
Supper model included an error in the indexing:

$$H'_k = H'_{k-1} - J_{9,7}(6k - 6 + c) + J_{9,7}(6k + 7 + c)$$  \hspace{1cm} (5.6)

where $H'_k$ is the $k$th ITD correlogram calculated by the incorrect indexing. Fig. 5.23 illustrates the effect of
this error in the indexing.

$I'_k$ can be shown to be equal to

$$I'_k = H_k + E_k$$  \hspace{1cm} (5.7)

where the $E_k$ is the error between $H_k$ and $I'_k$ and is given by

$$E_k = \sum_{n=1}^{k-1} C_9(6n + 7 + c) - \sum_{n=1}^{k-1} C_9(6n + c)$$  \hspace{1cm} (5.8)

The proof of Eqs. 5.7 and 5.8 is given in Appendix F. The magnitude of this error will vary with different
binaural streams. Fig. 5.25 illustrates the effect of correcting the indexing for the correlograms. The left hand
graph shows the results for one of the stimuli from the third listening test using the original indexing. The
right hand graph shows the same results when the indexing has been corrected. The dominant features in
the left hand graph are the spikes at $0^\circ$ and $-12^\circ$. Indeed, the spike at $0^\circ$ would be chosen by a peak-picking
algorithm as the angle of localisation for the stimulus. However, when the indexing for the correlograms is
corrected then the spike at $-12^\circ$ disappears and the spike at $0^\circ$ is greatly reduced. This demonstrates how
artifacts introduced by the errors in indexing can dominate the final results.
Figure 5.22: The graph shows how the correlograms are calculated. Each block represents the component correlogram $C_0(t)$, where time $t$ is on the x axis. The constant $c$ is the initial offset. Blocks above the horizontal lines show component correlograms that are added to the correlogram. For instance, the correlogram $H_2$ is equal to $\sum_{i=12}^{25} C_0(t + c)$.

Figure 5.23: The graph shows how the correlograms are miscalculated due to the errors in the indexing. Each block represents the component correlogram $C_0(t)$, where time $t$ is on the x axis. The constant $c$ is the initial offset. Blocks above the horizontal lines show component correlograms that are added to the correlogram. Conversely, blocks below show component correlograms that are subtracted from the correlogram. For instance, the correlogram $H'_2$ is equal to $\sum_{i=13}^{25} C_0(t + c) + C_0(13 + c) + C_0(19 + c) - C_0(6 + c)$. 
Figure 5.24: These graphs show AOT histograms for stimulus 18 from the third listening test. The graph on the left shows the results with all the modifications to the Supper model described in Sections 5.1.4 and 5.1.5 and with the IID and ITD histograms combined additively. The graph on the right shows the same results but taking the product of each of the IID and ITD histogram bins. The stimulus consisted of a trumpet playing the opening phrase of “Over the Rainbow”, which lasted 3 seconds. The sample was played through a loudspeaker 3m away from the central listening position at an angle 60° to the left. The listener was modelled facing forwards at the central listening position.

Figure 5.25: These graphs show AOT histograms for stimulus 4 from the third listening test. The graph on the left shows the results when the look-up tables had been modified as described in Section 5.1.4. The graph on the right shows the results when these same look-up table modifications and also with the corrections made to the indexing used to calculate the correlograms used in the ITD processing. A description of the stimulus is included in the caption for Fig. 5.14. The listener was modelled facing forwards at a point 75cm away and at an angle 135° to the right of the central listening position.
5.1.6 Combining the IID and ITD histograms

In Ben Supper's thesis the suggested method of combining the IID and ITD results is to sum their output histograms. The tacit assumption behind this is that the IID and ITD histograms both have the same scale. While this may have been true for the original implementation of the Supper model, it is doubtful if it is true for the modified model, which includes the use of IID and ITD look-up tables that have been normalised in two dimensions. Even if the assumption that both the IID and ITD output histograms have the same scale is true, the behaviour of the model as a result of summing the IID and ITD output histograms is not always exactly what is desired.

The IID and ITD cues may not always be in agreement. This is the case particularly for stimuli from reproduction systems using multiple loudspeakers. In this case any spurious spikes in either the IID or ITD histograms will still be present to a greater or lesser degree in the summed combined output. Now consider if the IID and ITD histograms are combined by multiplying corresponding bins for each pair of IID and ITD output histograms at each time slice. If a given peak in the IID output histogram agrees with a similar peak in the ITD output histogram then a corresponding peak will appear in the combined output histogram. Conversely, if a peak in one of the IID or ITD output histograms does not coincide with a peak in the other histogram then these bins will have very low values in the combined histogram (as the small values in the corresponding bins of the second histogram will attenuate the peaks of the first histogram when multiplied together). This means that combining the IID and ITD output histograms by taking the product of corresponding bins accentuates the angles where the two component histograms agree and attenuates those angles where they disagree.

Fig. 5.24 demonstrates the difference in the combined output histogram using the two methods (either summing or taking the product) for one of the stimuli from the third listening test. When the IID and ITD output histograms are summed there is a large spike at around 18°. When the two histograms are combined by taking the product of corresponding bins for each time slice, this spike disappears. As the stimulus was coming from a single loudspeaker 60° to the left, the spike at around 18° is spurious. In general it was found that combining the two histograms by taking the product of corresponding bins improved the performance of the Supper model.

5.1.7 Comparing the results of the original Supper model with the modified model

In this section a comparison is made between the performance of three different versions of the model. The first version being considered uses the original implementation of the Supper model and interprets the results from this by constructing an AOT histogram and then using a peak-picking method to obtain a single angle of localisation. In this version the IID and ITD results were combined, as suggested in Ben Supper's model. The second version of the model also uses the original implementation of the Supper model. However, it then uses the heuristic method described in Section 5.1.1 to interpret the results of the Supper model. The
third version has modified the Supper model as described in Sections 5.1.4 and 5.1.5 and the IID and ITD results were combined using multiplication.

Each stimulus used in the three listening tests was input into each of the three versions of the model and the obtained results were compared with the directional localisation results obtained from the subjects in the listening tests. Two groups of stimuli were omitted from the results of the first listening test discussed here. The first of these groups consists of the sine waves which were faded in and out. This group has been omitted because of the difficulty the subjects experienced in localising the stimuli. The second group to be omitted consists of those stimuli where two of the channels were misallocated. In a similar fashion, all the stimuli from the second and third listening tests that had been artificially widened have been omitted from the results discussed here. These stimuli have been omitted as they had proved relatively difficult to localise for both the subjects in the experiments and for the model.

The results are shown in Fig. 5.26. It can be seen from these graphs that the latest version of the model, which includes the modifications to the Supper model described in this document, performs the best. This can be seen in two ways. The first is how closely the model results have a linear relationship with the listening test results. As well as being able to see this from the graphs in Fig. 5.26, this is confirmed by the calculated coefficients of determination ($R^2$) values shown in Table 5.2. The second indicator of the performance of the models is the gradient of the line of best fit passing through the origin. Ideally this should be one, so that the predicted angles calculated by the models are identical to the angles obtained from the listening tests. Table 5.3 shows that the calculated gradients for the latest version of the model are indeed closer to one.

![Table 5.2: This table shows the coefficients of determination ($R^2$) between the results for the three listening tests and the model in the three states.](image)

<table>
<thead>
<tr>
<th>Listening test</th>
<th>Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Heuristic</td>
<td>Modified</td>
</tr>
<tr>
<td>1</td>
<td>0.735</td>
<td>0.790</td>
<td>0.902</td>
</tr>
<tr>
<td>2</td>
<td>0.911</td>
<td>0.937</td>
<td>0.992</td>
</tr>
<tr>
<td>3</td>
<td>0.897</td>
<td>0.929</td>
<td>0.987</td>
</tr>
</tbody>
</table>

![Table 5.3: This table shows the gradient of the line of best fit that passes through the origin between the results for the three listening tests and the model in the three states.](image)

<table>
<thead>
<tr>
<th>Listening test</th>
<th>Model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>Heuristic</td>
<td>Modified</td>
</tr>
<tr>
<td>1</td>
<td>1.26</td>
<td>1.15</td>
<td>1.03</td>
</tr>
<tr>
<td>2</td>
<td>1.40</td>
<td>1.39</td>
<td>1.17</td>
</tr>
<tr>
<td>3</td>
<td>1.46</td>
<td>1.42</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Other notable features in the graphs shown in Fig. 5.26 are the horizontal rows of data points. In the middle column of graphs, which show the results of using the original Supper model with the heuristic algorithm for
Figure 5.26: These nine graphs show the listening test results plotted against the localisation angles predicted by the model. The top row of graphs corresponds to the first set of listening tests, the second row to the second set of listening tests and the third row to the third set of listening tests. The first column of graphs shows the listening test results plotted against the results of using a peak-picking method to interpret the unmodified model. The second column shows the listening test results plotted against the results of the unmodified model interpreted using the heuristic method. The third column shows the listening test results plotted against the model after it has been modified as described in this document. In each graph the dashed line shows the ideal relationship and the dotted line is the line of best fit passing through the origin.
Table 5.4: Summary of the steps in the generation of the look-up tables and the modifications that have been made to the Supper model.

interpreting the output histograms, there are horizontal rows at ±80°. These can be explained as artifacts of the heuristic interpretation algorithm, which considers only the portion of the output histograms in the range [-80° 80°]. The other horizontal rows of data points in the top row of graphs in Fig. 5.26 are of more concern. In particular, the horizontal row of data points at 0° in the top right graph (the results from the modified model for the first listening test stimuli) need further investigation.

5.1.8 Summary

This section began by describing some of the problems that have been experienced using the Supper model. This was followed by a description of the initial heuristic interpretation of the results from the Supper model and some of the drawbacks of using this approach, and then a brief overview of the relevant parts of the algorithm of the Supper model. The largest part of this section described the investigation into the Supper model with the aim of improving the performance and integration of the Supper model with respect to the rest of the model developed in this project. Each modification which has been made to the Supper model has been described, together with a discussion of why the modification was made and an example of how the modification has improved the performance of the model. Table 5.4 summarises all the changes that
have been made to the Supper model. The details of the modifications are then followed by a short section comparing the performance of the model prior to the modification with the performance of the modified model. This demonstrated that the modifications described in this report improve the model's performance in terms of predicting the results of listening tests. Another benefit of the modification is that the output histograms can be converted into a single angle using relatively simple peak-picking method. This should be compared with the model before the modifications, where the heuristic algorithm used to interpret the results of the Supper model became increasingly complicated.

5.2 Validation of the modified Supper directional localisation model

This section validates the Supper localisation model, modified as described in Section 5.1, against the localisation results from the three listening test experiments described in Chapter 4. Table 5.5 summarises the results of the validation.

<table>
<thead>
<tr>
<th>Listening test</th>
<th>Description</th>
<th>$R^2$</th>
<th>RMSEP</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All stimuli except sine waves</td>
<td>0.881</td>
<td>8.24</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>All stimuli except sine waves and those with misallocated channels</td>
<td>0.902</td>
<td>7.68</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>All anechoic stimuli except sine waves and those with misallocated channels</td>
<td>0.932</td>
<td>8.00</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>All reflective stimuli except sine waves and those with misallocated channels</td>
<td>0.282</td>
<td>7.04</td>
<td>0.14</td>
</tr>
<tr>
<td>2</td>
<td>All stimuli</td>
<td>0.992</td>
<td>2.92</td>
<td>1.15</td>
</tr>
<tr>
<td>3</td>
<td>All stimuli</td>
<td>0.860</td>
<td>10.63</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>Single loudspeaker stimuli and panned stimuli</td>
<td>0.985</td>
<td>3.83</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>Artificially widened stimuli</td>
<td>0.000</td>
<td>14.80</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 5.5: The results of validating the modified Supper model against the results of the three listening test experiments. The last column contains the gradient of the line of best fit for the data, constrained to pass through the origin to preserve left-right symmetry.
5.2.1 Validation using the results from the first listening test experiment

Figs. 5.27 and 5.28 show results from the first listening test plotted against the results obtained by putting the stimuli from the listening test through the model. Each circle on these graphs corresponds to a combination of listening position and listening test stimulus. Each of the four graphs shows a different subset of the set of all listener position and stimulus combinations. The dashed lines in the graphs in Figs. 5.27 and 5.28 show the ideal relationship between the listening test results and the model results. The dotted lines in the same graphs show the line best fit through the data points. It is assumed that both the model and all the subjects from the listening test have left-right symmetry with respect to perceiving sound. Thus, a further constraint was applied to the line of best fit, which was that it should pass through the origin. The gradient, $p$, of this line of best fit is also shown on each graph, together with the coefficient of determination, $R^2$.

The left hand graph in Fig. 5.27 shows all the results except the sine wave stimuli. The sine wave stimuli have been excluded because of the difficulty localising these stimuli experienced by the subjects in the listening test. The gradient of the line of best fit, $p$, is 0.99, which is close to the ideal of unity. However, it can be seen from the graph that the majority of the data points are situated below the dashed line which shows the ideal relationship. Having more data points below the dashed line shows that the model tends to predict the location of the stimuli further to the left than the subjects of the listening test. If the assumption is made that both the model and all the subjects in the listening test have left-right symmetry then this is of less concern. The model was designed and built using the assumption of left-right symmetry, so if the model is not symmetric then it is a mistake in either the design or implementation of the model. The symmetry of the model was checked by reflecting the loudspeaker positions in the vertical plane passing the centre of the listening area dividing the listening area into left and right. The angles obtained from doing this were identical to the original results from the model but with the signs reversed, which confirms that the model is symmetric. Another possible cause of the model predicting the stimuli further to the left than the listening test results is the existence of systematic errors in the listening test itself. These could be caused by incorrectly positioning the scale or the positions in which the subjects were seated.

The right hand graph in Fig. 5.27 shows all the results with the exception of the sine stimuli and the six stimuli with misallocated signals. The six stimuli with misallocated signals have been excluded because the fact that the signal intended for the Rear Right loudspeaker was mistakenly routed to the Extra Left loudspeaker means that the resulting sound-field will be much less spatially coherent. In this graph the slope of the line of best fit remains close to unity and the removal of the six misallocated-channel stimuli improves the $R^2$ value from 0.88 to 0.90. The greatest concern in Fig. 5.27 is the horizontal row of data points in both graphs. This shows that a number of the stimuli have been assigned an angle of 0° when the results of the listening test show the same stimuli localised across a range of angles from approximately −15° to 20°. Fig. 5.28 shows the results separated into those where the stimuli were generated using anechoic conditions (the left hand graph) and those where the stimuli were generated echoic conditions (the right hand plot). The sine wave stimuli and misallocated-channel stimuli are excluded from both graphs. Here it can be seen that the echoic stimuli have all been localised at 0° with the exception of two data points. Furthermore, these two data points both correspond to the 24th stimulus. This stimulus was created by modelling a single source.
Figure 5.27: These two plots show the results from the first listening test against the results from the model. The left hand graph shows all the results except the sine wave stimuli. The right hand graph shows all the results except the sine wave stimuli and the stimuli with misallocated channels.

Figure 5.28: These two graphs show the first listening test results plotted against the model results. Both graphs omit the sine wave stimuli and the stimuli with misallocated channels. The left and right hand graphs correspond to the stimuli that were generated using an anechoic and an echoic environment respectively.
and single microphone in a reverberant room using CATT Acoustics. The original source signal was a three second sample of African percussion and the processed signal was played through the Centre loudspeaker. In the right hand graph in Fig. 5.28, the data point localised by the model at 10° corresponds to the 24th stimulus and the left listening position. The data point localised by the model at -15° corresponds to the 24th stimulus and the right listening position.

All the echoic stimuli, with the exception of the 24th stimulus, involve signals being played through either two or five loudspeakers. The model has assigned an angle of 0° to all these stimuli. It is clear from the right hand graph in Fig. 5.28 that the model is not able to correctly localise reflective stimuli that are presented over multiple loudspeakers.

5.2.2 Validation using the results from the second listening test experiment

The left hand graph in Fig. 5.29 shows the results for all the stimuli from the second listening test experiment plotted against the results obtained by running the model on the same stimuli signals. Again, each circle on these graphs corresponds to a combination of stimulus and listening position, although, as stated earlier, only a single listening position was used in the second listening test. The $R^2$ value of 0.99 and the root mean square error of prediction value of 2.9° show that the modified Supper model was able to very accurately predict the results of the second listening test experiment.

5.2.3 Validation using the results from the third listening test experiment

The right hand plot in Fig. 5.29 shows the results for all the stimuli from the third listening test experiment. Like the two plots in Fig. 5.27, this plot includes a horizontal row of data points passing through 0°. Again, this shows that a number of the stimuli have been assigned an angle of 0° when the results of the listening test show the same stimuli localised across a range of angles from approximately -50° to 10°. The left hand plot in Fig. 5.30 shows the results for all the stimuli that were either played through a single loudspeaker or else created using just the constant power panning law. The right hand plot shows the results for all the stimuli artificially widened using the combination of an all-pass filter and the constant power panning law. As with the echoic stimuli in the first listening test, it appears from this plot that the model is not able to correctly localise signals that have been artificially widened using this method. This contrasts with the results from the second listening test experiment, where the stimuli that had been artificially widened, albeit using a different method, were able to be localised by the model.

5.2.4 Summary

This section showed that the modified Supper model was able to predict the localisation results from the three listening test experiments with a high degree of accuracy, with the exception of two groups of stimuli. These exceptions are the reflective stimuli from the first listening test experiment and the artificially widened...
Figure 5.29: The left hand graph shows the results for all the stimuli from the second listening test plotted against the results from the model. The right hand graph shows the results for all the stimuli from the third listening test plotted against the results from the model.

Figure 5.30: These two graphs show the third listening test results plotted against the model results. The left hand graph shows the results for those stimuli that were either from a single loudspeaker or were constant power panned between a pair of loudspeakers. The right hand graph shows the results for those stimuli that were artificially widened.
stimuli from the third listening test experiment. For both of these groups of stimuli, the localisation model always returned a predicted angle of 0°.

5.3 The effect of reflections on the directional localisation model

In Section 5.2.1 it was seen that the modified Supper model was unable to predict the results of the reflective stimuli in the first listening test experiment. This section investigates further how the modified Supper model behaves with reflective signals. Two kinds of reflective signals are considered in this section, each of which is investigated separately. The first investigation is of loudspeaker signals created by modelling microphone techniques in reflective environments (processes I and II in Fig. 3.1). The second investigation is of the effects of having a reflective listening environment (processes VI and VII in Fig. 3.1). The reports of these two investigations are then followed by a short discussion of their results and also possible improvements to the model.

5.3.1 Localisation of reverberant loudspeaker signals

The majority of the stimuli used in the first listening test were generated by simulating the microphone capture of a sound-field. Each stimulus created in this way is uniquely defined by six variables. The first variable is the original signal, which in the first listening test consisted of either a sine tone, pink noise or an anechoic recording of music or speech. The second variable is the source location, i.e. the location in the modelled recording environment where the original signal is emitted. The third variable is the directivity pattern of the modelled sound source. In the first listening test the source was always chosen to be omnidirectional. The fourth variable is the microphone configuration, comprising the location, orientation, directivity pattern and frequency response of each modelled microphone. In the first listening test these were either a single omnidirectional microphone or the ORTF or Williams configurations (see Section 4.2.2). The fifth variable is the set of impulse responses from the source to the modelled microphones. Two sets of impulse responses were used in the creation of the stimuli for the first listening test: one set corresponded to an anechoic recording environment and the other set was generated using a model of a reflective room with the CATT-Acoustics software. The sixth variable is the loudspeaker configuration used to play back the recorded signals. In the first listening test these were either a mono loudspeaker or the TCS or FCS reproduction systems.

Consider the set of all possible stimuli generated using this method. Each stimulus in this set corresponds to a unique combination of original signal, source location, source directivity pattern, microphone configuration, loudspeaker configuration and impulse responses. In this section the term signal family will refer to subsets of this set of stimuli such that all the different variables are fixed except for the impulse responses. Consequently, each member of a given signal family is uniquely defined by a set of impulse responses. All the signal families discussed in this section also satisfy the following constraints. The first constraint is that the directivity pattern of the source is always omnidirectional, the second constraint is that the reproduction system is
either TCS or FCS and the final constraint is that the ORTF and Williams microphone configurations are always used for TCS and FCS respectively. If all three constraints are satisfied then each signal family is uniquely defined by a combination of original signal, reproduction system and source location.

The stimuli used in the first listening test were designed so that each stimulus that was created using a reverberant room model had a corresponding anechoic stimulus. All the other aspects of the stimuli were identical for each of these echoic/anechoic stimulus pairs; they used the same original signals, the same position of the source and the same microphone technique was modelled. Therefore the stimuli in each of these pairs both belong to the same signal family. Table 5.6 shows these stimulus pairs and the corresponding signal families. Some of the stimulus pairs were not strictly identical, but instead the location of the original source of each stimulus in the pair was the mirror image of the location of the source of the other. As both the room modelled for the microphone capture and the listening room had left-right symmetry (i.e. were symmetrical in the $y = 0$ plane, see Section 3.4), the results from these stimulus pairs can be still be compared by taking the absolute value of each angle. This does assume the listeners have no left/right bias. However, only the centre listening position can be used to compare the results of the stimulus pairs where the echoic and anechoic stimuli are mirror images of each other: this is because the left and right listening positions are not mirror images of each other.

The listening test results for each of these stimulus pairs are shown in Fig. 5.31. From this figure it can be seen that the stimuli pairs played through TCS exhibit different behaviour from the stimulus pairs played through FCS. The echoic stimuli played through TCS remain localised close to $0^\circ$ as the angle of the corresponding anechoic stimuli increases. In comparison, the localisation angles of the echoic stimuli played through FCS do tend to increase as the angles of the corresponding anechoic stimuli increase, albeit at a slower rate. Fig. 5.32 shows the location azimuths calculated by the model for the same set of stimulus pairs. Comparing this to Fig. 5.31 it can be seen that the angles for the anechoic stimuli are generally greater than those for the echoic stimuli for both the listening test and the model results. However, only three of the twenty-one anechoic stimuli have angles from the listening test equal to zero degrees. In comparison, when the model was used to predict the localisation azimuths for the same twenty-one anechoic stimuli, sixteen of the results were zero degrees.

Methodology

The set of stimulus pairs from the first listening test uses only two levels of reverberation: each stimulus pair consists of an anechoic stimulus and a reverberant stimulus using the set of impulse responses calculated using the CATT-Acoustics software. The room model used in CATT-Acoustics to generate the set of impulse responses had a reverberation time of 0.85 seconds. In order to investigate how the localisation model behaved with a range of reverberant loudspeaker signals, a set of impulse responses were created using the ODEON acoustics modelling software and a range of room models. The ten room models used all had identical dimensions, namely a width of 7.35m, a depth of 5.68m and a height of 2.49m, as shown in Fig. 5.33. These dimensions match those of the listening room at the University of Surrey, which meets the ITU-R BS.1116-1 recommendations [131]. Each of the ten room models had a different material assigned to the surfaces in
Table 5.6: Pairs of stimuli from the first listening test that were identical except for the reflectivity of the room model used when simulating the microphone capture of the original signals. The numbers for the echoic and anechoic stimuli refer to Table 4.10.

<table>
<thead>
<tr>
<th>Signal family</th>
<th>Stimulus pair Original signal</th>
<th>System</th>
<th>Distance (metres)</th>
<th>Angle (degrees)</th>
<th>Reflected in x = 0 plane</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>15 Anechoic 18 Echoic Pink noise</td>
<td>TCS</td>
<td>2.2</td>
<td>30</td>
<td>✓</td>
</tr>
<tr>
<td>B</td>
<td>16 Anechoic 19 Echoic Pink noise</td>
<td>FCS</td>
<td>3</td>
<td>-70</td>
<td>✓</td>
</tr>
<tr>
<td>C</td>
<td>17 Anechoic 20 Echoic Pink noise</td>
<td>FCS</td>
<td>2</td>
<td>-15</td>
<td>x</td>
</tr>
<tr>
<td>D</td>
<td>25 Anechoic 28 Echoic Cello</td>
<td>TCS</td>
<td>2.2</td>
<td>30</td>
<td>✓</td>
</tr>
<tr>
<td>E</td>
<td>26 Anechoic 29 Echoic Classical guitar</td>
<td>TCS</td>
<td>2</td>
<td>-15</td>
<td>x</td>
</tr>
<tr>
<td>F</td>
<td>27 Anechoic 30 Echoic African percussion</td>
<td>FCS</td>
<td>3</td>
<td>60</td>
<td>x</td>
</tr>
<tr>
<td>G</td>
<td>35 Anechoic 38 Echoic Female speech</td>
<td>FCS</td>
<td>3</td>
<td>-45</td>
<td>x</td>
</tr>
<tr>
<td>H</td>
<td>36 Anechoic 39 Echoic Male speech</td>
<td>TCS</td>
<td>3</td>
<td>-45</td>
<td>x</td>
</tr>
<tr>
<td>I</td>
<td>37 Anechoic 40 Echoic Female speech</td>
<td>FCS</td>
<td>3</td>
<td>60</td>
<td>x</td>
</tr>
</tbody>
</table>

Figure 5.31: Anechoic and echoic listening test localisation results. Comparison of the results from the first listening test for the pairs of stimuli that were identical except for the reflectivity of the room model used when simulating the microphone capture of the original signals. The plotted circles show the median localisation azimuths for the anechoic stimuli and the crosses show the median localisation azimuths for the echoic stimuli. The first line below the plots contains the name of the signal family, which corresponds to a pair of stimuli from the first listening test (see Table 5.6). The second line below the plots shows whether the results are from the centre (C), left (L) or right (R) listening positions. The bottom line shows whether the stimuli were played back through TCS (T) or FCS (F).
Figure 5.32: Anechoic and echoic localisation results from the model. Comparison of the model results for the pairs of stimuli from the first listening test that were identical except for the reflectivity of the room model used when simulating the microphone capture of the original signals. The plotted circles show the localisation azimuths calculated by the model for the anechoic stimuli and the crosses show the azimuths from the model for the echoic stimuli. The table below the plots is interpreted as in Fig. 5.31 and the order of the stimuli is identical to that in Fig. 5.31.

Figure 5.33: Plot showing the dimensions of the room models used in ODEON.
the room, with each material having different absorption characteristics. The details of these materials are shown in Table 5.7.

Each of these ten room models was used in ODEON to create a set of impulse responses. A total of ninety new stimuli were then created, each new stimulus corresponding to a combination of one of the ten sets of impulse responses and one of the nine signal families given in Table 5.6. Localisation azimuths were then calculated for all ninety new stimuli using the model.

Results

Table 5.8 shows the absolute values of the angles predicted by the modified Supper model for the ninety new stimuli. Fig. 5.34 shows the localisation results output by the model for the ten new stimuli that are members of signal family II (see Table 5.6) and the centre listening position. This figure shows that as the room models become more reverberant then the angles calculated by the model tend towards zero degrees, as can also be seen in Table 5.8. The results for signal family II shown in the figure are typical of all the signal families in Table 5.6, and all the signal families show this trend of the angles calculated by the model tending towards zero degrees as the reverberation of the room models increase. Note the abrupt change in the calculated angles where the absolute angles for materials 60 and 50 are 11° and 0° respectively. This is also typical of the results in Table 5.8.

Fig. 5.35 shows the IID and ITD results from the model for the new stimuli corresponding to material 50 and the signal families given in Table 5.6. This figure shows that twenty of the twenty-one stimuli resulted in an angle of 0° being calculated by the model from the IID cues. This should be compared with only five of the twenty-one stimuli having predicted locations at 0° when the model uses just the ITD cues. Fig. 5.36 shows the results for the same stimuli when the IID and ITD cues are combined by taking the product, as described in Section 5.1.6. A comparison of Figs. 5.35 and 5.36 shows that the IID cues tend to dominate the combined results. It is interesting to note that only one of the stimuli using FCS was not localised at 0°, and this stimulus was localised at 1° (see Table 5.8. In contrast, only one of the stimuli using TCS was localised at 0°. The presence of recorded reflections in the loudspeaker signals clearly makes the IID and ITD cues more difficult for the modified Supper model to use to localise the sound, and it may be that because each FCS stimulus has five loudspeakers active simultaneously that these cues become even less coherent. Each TCS stimulus has only two loudspeakers active simultaneously, which may contribute to the fact that fewer of these stimuli are localised at 0°, as can be seen in Fig. 5.36.

5.3.2 Localisation in a reverberant reproduction environment

All the directional localisations results presented in this thesis up to this section have used the interpolation of the Gardner-Martin IIRTF database (see Section 3.3.1 and [53]) to generate the binaural signals which were input into the Supper model. This assumes that the listening environment is either anechoic or else at least sufficiently non-reflective so that it can be approximated by an anechoic environment without signifi-
<table>
<thead>
<tr>
<th>Material ID</th>
<th>Description in ODEON</th>
<th>RT60 (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10% absorbent</td>
<td>0.92</td>
</tr>
<tr>
<td>20</td>
<td>20% absorbent</td>
<td>0.52</td>
</tr>
<tr>
<td>30</td>
<td>30% absorbent</td>
<td>0.34</td>
</tr>
<tr>
<td>40</td>
<td>40% absorbent</td>
<td>0.25</td>
</tr>
<tr>
<td>50</td>
<td>50% absorbent</td>
<td>0.19</td>
</tr>
<tr>
<td>60</td>
<td>60% absorbent</td>
<td>0.15</td>
</tr>
<tr>
<td>70</td>
<td>70% absorbent</td>
<td>0.12</td>
</tr>
<tr>
<td>80</td>
<td>80% absorbent</td>
<td>0.09</td>
</tr>
<tr>
<td>90</td>
<td>90% absorbent</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>100% absorbent</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 5.7: Details of the materials used in the room model in Odeon to alter the reverberation time.

Figure 5.34: Localisation azimuths output by the model for signal family II (see Table 5.6) and impulse responses calculated in ODEON for the range of materials given in Table 5.7. The plotted circles are the RT60 reverberation times for the ODEON room models corresponding to the different materials. The plotted crosses show the magnitude of the angles calculated by the model corresponding to the different materials. The centre listening position was used for all the model results in the figure.
<table>
<thead>
<tr>
<th>Reproduction system</th>
<th>Signal family</th>
<th>Listener position</th>
<th>Absolute angles (degrees)</th>
<th>Absorbency of surfaces (%)</th>
<th>1st exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>TCS</td>
<td>E</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>FCS</td>
<td>G</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>C</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>L</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.8: The absolute values of the angles predicted by the modified Supper model for the nine signal families and the different ODEON room models. Details of the materials used in the different ODEON models are given in Table 5.7. The first column contains the reproduction system (TCS or FCS), the second column contains the name of the signal family (this corresponds to a pair of stimuli from the first listening test: see Table 5.6) and the third column shows whether the centre (C), left (L) or right (R) listening position was used. The last two columns contain the absolute values of the mean angles from the first listening test for the anechoic and echoic stimuli from each signal family.
Figure 5.35: Comparison of the IID and ITD model results for signals in the nine signal families described in Table 5.6 generated using impulse responses calculated from the ODEON room model with material 50 (which is 50% absorbent). The plotted circles are the localisation azimuths calculated by the model using just the IIDs and the plotted crosses are the azimuths calculated by the model using only ITDs. The first line below the plots contains the name of the signal family, which corresponds to a pair of stimuli from the first listening test (see Table 5.6). The third line below the plots shows whether the centre (C), left (L) or right (R) listening positions were used. The bottom line shows whether the stimuli were played back through TCS (T) or FCS (F). The order of the results is the same as in Fig. 5.31.

Figure 5.36: The combined IID and ITD model results for signals in the nine signal families described in Table 5.6 generated using impulse responses calculated from the ODEON room model with material number 50 (which is 50% absorbent). The table below the plots is interpreted as in Fig. 5.35.
Figure 5.37: The RT60 reverberation times plotted against frequency for the ten rooms modelled in CATT, each room model created using a different absorption coefficient for the room surfaces. The horizontal black dotted line at 2.55s shows the ideal RT60 value for the room of these dimensions as recommended by ITU-R BS.1116-1 [131], with the solid black lines showing the boundaries of acceptable RT60 values.

significantly affecting the results. This section investigates the effects of a reflective listening environment on the performance of the modified Supper model.

Methodology

It was discussed in Section 3.3.1 how acoustics modelling software such as CATT-Acoustics or ODEON could be used to generate HRTFs for more reflective environments, and this was the method used in this investigation. Ten different room models were created in CATT-Acoustics. Each of these had the same dimensions as the listening room at the University of Surrey (see Fig. 5.33), which meets the ITU-R BS.1116-1 recommendations [131]. As in the previous section, each of the ten room models had a different material assigned to the surfaces in the room, with each material having different absorption characteristics. Fig. 5.37 shows the RT60 reverberation times plotted against frequency for each of the ten models. The loudspeaker positions and listener positions from the third listening test were then added to each of the ten models. CATT-Acoustics was then used to create HRTFs for each combination of loudspeaker and listener position in each room model, giving a total of $3 \times 8 \times 10 = 240$ different HRTFs. The HRTFs corresponding to each room model were then used to generate binaural signals for each combination of listening position and stimulus from the third listening test experiment. Only the non-widened stimuli were considered, as it has already been seen in Section 5.2.3 that the modified Supper model is unable to localise the stimuli that had
been processed using the widening algorithm used for the third listening test experiment. This resulted in a total of \((10 \text{ room models}) \times (18 \text{ stimuli}) = 180\) different pairs of binaural signals, which were then input into the modified Supper model.

Results

All the localisation model results that have been presented before this section have been calculated using binaural signals generated using the method of interpolating the Gardner-Martin HRTF database [53] described in Section 3.3.1. These interpolated HRTFs all correspond to an anechoic listening environment. This section has described how binaural signals were created for a number of listening environments with different reflective properties using the CATT-Acoustics software. One of the listening environments modelled in CATT-Acoustics has room surfaces with 100% absorption, i.e. the room is anechoic. Table 5.9 shows a comparison of using these two different methods of generating the binaural signals in an anechoic environment together with the modified Supper model to predict the localisation results from the third listening test experiment. The artificially widened stimuli were omitted, as it was seen in Section 5.2.3 that the model was not able to predict the results for these stimuli. This table shows that the two methods of generating the binaural signals give comparable results, with the interpolated Gardner-Martin HRTFs giving a lower root mean square error of prediction. This slightly better performance may be because the Gardner-Martin HRTF database was also used to train the look-up table in the modified Supper model (see Section 5.1.3).

Table 5.10 contains a summary of using the model with the ten different listening environments modelled in CATT-Acoustics to predict the intended angles of the stimuli from the third listening test experiment. The artificially widened stimuli were omitted and the intended angles for the panned stimuli were calculated using the equidistant method described in Section 4.4.5. The results in this table show that, as the modelled listening environment becomes more reflective, the model no longer localises the stimuli at their intended locations. It is interesting to note that the results for the room with 90% absorbent surfaces actually predicted the intended locations better than the anechoic (100% absorbent) room.

Figs. 5.38 and 5.39 show the model results for four of the different listening environments plotted against the intended angles. Again, the artificially widened stimuli were omitted and the intended angles for the panned stimuli were calculated using the equidistant method. The absorbency of the surfaces in the four room models are: 10% (left hand plot, Fig. 5.38), 40% (right hand plot, Fig. 5.38), 70% (left hand plot, Fig. 5.39) and 100% (right hand plot, Fig. 5.39). From these plots it can be seen that the modified Supper model tends to predict more and more of the stimuli at 0° as the listening room becomes more reflective. This is very similar to the results of the investigation into the effect of reverberant loudspeaker signals in Section 5.3.1, where the angles calculated by the model tended to 0° as the room models used to generate the loudspeaker signals became more reflective.

Finally, note from Fig. 5.37 that the room model with surfaces of 40% absorbency has RT60 values that meet the listening room recommendations in ITU-R BS.1116-1 [131] for almost all frequencies. As the listening room at the Institute of Sound Recording at the University of Surrey also meets these recommendations,
### Table 5.9: Comparison of using the two methods of calculating the binaural signals in an anechoic environment with the modified Supper model to predict the localisation results for the third listening test experiment.

<table>
<thead>
<tr>
<th>Description</th>
<th>R value</th>
<th>RMSEP</th>
<th>Gradient of line of best fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpolation of Gardner-Martin database [53]</td>
<td>0.99</td>
<td>3.83</td>
<td>1.20</td>
</tr>
<tr>
<td>Anechoic room model in Catt-Acoustics</td>
<td>0.99</td>
<td>8.31</td>
<td>1.19</td>
</tr>
</tbody>
</table>

### Table 5.10: The results of fitting the localisations from the model for differently reverberant reproduction environments to the intended angles for the third listening test. The artificially widened stimuli were omitted. The intended angles for the panned stimuli were calculated using the equidistant method described in Section 4.4.5.

<table>
<thead>
<tr>
<th>Absorbency of room surfaces</th>
<th>RT60 (seconds)</th>
<th>R value</th>
<th>RMSEP degrees</th>
<th>Gradient of line of best fit</th>
<th>Percentage localised at 0°</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.94</td>
<td>0.09</td>
<td>32.55</td>
<td>0.00</td>
<td>94%</td>
</tr>
<tr>
<td>20</td>
<td>0.48</td>
<td>0.21</td>
<td>32.18</td>
<td>0.01</td>
<td>91%</td>
</tr>
<tr>
<td>30</td>
<td>0.31</td>
<td>0.58</td>
<td>30.04</td>
<td>0.09</td>
<td>76%</td>
</tr>
<tr>
<td>40</td>
<td>0.22</td>
<td>0.53</td>
<td>28.61</td>
<td>0.17</td>
<td>54%</td>
</tr>
<tr>
<td>50</td>
<td>0.18</td>
<td>0.62</td>
<td>27.04</td>
<td>0.22</td>
<td>46%</td>
</tr>
<tr>
<td>60</td>
<td>0.14</td>
<td>0.75</td>
<td>22.25</td>
<td>0.46</td>
<td>33%</td>
</tr>
<tr>
<td>70</td>
<td>0.11</td>
<td>0.86</td>
<td>17.30</td>
<td>0.71</td>
<td>20%</td>
</tr>
<tr>
<td>80</td>
<td>0.09</td>
<td>0.93</td>
<td>12.21</td>
<td>0.95</td>
<td>17%</td>
</tr>
<tr>
<td>90</td>
<td>0.07</td>
<td>0.99</td>
<td>5.89</td>
<td>1.12</td>
<td>13%</td>
</tr>
<tr>
<td>100</td>
<td>0.02</td>
<td>0.99</td>
<td>7.25</td>
<td>1.18</td>
<td>9%</td>
</tr>
</tbody>
</table>
Figure 5.38: These two plots show the localisation results from the model using differently reverberant reproduction environments plotted against the intended angles for the third listening test. The surfaces of the reproduction room were modelled as 10% and 40% absorbent for the left and right plots respectively. The artificially widened stimuli were omitted. The intended angles for the panned stimuli were calculated using the equidistant method described in Section 4.4.5.

Figure 5.39: These two plots show the localisation results from the model using differently reverberant reproduction environments plotted against the intended angles for the third listening test. The surfaces of the reproduction room were modelled as 70% and 100% absorbent for the left and right plots respectively. The artificially widened stimuli were omitted. The intended angles for the panned stimuli were calculated using the equidistant method described in Section 4.4.5.
the 40% absorbency room model should be similar to the listening room in which the three listening test experiments were performed. However, the right hand plot in Fig. 5.38 shows that the modified Supper model using this room model for the listening environment is not a very good predictor of the intended angles for the third listening test experiment. This is particularly evident when compared to the case with the anechoic room model, as shown in the right hand plot of Fig. 5.39. It should also be compared to the localisation results from the third listening test experiment, which were much closer to the intended angles (see Fig. 4.29).

5.3.3 Discussion

The investigations in Sections 5.3.1 and 5.3.2 both introduced aspects of reflective acoustical environments into the binaural signals which were then input into the modified Supper model. In both sections, the investigations used software to create models of reflective rooms which were then used to create impulse responses. The difference between the two investigations is that in Section 5.3.1 the reflective room models were used in the modelling and capture of the original sound-field (processes I and II in Fig. 3.1 in Section 3.1), while in Section 5.3.2 the reflective room models were used in the modelling of the reproduction system and the generation of the binaural signals (processes VI and VII in Fig. 3.1).

In both investigations the behaviour of the modified Supper model was similar: as the room models became more reflective, so the angles output by the modified Supper model tended towards 0°. This behaviour is also seen in real listeners: Fig. 5.31 shows the difference between the listeners’ localisation responses to the anechoic and reflective stimuli in the first listening test. However, a comparison of Figs. 5.31 and 5.32 shows that the behaviour of the model is more exaggerated: the listeners’ responses moved only part of the way towards 0° for the reflective stimuli, while the majority of the reflective stimuli were localised actually at 0° by the model.

The first listening test experiment was the only experiment to include reflective stimuli, and, as shown in Section 5.2 and in Figs. 5.31 and 5.32, the modified Supper model was not able to predict the listeners’ localisation responses to these stimuli. One way in which the performance of the modified SUpper model may be improved with respect to the reflective signals is to incorporate the onset detector from Supper’s original localisation model [158]. Supper defines an auditory onset as any region of time where the direct sound is the dominant part of the sound-field, before the arrival of the reflected energy. This is an extension of the usual definition of an auditory onset as the period of time when a new auditory event begins, with a steep and significant increase in sound energy [10]. Supper’s motivation for creating an onset detector for use with his localisation model was that the IIDs and ITDs cues for localisation will be most unambiguous when the least amount of reflections are present in the binaural signals. Indeed, this section has demonstrated that the model does not perform well with signals which include reflections. As the direct sound dominates the periods of time immediately following an auditory onset, detecting these onsets will allow the localisation model to concentrate on these regions of the binaural signals, leading to more robust localisation. From this it can be seen that incorporating Supper’s onset detector into the modified Supper model may directly help with the issues involving reflective signals described in this section.
One possible method of improving the performance of the modified Supper model for a specific reflective reproduction environment is to use HRTFs measured in the environment to populate the IID and ITD look-up tables used in the localisation algorithm (see Section 5.1.3). This could be considered analogous to human listeners learning and remembering the reflective properties of a given room or hall, which are then recalled upon revisiting the space.

5.3.4 Summary

This section has described two investigations into the effects of reflective signals on the performance of the modified Supper model. The first investigation was concerned with loudspeaker signals created by modelling microphone techniques in reflective environments. The second investigation was concerned with modelling different reflective listening environments. In both investigations the angles output by the modified Supper model tended towards 0° as the modelled rooms became more reflective. Inspection of the results from the first listening test experiment showed that the model’s behaviour with reflective signals appeared to be an exaggerated version of the behaviour of human listeners. This was followed by a discussion of possible changes to the modified Supper model which may improve its performance when presented with reflective signals.
5.4 How different reproduction systems and formats affect localisation

Section 5.2 showed the validation of the localisation model, i.e. the modified Supper model together with the framework described in Chapter 3 is able to predict how listeners perceive the localisation of reproduced sound sources, at least in anechoic and near anechoic conditions. Consequently, the localisation model can also be used as a tool to assess how well different audio reproduction systems can accurately reproduce localised sound sources for a listener positioned at the sweet spot, as described in this section.

5.4.1 Methodology

Four different loudspeaker layouts for the audio reproduction systems were considered: corresponding to mono, TCS, FCS and 32 channel WFS. All the loudspeaker layouts had a radius of 2.2m. Ten different intended angles were considered, from 0° (directly in front of the sweet spot) to 90° (to the right of the sweet spot) in 10° increments. Only angles in the front right quadrant were considered as the model has left/right symmetry. The signals were created by modelling a source located at a distance of 3m from the sweet spot and at the intended angle. The capture of the resulting sound field was then modelled in an anechoic environment using a microphone array. The six signals used as the sources in the original sound field were the pink noise and anechoic recordings of African percussion, cello, classical guitar and male and female speech described in Table 4.1 in Section 4.1. The microphone techniques used were a single omnidirectional microphone at the sweet spot for the mono reproduction system, ORTF [157] for TCS, the Williams-Dü array [167] for FCS and a circular array of 32 cardioid microphones for WFS (see Section 3.5.2). In addition to these signals, the constant power panning law was used to create signals at the ten intended angles for the six original signals for TCS and FCS. Note that when using TCS the range of angle to which a source can be panned is limited to ±30°. Consequently, for those intended angles that fell outside this range (i.e. were greater than 30°), the original signals were panned as far right as possible, i.e. the signals were panned to the right hand loudspeaker.

Six reproduction systems (each a combination of a loudspeaker layout and a method of generating the signals to the loudspeakers) were used, together with six original signals and ten different intended angles, giving a total of 6 x 6 = 360 stimuli. These stimuli were input through the model, using an anechoic reproduction environment and the listener situated at the sweet spot facing forwards. Finally, the model was also used to predict how a listener would localise the original sound fields (i.e. the original signals modelled as sources at a distance of 3m and at the intended angle in an anechoic environment). This resulted in a further 6 x 10 = 60 predicted angles, giving a total of 420 angles predicted by the model. These results were then recalculated with the listener still located at the sweet spot but facing 45° to the right.
5.4.2 Results

Fig. 5.40 shows the predicted localisations of the pink noise stimuli for the six different reproduction systems. Appendix G shows the same results plotted on plans of the listening area in order to illustrate the localised source angles relative to the listener and the loudspeakers. Table 5.11 shows the errors, i.e. the differences between the intended angle and the predicted angle for each stimulus. The listener was facing forwards at the sweet spot for the results in this table, Fig. 5.40 and the plots in Appendix G. From Table 5.11 it can be seen that for each reproduction system the six original signals were mostly localised at similar locations by the model. This is especially true of the Mono reproduction system, where all the signals were localised at 0° regardless of the intended angle of the original signal. The stimuli created from the anechoic recording of African percussion do appear to have larger errors compared to the stimuli created from the other five original signals. This can be seen especially with the original sound field results, where the African percussion stimuli at angles of 70° and 80° were both localised close to 0°. Fig. 5.40 shows that the signal panned to 30° (i.e. the right hand loudspeaker) was actually localised outside the angle spanned by the loudspeakers. This is likely to be an artifact of the model: Section 5.2 showed that the line of best fit passing though the origin had a gradient greater than one. Consequently, the model tends on average to predict the localisation angles of sound sources closer to ±90° than real listeners. Furthermore, the magnitude of this error increases as the intended angle tends towards ±90°.

Other artifacts of the model are the angles predicted at 0° when the intended angle is close to ±90°. Although it is possible that this accurately represents the behaviour of the reproduction system, two factors suggest that these are more likely to be artifacts of the model. The first of these is that the original sound field consisting of the African percussion at an actual angle of 80° was localised by the model at 0°, which cannot be attributed to the limitations of the reproduction system, as no reproduction system was involved. The second factor suggesting that some of the angles predicted at 0° are artifacts of the model is that this behaviour was exhibited in the original, unmodified Supper model (see Section 5.1.4). So, although steps have been taken to reduce this behaviour, cases of the model returning spurious results of 0° have a precedent. Similarly, Section 5.3 shows that when the binaural signals include reflected signals at some point in the recording chain then the model has difficulty localising the sources and returns an angle of 0°. This suggests that a predicted angle of 0° actually has two possible causes. The first cause is the one intended in the design of the model, i.e. the source is well localised by the model, and here the results are representative of where real listeners would also localise the sound. However, there appears to be an alternative set of circumstances that also give rise to a predictive angle of 0°, namely that the model has difficulty localising the sound source. This behaviour is consistent with the results in Section 5.3, where the angles output by the modified Supper model tended to 0° as the room models became more reflective and the IID and ITD cues became more ambiguous.
Figure 5.40: Summary of the perceived localisation angles calculated by the model for intended angles in the range 0° to 90° for the six different reproduction systems and the original sound-field. These results were calculated with the listener situated at the sweet spot facing forward (0°).

Figure 5.41: Summary of the perceived localisation angles calculated by the model for intended angles in the range 0° to 90° for the six different reproduction systems and the original sound-field. These results were calculated with the listener situated at the sweet spot facing 45° to the right.
<table>
<thead>
<tr>
<th>Reproduction system</th>
<th>Original signal</th>
<th>Angle of source (degrees)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 10 20 30 40 50 60 70 80 90</td>
<td>A S</td>
</tr>
<tr>
<td><strong>Mono</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>Cello</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>Female speech</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>Classical guitar</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>Male speech</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>Pink noise</td>
<td>0 -10 -20 -30 -40 -50 -60 -70 -80 -90</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td><strong>TCS (mics)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>0 -7 -11 -19 -25 -32 -39 -40 -48 -55</td>
<td>27.6</td>
<td>27.6</td>
</tr>
<tr>
<td>Female speech</td>
<td>0 -6 -10 -15 -23 -28 -45 -39 -47 -56</td>
<td>26.4</td>
<td>26.4</td>
</tr>
<tr>
<td>Male speech</td>
<td>0 -7 -11 -16 -26 -27 -33 -40 -48 -56</td>
<td>26.4</td>
<td>26.4</td>
</tr>
<tr>
<td>Pink noise</td>
<td>0 -7 -11 -27 -25 -27 -34 -40 -48 -56</td>
<td>27.5</td>
<td>27.5</td>
</tr>
<tr>
<td><strong>TCS (pan)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>0 1 1 4 -6 -16 -26 -36 -46 -56</td>
<td>19.2</td>
<td>18.0</td>
</tr>
<tr>
<td>Cello</td>
<td>0 1 3 6 -4 -14 24 -34 -44 -54</td>
<td>18.4</td>
<td>16.4</td>
</tr>
<tr>
<td>Female speech</td>
<td>0 1 3 7 -3 -13 -23 -33 -43 -53</td>
<td>17.9</td>
<td>15.7</td>
</tr>
<tr>
<td>Classical guitar</td>
<td>0 2 3 7 -3 -13 -23 -33 -43 -53</td>
<td>18.0</td>
<td>15.6</td>
</tr>
<tr>
<td>Male speech</td>
<td>0 1 3 6 -4 -14 -24 -34 -44 -54</td>
<td>18.4</td>
<td>16.4</td>
</tr>
<tr>
<td>Pink noise</td>
<td>0 1 3 7 -3 -13 -23 -33 -43 -53</td>
<td>17.9</td>
<td>15.7</td>
</tr>
<tr>
<td><strong>FCS (mics)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>0 -6 -10 -15 -24 -29 -45 -38 -42 -46</td>
<td>25.5</td>
<td>25.5</td>
</tr>
<tr>
<td>Cello</td>
<td>0 -5 -8 -16 -24 -34 -48 -30 -28 -90</td>
<td>28.5</td>
<td>28.5</td>
</tr>
<tr>
<td>Female speech</td>
<td>0 -5 -10 -16 -24 -30 -35 -32 -28 -99</td>
<td>26.9</td>
<td>26.9</td>
</tr>
<tr>
<td>Classical guitar</td>
<td>0 -5 -10 -16 -24 -29 -47 -32 -30 -89</td>
<td>28.2</td>
<td>28.2</td>
</tr>
<tr>
<td>Male speech</td>
<td>0 -5 -8 -15 -24 -30 -59 -37 -26 -90</td>
<td>29.4</td>
<td>29.4</td>
</tr>
<tr>
<td>Pink noise</td>
<td>0 -5 -10 -15 -24 -31 -35 -40 -31 -89</td>
<td>28.0</td>
<td>28.0</td>
</tr>
<tr>
<td><strong>FCS (pan)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African percussion</td>
<td>0 2 3 4 3 -3 -9 -14 -19 -21</td>
<td>7.8</td>
<td>5.4</td>
</tr>
<tr>
<td>Cello</td>
<td>0 2 3 6 2 -3 -8 -15 -19 -23</td>
<td>8.1</td>
<td>5.5</td>
</tr>
<tr>
<td>Female speech</td>
<td>0 2 3 7 2 -3 -9 -13 -18 -21</td>
<td>7.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Classical guitar</td>
<td>0 2 3 7 3 -3 -7 -13 -19 -21</td>
<td>7.8</td>
<td>4.8</td>
</tr>
<tr>
<td>Male speech</td>
<td>0 2 3 6 2 -3 -9 -14 -19 -21</td>
<td>7.9</td>
<td>5.3</td>
</tr>
<tr>
<td>Pink noise</td>
<td>0 2 3 7 3 -3 -8 -13 -19 -21</td>
<td>7.9</td>
<td>4.9</td>
</tr>
</tbody>
</table>

Table 5.11: (Table continued on next page)
<table>
<thead>
<tr>
<th>Reproduction system</th>
<th>Original signal</th>
<th>Angle of source (degrees)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>WFS</td>
<td>African percussion</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Original sound field</td>
<td>African percussion</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5.11: The differences between the intended angle and the predicted angle for the reproduced sound sources. The listener was situated at the sweet spot facing forward (0°). All the angles are in degrees. Errors with a negative value show the predicted angle was to the left of the intended angle. All the sources were intended to be located in the front right quadrant, so the negative errors show that the predicted angle was close to 0° than the intended angle (no sources were localised by the model in the front left quadrant). Similarly, positive errors show the predicted angle was to the right of the intended angle (i.e. closer to 90° than the intended angle). The two columns on the far right show the mean errors calculated for each combination of reproduction system and original signal calculated across the ten different intended angles. The means in the A column were calculated using the absolute values of the errors, while the means in the S column were calculated using the signed values of the errors. (Table continued from previous page)
<table>
<thead>
<tr>
<th>Reproduction system</th>
<th>Original signal</th>
<th>Angle of source (degrees)</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0  10  20  30  40  50  60  70  80  90</td>
<td>A  S</td>
</tr>
<tr>
<td>TCS (mics)</td>
<td>African percussion</td>
<td>-1  -8  -13  -19  -9  -7  -27  -39  -50  -61</td>
<td>23.4  -23.4</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>-1  -8  -13  -19  -5  -6  -29  -39  -50  -61</td>
<td>23.1  -22.1</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>-2  -8  -13  -19  5  -6  -28  -39  -50  -62</td>
<td>23.2  -22.2</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>-1  -10  -13  -19  4  -7  -30  -39  -49  -61</td>
<td>23.3  -22.5</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>-1  -8  -13  -19  4  -6  -28  -39  -50  -62</td>
<td>23.0  -22.2</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>-1  -8  -13  -18  5  -6  -28  -39  -50  -61</td>
<td>22.9  -21.9</td>
</tr>
<tr>
<td>TCS (pan)</td>
<td>African percussion</td>
<td>-1  -1  -3  -13  -23  -33  -43  -53  -63</td>
<td>23.6  -23.6</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>-1  20  10  -3  -13  -23  -33  -43  -53  -63</td>
<td>26.2  -20.2</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>-2  -2  -3  -13  -23  -33  -43  -53  -63</td>
<td>23.8  -23.8</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>0  -2  -4  -13  -23  -33  -43  -53  -63</td>
<td>23.7  -23.7</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>-1  -2  -2  -12  -22  -32  -42  -52  -62</td>
<td>22.9  -22.9</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>-1  -3  10  -3  -13  -23  -33  -43  -53  -63</td>
<td>24.5  -22.5</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>-7  -11  -20  -26  -34  -37  -39  -38  -33  -31</td>
<td>27.6  -27.6</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>-7  -11  -17  -25  -34  -28  -40  -39  -34  -45</td>
<td>28.0  -28.0</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>-8  -11  -19  -25  -34  -37  -33  -38  -36  -43</td>
<td>28.4  -28.4</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>-7  -10  -17  -25  -34  -37  -33  -38  -34  -44</td>
<td>27.9  -27.9</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>-7  -9  -15  -25  -33  -37  -33  -37  -35  -44</td>
<td>27.5  -27.5</td>
</tr>
<tr>
<td>FCS (pan)</td>
<td>African percussion</td>
<td>-8  -7  -7  -2  -2  -2  -4  -4  0</td>
<td>3.9  -3.9</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>-9  -7  -6  -3  -2  -5  -1  -2  -2  -1</td>
<td>3.8  -3.8</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>-9  -5  -6  -3  -2  -5  -1  -2  -2  -1</td>
<td>3.6  -3.6</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>-9  -8  -6  -3  -2  -2  -1  -2  -2  -1</td>
<td>3.6  -3.6</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>-8  -7  -6  -2  -2  -5  -1  -2  -2  0</td>
<td>3.5  -3.5</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>-10  -7  -6  -3  -2  -5  -1  -2  -2  0</td>
<td>3.8  -3.8</td>
</tr>
</tbody>
</table>

Table 5.12: (Table continued on next page)
Table 5.12: The differences between the intended angle and the predicted angle for the reproduced sound sources. The listener was situated at the sweet spot facing 45° to the right. All the angles are in degrees. Errors with a negative value show the predicted angle was to the left of the intended angle. All the sources were intended to be located in the front right quadrant, so the negative errors show that the predicted angle was close to 0° than the intended angle (no sources were localised by the model in the front left quadrant). Similarly, positive errors show the predicted angle was to the right of the intended angle (i.e. closer to 90° than the intended angle). The two columns on the far right show the mean errors calculated for each combination of reproduction system and original signal calculated across the ten different intended angles. The means in the A column were calculated using the absolute values of the errors, while the means in the S column were calculated using the signed values of the errors. (Table continued from previous page)
In order to try to differentiate between the errors introduced by the reproduction systems and the errors introduced by the model, Table 5.12 shows the errors calculated when the listener was located at the sweet spot facing 45° to the right. Similarly, Fig. 5.41 shows the predicted localisations of the pink noise stimuli for the six different reproduction systems when the listener was located at the sweet spot facing 45° to the right. In this table the issue of sources being located at 0° (relative to the listener) has disappeared for the sources located close to 90° (relative to the listening environment). For instance, the original sound-field with African percussion located at 80° had an error of -80° (corresponding to an angle of 0° relative to both the listener and the environment) when the listener was facing 0°. When the listener orientation was rotated to 45°, the same stimulus (original sound-field with African percussion located at 80°) gave an error of 8°, corresponding to an angle of 88° relative to the environment and 43° relative to the listener. This shows that the results seen in Table 5.11 where the angles were predicted at 0° when the intended angle was close to 90° are indeed artifacts of the model rather than being due to the reproduction systems themselves.

From Tables 5.11 and 5.12 and Figs. 5.40 and 5.41 it can be seen that the mono reproduction system performed worst, followed by the TCS (mics) and FCS (mics) reproduction systems. From inspecting the last two columns in both tables it appears that FCS (pan) is the reproduction system best able to accurately place sources. However, these results are affected by the errors introduced by the model, as can be seen by the results for the original sound-field in Table 5.11. Table 5.12 does not have the issue of sources with intended angles close to 90° being spuriously located at 0°, and from this table it can be seen that the mean errors in the last two columns for the 32-channel WFS system are almost as small as the mean errors for the FCS (pan) reproduction system. Inspection of the error values for the individual source angles shows that the errors for the WFS system are actually a closer match to the error values for the original sound-field than the errors for FCS (pan). This shows that actually the WFS system positions the sources closer to their intended positions and that the errors for WFS seen in Table 5.12 are introduced by the model rather than the reproduction system. Indeed, it was seen in Section 5.2 that the line of best fit passing through the origin for the model predictions plotted against the listening test results had a gradient greater than one, which would lead to prediction errors like those seen in the WFS and original sound-field results in Table 5.12. The only reason that FCS pan appears to perform better at accurately placing sources is that the errors introduced by the reproduction system are to some extent cancelled out by the errors introduced by the localisation model.

5.4.3 Summary

This section has described an investigation using the modified Supper model to assess the ability of different audio reproduction systems to accurately reproduce placed sound sources for listeners at the sweet spot. The model was used to predict localisation angles for every combination of seven reproduction systems (including the original sound-field), six original signals, ten different source locations and two different listener orientations. The discussion of the results concluded that the 32-channel WFS system was the most accurate system at reproducing source locations. However, interpreting the results of the model predictions was not straightforward due to the systematic errors introduced by the model.
5.5 Localisation across the listening area

One of the advantages of the model framework described in Chapter 3 is that it is not limited to the centre listening position (the "sweet spot"). For audio reproduction systems in domestic environments the listener is often not situated at the ideal centre position, especially when there are multiple listeners in the listening environment. This means that it is desirable to obtain measures about spatial percepts not only at the sweet spot but across the listening area. This section describes how the model was used to obtain results across the listening area, followed by a number of these results and a discussion of the results.

5.5.1 Methodology

The listening area was sampled using a grid with a spacing of 10cm between adjacent points. The grid was limited to points within a radius of 1.5m from the centre of the listening area, as shown in Fig. 5.42, resulting in 709 different points. The grid was aligned to include a point at the centre of the listening area and also so that the axes corresponded to the x and y axes of the listening area (see Fig. 3.7).

Figure 5.42: The centre of each square in the grid corresponds to a point at which the listening area was sampled. The circular grid has a radius of 1.5m. An FCS loudspeaker array with a radius of 2.2m is also shown. This sampling grid was also used with the mono, TCS and WFS loudspeaker configurations.
The same reproduction systems (each a combination of a loudspeaker configuration and a method of generating the signals to the loudspeakers) were used as in Section 5.4: mono, TCS and FCS with constant power panning, TCS and FCS with the ORTF and Williams-Dû microphone techniques, and thirty-two channel WFS. All the loudspeaker configurations used a radius of 2.2m. The stimuli were all created for a source positioned at a distance of 3m and with a range of different angles. For each stimulus, the listener's head was modelled at each of the different points in the grid and the modified Supper model was used to calculate a directional localisation.

5.5.2 Results

Figs. 5.43 to 5.52 show plots of results from sampling across the listening area for pink noise sources located at \((-20^\circ, 3)\) and \((40^\circ, 3)\) for each of the different reproduction systems. The left hand plot in each figure has arrows showing the predicted angle of localisation at each point sampled across the listening area. The \textit{interp2} function in Matlab was used to interpolate the two dimensional array of angles calculated by the model (corresponding to a grid with 10cm spacing) by a factor of ten, resulting in a two dimensional array of angles corresponding to a grid with 1cm spacing. The interpolation was performed using cubic splines. The errors were then calculated for this interpolated array as the absolute value of the difference between the interpolated angles calculated by the model and the intended angle of the source from the corresponding position in the listening area. This interpolated array of errors was then used to determine the error contours which were then added to the left hand plots in Figs. 5.43 to 5.52. The right hand plot in each of these figures clarifies this by showing how the different bands of the error values are distributed across the listening area.

From these plots it can be seen that the reproduction systems involving the use of simulated microphone techniques are much worse at accurately positioning source images than the reproduction systems using either panning or WFS. The TCS (pan) reproduction system has both the \(40^\circ\) and \(-20^\circ\) source localised in front of the listeners close to \(0^\circ\) for most of the listening area (see Figs. 5.45 and 5.46). This is also the case for the FCS (pan) system (see Figs. 5.49 and 5.50).

The localisation and error patterns are similar for the two reproduction systems involving the use of constant power panning. For both systems the plots for the source at \((-20^\circ, 3)\) show good regions of localisation in a vertical band running down the middle of the plot and also on the left hand side of the plot (see Figs. 5.43 and 5.47). Similarly, in the plots for the source at \((40^\circ, 3)\) both systems show a good region of localisation running diagonally from top right to bottom left in the plots, and the left hand side of the listening area has better localisation than the right hand side (see Figs. 5.44 and 5.48). Note that, because of the limits of the TCS loudspeaker configuration, the signal for the source at \((40^\circ, 3)\) is panned only to the right hand loudspeaker positioned at \(30^\circ\).

The localisation and error patterns for the WFS system are noticeably different from those for the systems using panning. Fig. 5.51 shows the results for the WFS system with the original source at \((-20^\circ, 3)\). In this plot almost all of the listening area has errors less than \(10^\circ\), and a large proportion has errors less than \(2.5^\circ\).
Figure 5.43: TCS, constant power panning, 20° left. The localisation results from the modified Supper model for a one second burst of pink noise positioned 20° to the left using the constant power panning law and replayed over TCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.

Figure 5.44: TCS, constant power panning, 40° right. The localisation results from the modified Supper model for a one second burst of pink noise positioned as close to 40° to the right as possible using the constant power panning law and then replayed over TCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.
Figure 5.45: TCS, ORTF microphone array, $(-20^\circ,3)$ The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at $(-20^\circ,3)$ and recorded using an ORTF microphone array and then replayed over TCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.

Figure 5.46: TCS, ORTF microphone array, $(40^\circ,3)$ The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at $(40^\circ,3)$ and recorded using an ORTF microphone array and then replayed over TCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.
Figure 5.47: FCS, constant power panning, 20° left. The localisation results from the modified Supper model for a one second burst of pink noise positioned 20° to the left using the constant power panning law and replayed over FCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.

Figure 5.48: FCS, constant power panning, 40° right. The localisation results from the modified Supper model for a one second burst of pink noise positioned as close to 40° to the right as possible using the constant power panning law and then replayed over FCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.
Figure 5.49: **FCS, Williams microphone array, (-20°, 3)** The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at (-20°, 3) and recorded using a Williams microphone array and then replayed over FCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.

Figure 5.50: **FCS, Williams microphone array, (40°, 3)** The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at (40°, 3) and recorded using a Williams microphone array and then replayed over FCS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.
Figure 5.51: WFS, (-20°, 3) The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at (-20°, 3) and was recorded and replayed using 32 channel WFS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.

Figure 5.52: WFS, (40°, 3) The localisation results from the modified Supper model for a one second burst of pink noise where the original source was modelled positioned at (40°, 3) and was recorded and replayed using 32 channel WFS. The left hand plot shows the azimuths at the grid of points sampled across the listening area with error contours. The right hand plot shows the errors between the intended azimuth and the perceived azimuth predicted by the model.
Fig. 5.52 shows the results for the WFS system with the source at (40°, 3). Unlike the equivalent results for TCS (pan) and FCS (pan), here it is the right hand side of the listening area that has the smallest errors. Of particular note is the band with large errors seen as the lower of the two dark blue regions in the right hand plot in Fig. 5.52. From inspecting this region in the left hand plot of this figure it can be seen that the source is localised around 0° in this region. The fact that the angles of localisation within this region are so different from those in the surrounding area suggests that this is an artifact of the WFS system, especially considering that none of the other reproduction systems exhibit this behaviour. It is not altogether surprising that WFS has artifacts like this while the other systems do not. The sound-fields for reproducing single sources with the TCS (pan), TCS (mics) and FCS (pan) systems are a result of the interaction of the sound from only two real sound sources (the active loudspeakers). The WFS sound-field is the result of the sound 32 loudspeakers interacting, so the potential for loudness and phase anomalies at some points in the listening area is greatly increased. Indeed, it is known that WFS is unable to accurately reproduce the original sound-field above the spatial aliasing frequency [43, 116]. For a circular array of 32 equally spaced loudspeakers with a radius of 2.2m, such as that used for the WFS system, the distance between adjacent loudspeakers, \( \Delta_{\text{sample}} \), is 43cm. Therefore, from Equation 3.8 in Section 3.5.2, the spatial aliasing frequency of the 32 channel WFS system is

\[
 f_{\text{aliasing}} = \frac{c}{2\Delta_{\text{sample}}} \approx \frac{330}{2 \times 0.43} = 384\text{Hz} \tag{5.9}
\]

This means that the artifacts seen in Fig. 5.52 may be due to spatial aliasing, as the pink noise signal includes frequencies above the spatial aliasing frequency.

The original signals used in this investigation were not just limited to pink noise. Tables 5.13 and 5.14 summarise the results for the (-20°, 3) and (40°, 3) source positions for the six different original signals. Both these tables contain the percentages of the listening area where the error (i.e. the absolute difference between the intended and predicted angles) is below a given threshold. Five different threshold angles were used in both tables: 2.5°, 5°, 10°, 20° and 30°. In both tables it can be seen that for each reproduction system the results for the different original signals are similar.

Table 5.13 contains the results positioned at (-20°, 3). In this table the reproduction systems using constant power panning can be seen to perform better than the two reproduction systems using modelled microphone techniques. The model predicted 27% of the listening area within 2.5° for the pink noise source using TCS (pan) compared with 10% for the TCS (mics) system. Similarly, the model predicted 14% of the listening area within 2.5° for the pink noise source using FCS (pan) compared with 3% for the FCS (mics) system. It is interesting to note that from looking at the results using the 2.5° threshold it appears that TCS (pan) performs better than FCS (pan). This is slightly surprising, given that the signal for TCS (pan) is panned between loudspeakers at -30° and 30°, while the signal for FCS (pan) is panned between loudspeakers at 0° and -30°. It would be expected that the positioning of the source image would be more accurate for FCS (pan), as it uses a more closely spaced pair of loudspeakers. However, recall from Section 5.2 that the model tends to predict the angles slightly closer to ±90° than the results from the listening tests. These small errors in the output from the model may dominate the results when using small error thresholds. This supported by the results in the column in Table 5.13 for the 5° threshold: here it can be seen that FCS (pan) has a
<table>
<thead>
<tr>
<th>Reproduction system</th>
<th>Original signal</th>
<th>Threshold angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2.5</td>
</tr>
<tr>
<td>FCS (mics)</td>
<td>African percussion</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>3</td>
</tr>
<tr>
<td>FCS (pan)</td>
<td>African percussion</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>14</td>
</tr>
<tr>
<td>TCS (mics)</td>
<td>African percussion</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>10</td>
</tr>
<tr>
<td>TCS (pan)</td>
<td>African percussion</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Classical guitar</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Female speech</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Male speech</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Pink noise</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 5.13: This table contains the percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold. The original source was modelled positioned at (−20°, 3) and the listener was facing forwards for all listening positions. Each of the five columns on the right of the table corresponds to a different threshold angle, e.g. the third column in the table shows the percentage of the listening area where the difference between the intended and predicted angles is less than 2.5°.
Table 5.14: This table contains the percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold. The original source was modelled positioned at (40°, 3) and the listener was facing forwards for all listening positions. Each of the five columns on the right of the table corresponds to a different threshold angle, e.g. the third column in the table shows the percentage of the listening area where the difference between the intended and predicted angles is less than 2.5°.
larger percentage of the listening area with errors less than 5° than TCS (pan).

Table 5.14 contains the results for sources positioned at (40°, 3). By comparing this table's columns for the 2.5°, 5° and 10° thresholds with the corresponding columns in Table 5.13 it can be seen that the intended and predicted angles are in less agreement for the source located at (40°, 3) than they are for the source located at (−20°, 3). There are two possible explanations for this. This first is that the model performs worse as the angle of the source relative to the listener's head moves to the side, closer to ±90°. The second explanation is that the performance of the reproduction system deteriorates as the intended location of the sources moves to the side to ±90°.

Table 5.15 summarises the results for a wider range of source angles. Only pink noise has been used in this table, as it was seen in Tables 5.13 and 5.14 that using different original signals did not make a large difference to the results. In this table it can be seen that the errors increase as the position of the source moves to the side. Again, this could be due to either the poorer performance of the model as the angle of the source relative to the listener's head increases, the inability of the reproduction systems to accurately place images at the side, or a combination of the two. However, the same model was used to predict the localisations for all the reproduction systems considered, so the errors due to the model were consistent for all the different systems. This means that a comparison can still be made between the results for the different reproduction systems.

It was noted in the discussion of the plots in Figs. 5.43 to 5.52 that the TCS (mics) and FCS (mics) systems performed much worse than the other reproduction systems considered. This is confirmed in Table 5.15, where the proportion of the listening area with small errors decreases rapidly as the position of the original source moves to the side. From the table it can be seen that the WFS system has a consistently larger proportion of the listening area within the error thresholds for the 0° to 30° source angles than the other systems. The performance of the WFS system appears to deteriorate for source angles in the 50° to 60° range, and it appears to be out-performed by the FCS (pan) system for this range of angles when using error thresholds less than 10°. However, Table 5.15 also shows that the TCS (pan) system has the largest proportion of the listening area with errors less than 2.5° for the source angle 40°. As mentioned previously, this signal was panned to the right hand loudspeaker only of the TCS setup, so the actual position of the source image is known to be 30°. Indeed, the loudspeaker signals are identical for all the source angles greater than or equal to 30°. If the localisation model had no errors then the 30° source angle would have the best localisation. However, this is not the case and the best localisation with the TCS (pan) system is for the 40° source angle. This is consistent with the results of Section 5.2, where it was seen that the errors due to the model increase as the source location moves towards ±90°, with the model predicting the angles closer to ±90° than the listening test results.

It was then decided to calculate a number of results across the listening area where the position of the original source and the reproduction system were both kept constant and only the orientation of the listener was varied. Doing this meant that the physical reproduced sound-field remained constant and so any changes in the predicted angles over these different results will be due only to the changes in the listener orientation. The reproduction system used in this investigation was 32-channel WFS: this was chosen as it showed the best
<table>
<thead>
<tr>
<th>Threshold</th>
<th>Reproduction system</th>
<th>Source angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2.5</td>
<td>FCS (mics)</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>FCS (pan)</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>TCS (mics)</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>TCS (pan)</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>WFS</td>
<td>47</td>
</tr>
<tr>
<td>5</td>
<td>FCS (mics)</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>FCS (pan)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>TCS (mics)</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>TCS (pan)</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>WFS</td>
<td>86</td>
</tr>
<tr>
<td>10</td>
<td>FCS (mics)</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>FCS (pan)</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>TCS (mics)</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>TCS (pan)</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>WFS</td>
<td>99</td>
</tr>
<tr>
<td>20</td>
<td>FCS (mics)</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>FCS (pan)</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>TCS (mics)</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>TCS (pan)</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>WFS</td>
<td>100</td>
</tr>
<tr>
<td>30</td>
<td>FCS (mics)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>FCS (pan)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>TCS (mics)</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>TCS (pan)</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>WFS</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.15: This table contains the percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold. The original source was pink noise positioned at a distance of 3m from the centre of the listening area for a range of angles and the listener was facing forwards for all listening positions. Each of the eight columns on the right of the table corresponds to a different source angle.
Table 5.16: This table contains the percentages of the listening area where the difference between the intended and predicted angles is less than a given threshold. The original source was pink noise positioned at (0°, 3) and the reproduction system used was 32-channel WFS. Each row in the table corresponds to a different orientation of the listener's head.

<table>
<thead>
<tr>
<th>Head angle (degrees)</th>
<th>Threshold angle (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5</td>
</tr>
<tr>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>11.25</td>
<td>44</td>
</tr>
<tr>
<td>22.5</td>
<td>31</td>
</tr>
<tr>
<td>33.75</td>
<td>16</td>
</tr>
<tr>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>56.25</td>
<td>3</td>
</tr>
<tr>
<td>67.5</td>
<td>4</td>
</tr>
<tr>
<td>78.75</td>
<td>17</td>
</tr>
<tr>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

performance in terms of accurately positioned across the listening area. The original source was again pink noise. Nine different head orientations were used, from 0° to 90° in increments of 11.25°. This incremental angle was chosen as it is also the angle between adjacent loudspeakers in the circular loudspeaker array used for 32-channel WFS. This means that, due to the symmetry of the 32-channel WFS system, this is equivalent to keeping the orientation of the head fixed at 0° and moving the position of the original source to the left from 0° to -90° in increments of 11.25°: in each case the source will be in line with one of the loudspeakers in the array.

The results of these calculations are shown in Table 5.16, which shows that, even with the same reproduced sound-field, the proportion of the listening area where the source is well localised decreases as the position of the source moves to the side towards ±90° in relation to the listener's head. It can also be seen that the proportion of the listening area with errors less than 2.5° is particularly small for the 45°, 56.25° and 67.5° head angles. This confirms that the similar behaviour seen in Table 5.15 is the result of errors in the model rather than the limitations of the WFS reproduction system.

Note that for the larger head angles the source at (0°, 3) is in the rear hemisphere in relation to the listener (i.e. behind the listener) for some listener positions in the listening area. The modified Supper model assumes that the sources are all in the front hemisphere, i.e. in front of the listener. In order to see whether this was affecting the results in Table 5.16, post-processing was applied to the angles output by the modified Supper model so that when the source was in the listener's rear hemisphere then the angle was modified by being reflected in the vertical plane that passes through both the listener's ears. The following pseudo code shows...
this post-processing (all angles are relative to the modelled listener):

\[
\text{IF (source in rear hemisphere) THEN} \\
\theta_{\text{new}} = 180^\circ - \theta_{\text{old}} \\
\text{ELSE} \\
\theta_{\text{new}} = \theta_{\text{old}} \\
\text{END IF}
\]

However, it was found that this did not have a large effect on either the position or size of the error contours for the results which are summarised in Table 5.16. When the sources were in the modelled listener’s rear hemisphere they were also close to ±90° relative to the listener. This meant that either the effects of the post processing on an individual angle were small (e.g. changing 87° to 93°) or else the errors due to the model were already so large that the post-processed angles also had large errors. The subject of differentiating between front and rear sources with the modified Supper model is considered in more detail in Section 7.2.2.

5.6 Summary

This chapter has described how the directional localisation model developed by Supper [158] was modified to both fit into the framework described in Chapter 3 and also to improve its performance. The modified Supper model was then validated against the directional localisation results from the three listening test experiments described in Chapter 4. The model was shown to be able to predict accurately the localisation results, with the exception of the reflective stimuli from the first listening test experiment and the artificially widened stimuli from the third listening test experiment.

The difficulty experienced by the model in predicting these reflective stimuli led to an investigation of the effects of reflections on the performance of the model. This was divided into the effects of loudspeaker signals created using microphone techniques in a reflective environment and the effects of a reflective listening environment. In both cases the angles predicted by the model tended to 0° as the modelled rooms became more reflective. This appears to be an exaggerated version of the behaviour exhibited by real listeners in the first listening test experiment. Contained within this section was a discussion of changes that could be made to the model to improve its performance with reflective signals, principally incorporating Supper’s onset detector [158].

This was followed by an investigation comparing the abilities of different reproduction systems to accurately reproduce source locations for a listener at the sweet spot. Here it was found that the interpretation of
the results was not straightforward due to the errors from the model (i.e. predicting angles with a slightly larger magnitude than the angles obtained from the listening tests and spurious predicted angles of 0° when the source is close to ±90°). However, comparing the model results for the different reproduction systems with the model results of the original sound-fields showed that, of the reproduction systems considered, the 32-channel WFS system was able to produce the most accurately placed sources.

Finally, it was demonstrated how a combination of spatially sampling the listening area and using the localisation model could be used to assess the spatial performance of different reproduction systems across the area. Again, it was seen how the limitations of the model affected these results. Taking these limitations into account when interpreting the results again showed the 32-channel WFS system to produce the most accurately positioned sources across the listening area.
Chapter 6

Source width

This chapter describes how two existing models were used to predict source widths for the stimuli used in the third listening test. The two models considered were the Supper model [158] after it had been modified as described in Section 5.1 and the source width model developed by Mason [98].

The first model for predicting source width described in this chapter is the modified Supper model. This investigation was undertaken as the Supper model has been extensively modified (as described in the previous chapter), the histogram output has already been calculated for the prediction of directional localisation and Supper has already described how his model may be used to predict source width [158].

Much of the literature on the perception of source width has approached the subject from the point of view of concert hall acoustics [119]. Indeed, auditory source width (ASW), as defined by Bradley and Soulodre [25], has been found to be principally related to the early lateral reflections. The investigation into the perception of localisation described in Chapter 5 used Supper's localisation model, the design of which is very influenced by psychoacoustics: some of the stages explicitly model processes in human listening, such as the use of binaural signals and the separation of these signals into critical bands. Indeed, the model framework described in Chapter 3 and illustrated in Fig. 3.1 explicitly uses a psychoacoustical approach, in particular in stage VIII, the translation from binaural signals to the perceptual domain. As such, this approach will also be used for the prediction of source width.

The relationship between psychoacoustics and concert hall acoustics has been investigated, not least by Bradley [23], who concluded that the IACC in critical bands lower than 2kHz was significantly related to the $LF_{80}$ measure of ASW (see Section 2.4). This relationship between perceived width and the IACC forms the basis of Mason's source width model, the second of the two models discussed in this chapter. It is important to note that this relationship is an outcome of research in the field of concert hall acoustics. It is repeatedly noted in the literature that the early lateral reflections common in concert hall acoustics which give a sense of width have an inverse relation to IACC values. However, it is not claimed that the relationship between the IACC and perceived width is valid for all binaural signals. It is possible that the primary cue for width
perception is not the IACC, and that the behaviour observed in the IACC is a consequence of the other factors present in wide sounds in naturally occurring acoustical situations (e.g. in concert halls), such as the presence of early lateral reflections. This means that wide sources may imply low IACC values, but the converse may not necessarily be true, i.e. low IACC values may not imply wide sources. This is not an issue in the field of concert hall acoustics, as perceptual cues caused by the acoustics of the hall will always be present (e.g. early lateral reflections). It is an issue, however, in the case of reproduced audio, where the signals can be artificially generated and so aspects of the reproduced sound field (such as the IACCs of the binaural signals heard by a listener) can be manipulated independently of some of the cues that would also be present in a naturally occurring sound field. In particular, this may be an issue when using models based on the IACC to predict the width results from the third listening test experiment, as some of the stimuli used in this experiment were artificially widened.

The chapter concludes with a discussion of how the performance of the two width models could be improved and a more general discussion on the problems of predicting source width, including the design of listening test experiments for obtaining subjective source width data.

6.1 Predicting source width with the modified Supper model

The histograms from the modified Supper model have already been calculated as part of the prediction of directional localisation. Also, there is already a precedent of using the Supper model as a predictor for perceived source width [158]. Both of these facts informed the decision to undertake an investigation into the use of the modified Supper model to predict perceived source width. This section describes two methods used to interpret the modified Supper model output histograms in order to predict perceived source width. This is followed by the results of using these two methods to predict the source width results obtained in the third listening test and a discussion of these results.

6.1.1 Interpreting the output of the modified Supper model

One method Supper proposed for predicting source width was to inspect the histograms output by his model [158]. He argues that wide sources should be more difficult for the model to localise, resulting in the output histograms being more blurred and also in fluctuations in the predicted localisation. This corresponds to an increase in the spread of values in the averaged over time (AOT) histograms. Supper also plots the interquartile range of the output histograms, which suggests that this might be a suitable measure of source width. This method was adopted as a starting point for investigating how the output of the modified Supper model could be used to predict perceived widths. The method of using the interquartile range of the histogram output of the modified Supper model will be referred to as the IQR method in order to facilitate the discussion. It was seen in Section 5.1.6 how the method of combining the localisation histograms from the IIDs and ITDs had a noticeable impact on the ability of the model to predict successfully the location results from listening tests. For this reason the interquartile ranges were calculated for four different histograms
(the two histograms from the separate IID and ITD cues and the two histograms created by taking either the sum or the product of the IID and ITD histograms). Each of these four interquartile ranges was calculated for all the stimuli used in the third listening test. These four sets of interquartile range values were then compared to the width results from the third listening test in order to assess their suitability as a predictor of perceived width.

An alternative measure of the spread of a distribution is the standard deviation, and the second method of interpreting the histograms output by the modified Supper model to predict width is to calculate their standard deviation. This method will be referred to as the SD method. In a similar manner to the IQR method, the SD method was applied to the four sets of histograms created from just the IID cues, just the ITD cues, combining the IID and ITD cues by taking the sum (IID+ITD) and finally combining the two types of cues by taking the product (IID*ITD).

6.1.2 Using the modified Supper model to predict the width results of the third listening test

This section describes the results of using the SD and IQR methods of interpreting the modified Supper model to predict the source width results obtained in the third listening test experiment. Table 6.1 contains values of the Root-Mean-Square Error of Prediction (RMSEP), which has units in degrees, and Pearson product-momentum correlation coefficients (R values) results from using the output of the modified Supper model to predict the width results from the third listening test. The results in this table are qualitatively similar for both the IQR and the SD methods. For both methods, using the histograms created using just the ITD cues resulted in a relatively high degree of correlation between the predicted widths and the widths from the listening test. However, this was offset by a correspondingly high RMSEP value. Similarly, for both methods the histograms created by taking the product of the IID and ITD histograms resulted in a much lower RMSEP than for the other histograms, but still a relatively high R value. This offers further support for the method of using the product to combine the IID and ITD histograms, which was also found to improve the directional localisation performance of the model, as discussed in Section 5.1.6. From the table it can be seen that the SD method is able to predict width better than the IQR method, giving both higher R values and lower errors for all the output histograms being considered. However, despite relatively low RMSEP values, an R value of 0.35 is still low, showing that the model can only explain 12% of the variance in the listening test width results.

Table 6.2 shows the Pearson product-momentum correlation coefficients (R values) from predicting the width results from the third listening test with the SD method of interpreting the IID*ITD output of the modified Supper model. This table also shows the R values from separating the listening test results into the different types of signals (Single, Panned and Widened) and the three different listening positions and using the model to predict these subsets of the results. From this table it can be seen that the model was able to explain the variance better for some combinations of signal type and listener position than others. There is no obvious pattern to these results: the highest R values are for the single loudspeaker stimuli at the right...
listener position (0.69) and the panned stimuli at the left listener position (0.67), yet the combination of single loudspeaker stimuli and left listener position gave the lowest R value (0.08).

Table 6.3 shows the root-mean-square error of prediction from using the modified Supper model to predict the width results from the third listening test. The listening test results have been decomposed into the same subsets as Table 6.2. The largest error value is for the combination of the widened stimuli and the left listening position (13°). However, even this is relatively small when considering the fact that the largest standard deviation for the listening test results with the widened stimuli and the left listening position is 33°.

Fig. 6.1 shows the widths predicted by the modified Supper model plotted against the width results from the third listening test. The predicted results in both these figures were calculated using the SD method from the histograms created taking the product of the IID and ITD histograms. Fig. 6.2 shows the same results with the addition of the confidence intervals for the listening test results, which show how varied and inconsistent the listening test results are. Note that for three combinations of stimulus and listening position the line of best fit does not pass through the 95% confidence intervals (stimuli 10 and 14 at the left listening position and stimulus 14 at the right position).
Table 6.1: The Root-Mean-Square Error of Prediction (RMSEP) and Pearson product-momentum correlation coefficient (R) results from using the output of the modified Supper model to predict the width results from the third listening test. The first column shows whether the interquartile range (IQR) or the “standard deviation” method (SD) was used to interpret the histograms output by the modified Supper model. The second column shows which histogram output by the modified Supper model was used to predict the width (the histograms resulting from either the IID cues, the ITD cues or a combination of the two).

<table>
<thead>
<tr>
<th>Method</th>
<th>Model results</th>
<th>RMSEP</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQR method</td>
<td>iid*itd</td>
<td>8.89</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>iid+itd</td>
<td>20.33</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>itd</td>
<td>42.85</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>iid</td>
<td>17.78</td>
<td>0.02</td>
</tr>
<tr>
<td>SD method</td>
<td>iid*itd</td>
<td>7.23</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>iid+itd</td>
<td>19.32</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>itd</td>
<td>24.97</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>iid</td>
<td>12.93</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 6.2: The correlation coefficients (R) between the width results of the third listening test and the predicted widths from using the SD method of interpreting the combined IID and ITD histogram output from the modified Supper model. The IID and ITD results were combined by taking the product.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centre</td>
</tr>
<tr>
<td>Single</td>
<td>0.38</td>
</tr>
<tr>
<td>Panned</td>
<td>0.15</td>
</tr>
<tr>
<td>Widened</td>
<td>0.44</td>
</tr>
<tr>
<td>All processes</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 6.3: The root-mean-square error of prediction (RMSEP) from using the SD method of interpreting the combined IID and ITD histogram output from the modified Supper model to predict the width results of the third listening test. The IID and ITD results were combined by taking the product.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centre</td>
</tr>
<tr>
<td>Single</td>
<td>5.65</td>
</tr>
<tr>
<td>Panned</td>
<td>7.01</td>
</tr>
<tr>
<td>Widened</td>
<td>5.76</td>
</tr>
<tr>
<td>All processes</td>
<td>6.17</td>
</tr>
</tbody>
</table>
Figure 6.1: The results of using the standard deviations from the histograms output from the modified Supper model to predict the source width results from the third listening test. The top left graph shows the results for all three listening positions, the top right graph shows the results from the centre listening position and the results from the left and right listening positions are shown in the bottom left and bottom right graphs respectively. In the top left graph the results for the centre, left and right listening positions are plotted in red, blue and green respectively. In the other graphs the stimuli created by either being played through a single loudspeaker, pair-wise panning or the widening algorithm have been plotted in red, blue and green respectively. In all the graphs the outliers have been labelled to show the listener position (centre (C), left (L) and right (R)) and the stimulus number. The line of best fit was plotted through the data in each graph.
Figure 6.2: The results of using the standard deviations from the histograms output from the modified Supper model to predict the source width results from the third listening test, together with 95% confidence intervals. These graphs show the same results as the graphs in Fig. 6.1 with the addition of the 95% confidence intervals added. The confidence intervals are plotted as dotted lines for those combinations of listening position and stimulus whose width results from the listening test were judged not to be normal by the Shapiro-Wilk test (see Section E.3). The top left graph shows the results for all three listening positions, the top right graph shows the results from the centre listening position and the results from the left and right listening positions are shown in the bottom left and bottom right graphs respectively. In the top left graph the results for the centre, left and right listening positions are plotted in red, blue and green respectively. In the other graphs the stimuli created by either being played through a single loudspeaker, pair-wise panning or the widening algorithm have been plotted in red, blue and green respectively. The line of best fit was plotted through the data in each graph.
6.2 The Mason source width model

This section contains a brief description of the design of the Mason width model together with a description of how the model was integrated into the framework described in Chapter 3. The section also contains the results of using the Mason width model to predict the width results from the third listening test together with a discussion of these width predictions.

Predicting source width is only one part of the functionality of the model developed by Mason, which also predicts the localisation of sound sources using IIDs and ITDs from correlograms calculated in each critical band. There are similarities between the Mason model and the Supper model. Indeed, the architectures of both models are identical up to the point when the IACCs have been calculated: the input binaural signals are filtered into critical bands, each band is then rectified and low pass filtered, followed by the calculation of the IACC from correlograms for each critical band. The main difference between the localisation algorithms in the Supper model and Mason models is that the Supper model uses fuzzy logic techniques to determine the angle of localisation, and so the output of the Supper includes information on the degree to which the binaural signal input to the model belongs to each of the sets of binaural signals which are localised at a particular angle. In contrast, the localisation algorithm used in Mason's model returns a single angle for each IID or ITD in each critical band.

An overview of the parts of the Mason model relevant to the calculation of source width is shown in Fig. 6.3. The main idea of the Mason model is that for each critical band the IACC has a linear relationship to the perceived source width. As in the listening tests described in Chapter 4, the Mason model assumes that source width can be perceived as a spanned angle, so the Mason model calculates a spanned angle for each critical band. For each discrete time frame these values of spanned angles are then combined into a single histogram across the possible range of spanned angles, as shown in Fig. 6.4. In these histograms, the value for a given angle is equal to the number of critical bands where the calculated spanned angle is larger than the given angle. For instance, in the histogram at the bottom of Fig. 6.4 it can be seen that the spanned angle of 10° has a histogram value of 22. This shows that the widths (spanned angles) greater than 10° were calculated for all twenty-two critical bands, as can be seen in the horizontal bar chart above the histogram. In contrast, the spanned angle of 80° has a value of one assigned to it in the histogram, as only a single critical band had a calculated spanned angle greater than 80°.

The example shown in Fig. 6.4 assumed that the signals in each critical band all had equal loudness. In practice this rarely happens, and so the contributions of each critical band signal to the final histogram are weighted by their loudness. The output of the Mason model is a series of these histograms, and is displayed as an intensity plot, such as that shown in Fig. 6.5. This can be interpreted in a relatively intuitive manner: time is shown on the x-axis and the width of the sound source is represented by the shaded area. The more the IACC cues agree with each other across the different critical bands, then the sharper the transition from the high values to the low values in the histogram. This, in turn, leads to a more defined edge to the shaded area in the intensity plot. Conversely, if the calculated widths vary greatly over the critical bands then there is a longer transition between the high and low values in the histogram, leading to more blurred, less...
Figure 6.3: The structure of the Mason source width model. Note that the frequency and loudness compensation are both optional. Only the parts of the Mason model relating to the calculation of source width are shown in the diagram: the model also calculates source location using IIDs and ITDs in a similar way to the Supper model.

defined boundary to the shaded area in the intensity plot. In addition to this, the loudness weighting means that the intensity of the output plot from the Mason model is proportional to the loudness of the binaural signals input to the model. This can be seen in Fig. 6.5, which was calculated using the binaural signals corresponding to the ninth stimulus from the third listening test at the central listening position. The gap between the words, where the signals were much quieter, corresponds to a gap in the shaded area on the plot.
Figure 6.4: Combining the results across frequency. This diagram shows how the spanned angles calculated for each critical band are combined in the Mason width model. A horizontal bar chart is created of the calculated spanned angles in each critical band. This is equivalent to populating a two dimensional array where each row corresponds to a critical band and each column corresponds to a spanned angle. Each element in the array is assigned a value of one if the angle calculated by the model for that critical band is greater than the or equal to the angle corresponding to the array element. If this is not the case then the element is assigned a value of zero. The model results are then combined by summing the columns in the horizontal bar chart (i.e. summing the columns in the two dimensional array), as illustrated by the graph at the bottom of this figure.
6.2.1 Integration of the Mason source width model into the framework

The previous sections included a discussion of some of the similarities between the architecture and the algorithms of the Supper localisation model and the Mason width model. Moreover, both the Supper and Mason models take binaural signals as their input and both output a series of histograms. In the context of the model framework described in Chapter 3 and Fig. 3.1 in particular, both the Supper and Mason models correspond to stage VIII, the translation to the perceptual domain.

Chapter 4 described a series of listening tests which were undertaken in order to provide a data set of localisation and source width results against which the models being investigated could be validated. One of the major elements in the design of these listening test experiments was allowing the subjects to easily and clearly communicate their perceptions of localisation and source width to the investigator, and attempting to minimise any sources of error in this communication. This aspect of the experimental design corresponds to stage IX in Fig. 3.1: obtaining measures of spatial attributes.

For each combination of subject, stimulus, listening position and spatial percept being investigated, the graphical user interfaces employed in the experiments output a single number: the azimuth of the sound source position for localisation and the angle subtending the left and right edges of the sound source for width. Encouraging the listeners to effectively summarise each of their perceptions into a single measure simplified the act of communication between the listeners and the investigator, thus diminishing the likelihood of errors being introduced. However, a single measure cannot convey the complexity of the listeners' perceptions, e.g. such aspects of each listener’s perception as the ease of localisation, or whether the sound is perceived to be inside or outside the head. Some of this complexity can be inferred by analysing the results of having a
number of subjects undertake the experiments. For instance, in the first listening test the localisation results for the sine wave stimuli varied widely across the range of subjects, which implies that these stimuli are hard to localise.

In order to be able to verify the Supper and Mason models their output must be of the same form as the listening test results. Now, for each combination of stimulus and listening position in each of the listening test experiments, each listener gave a single value for each percept being measured, giving a distribution of these measures for each combination. A single measure can be obtained from these distributions by taking the mean or the median. This gives two possible forms for the measures output by the Supper and Mason models: either a single value which corresponds to the angles from the listening tests (azimuth for localisation and subtended angle for width) or, alternatively, a distribution of these angles.

The validation of the models is made easier if the first of these two options is taken. Indeed, this was the option used for localisation in Chapter 5 and also for the investigation into using the Supper model to predict source width described in Section 6.1. Thus, the integration of the Mason source width model into the model framework requires that the output of the Mason model is transformed from a time series of histograms to a time series of single values. This corresponds to stage IX in Fig. 3.1.

Recall that the output of the unmodified Mason source width model is a series of cumulative histograms similar in form to the histogram shown at the bottom of Fig. 6.4. These histograms are always monotonically decreasing, i.e. the value for the angle \( x \) is equal or less than the value for the angle \( x - 1 \) for all values of \( x \) greater than zero. In these histograms, angles have the highest value when the results for all twenty-two critical bands agree that the source is wider than the angle, and angles have a value of zero when the results for all the critical bands agree that the source is narrower than the angle.

Fig. 6.6 shows an alternative method of calculating a histogram from the widths calculated in each of the critical bands. This method is identical to that shown in Fig. 6.4 except for one detail. In the previous method a horizontal bar-chart was first created, the columns of which were then summed to create the output histogram. In comparison, in the method illustrated in Fig. 6.6 a two dimensional array is created where only the elements corresponding to the width angles calculated for each critical band are assigned positive, non-zero values. In a similar manner to the previous method, the columns of this array are then summed to give the final histogram, as shown in the histogram at the bottom of Fig. 6.6.

Note that the histograms generated using the previous method are a form of cumulative plot of the distribution shown in the histograms generated using the new method. Compared to the previous histograms, these new histograms have a more intuitive interpretation: the histogram value for a given angle is high if the IACC cues from multiple critical bands correspond to the angle, and the value is zero if none of the critical bands have IACC cues corresponding to the angle. Therefore, angles with high histogram values are more likely to be the perceived width and angles with low histogram values are unlikely to be the perceived width. The meaning of these new histograms has similarities with that of the histograms output by the Supper model. As such, similar methods of extracting a single angle from the histograms can be employed, in particular the peak-picking method described in Section 5.1.1. For example, in the histogram at the
Figure 6.6: Generating width histograms. This diagram shows how the spanned angles calculated for each critical band are combined to generate histograms of the width results. A two dimensional array is populated where each row corresponds to a critical band and each column corresponds to a spanned angle. Each element in the array is assigned a value of one if the angle calculated by the model for that row (i.e. critical band) is equal to the angle corresponding to the array element. If this is not the case then the element is assigned a value of zero. The model results are then combined by summing the columns in the array, as illustrated by the bottom graph. This should be compared with Fig. 6.4. Note that the graph at the bottom of Fig. 6.4 is equivalent to the cumulative frequency of the graph at the bottom of this figure.
bottom of Fig. 6.6 the peak-picking method would extract 24°, as this is the location of the largest peak in the histogram.

As in the previous chapter, only stationary stimuli will be considered, allowing the entire time series to be considered when extracting a single value for the source width. As has already been noted, all the stimuli used in the three listening tests have been stationary. The series of histograms will be combined into a single histogram by averaging the histograms over time, as shown in Fig. 6.7. This is identical to the method used with the series of histograms output by the Supper model in the previous chapter and also in Section 6.1.

Figure 6.7: These graphs show the intermediate results for stimulus 9 from the third listening test. The graph on the left shows the time series of histograms calculated from the output of the Mason model as described in Section 5.1 and Fig. 6.6. The graph on the right shows the Averaged Over Time (AOT) histogram.

6.2.2 Using the extended Mason source width model to predict the width results of the third listening test

This section describes the results of using the Mason model to predict the source width results obtained in the third listening test experiment. The results are shown in Table 6.4 along with details of the stimuli.

Table 6.5 shows the Pearson product-momentum correlation coefficients (R values) from predicting the width results from the third listening test with the Mason model. This table also shows the R values from separating the listening test results into the different types of signals (Single, Panned and Widened) and the three different listening positions and using the Mason model to predict these subsets of the results. As with the corresponding table for the modified Supper model, Table 6.2, it can be seen that the Mason model was able to explain the variance better for some combinations of signal type and listener position than others. The R values for the single and panned sources are generally higher for the Mason model than those for
### Table 6.4: The source width results from the third listening test and the results predicted by the Mason source width model.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Original signal</th>
<th>Processing</th>
<th>Loud-speakers</th>
<th>Listening test</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>widths (degrees)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L</td>
<td>C</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>PN</td>
<td>S</td>
<td>1</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>PN</td>
<td>S</td>
<td>4</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>PN</td>
<td>S</td>
<td>8</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>PN</td>
<td>P</td>
<td>5,7</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>PN</td>
<td>P</td>
<td>1,2</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>PN</td>
<td>P</td>
<td>2,3</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>PN</td>
<td>W</td>
<td>1,2</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>PN</td>
<td>W</td>
<td>5,8</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>FS</td>
<td>S</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>MS</td>
<td>S</td>
<td>7</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>FS</td>
<td>S</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>MS</td>
<td>P</td>
<td>1,2</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>FS</td>
<td>P</td>
<td>4,5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>MS</td>
<td>P</td>
<td>5,7</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>MS</td>
<td>W</td>
<td>1,2</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>FS</td>
<td>W</td>
<td>5,8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>17</td>
<td>CG</td>
<td>S</td>
<td>2</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>18</td>
<td>T</td>
<td>S</td>
<td>6</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>19</td>
<td>C</td>
<td>S</td>
<td>5</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>AP</td>
<td>P</td>
<td>1,2</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>CG</td>
<td>P</td>
<td>5,7</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>22</td>
<td>T</td>
<td>P</td>
<td>4,8</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>C</td>
<td>W</td>
<td>1,2</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>24</td>
<td>AP</td>
<td>W</td>
<td>5,8</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>25</td>
<td>PN</td>
<td>W</td>
<td>5,6</td>
<td>17</td>
<td>39</td>
</tr>
<tr>
<td>26</td>
<td>MS</td>
<td>W</td>
<td>5,6</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>27</td>
<td>T</td>
<td>W</td>
<td>5,6</td>
<td>22</td>
<td>34</td>
</tr>
</tbody>
</table>

**Key:**

<table>
<thead>
<tr>
<th>Original signal</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN</td>
<td>S</td>
</tr>
<tr>
<td>FS</td>
<td>P</td>
</tr>
<tr>
<td>MS</td>
<td>W</td>
</tr>
<tr>
<td>CG</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>AP</td>
<td></td>
</tr>
</tbody>
</table>

*S* Played through a single loudspeaker  
*P* Pair-wise panned  
*W* Widening algorithm
the modified Supper model (the correlation for the combination of panned signals and left position being an exception). In Table 6.5, the correlations for the artificially widened signals are noticeably lower than for the other signal types. Overall, the Mason source width model has higher correlations than the modified Supper model, showing that the Mason model is better able to explain the variation in the width results from the third listening test.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Centre</th>
<th>Left</th>
<th>Right</th>
<th>All positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>0.57</td>
<td>0.40</td>
<td>0.78</td>
<td>0.55</td>
</tr>
<tr>
<td>Panned</td>
<td>0.54</td>
<td>0.43</td>
<td>0.70</td>
<td>0.39</td>
</tr>
<tr>
<td>Widened</td>
<td>0.22</td>
<td>0.03</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>All processes</td>
<td>0.23</td>
<td>0.50</td>
<td>0.68</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 6.5: The correlation coefficients ($R$) between the width results of the third listening test and the predicted widths from the Mason model.

Table 6.6 shows the root-mean-square error of prediction (RMSEP) from using the modified Supper model to predict the width results from the third listening test. The listening test results have been decomposed into the same subsets as Table 6.5. These results are comparable to the corresponding RMSEP results for the modified Supper model in Table 6.3.

<table>
<thead>
<tr>
<th>Processing</th>
<th>Centre</th>
<th>Left</th>
<th>Right</th>
<th>All positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>4.57</td>
<td>5.09</td>
<td>6.07</td>
<td>5.28</td>
</tr>
<tr>
<td>Panned</td>
<td>4.08</td>
<td>9.85</td>
<td>8.83</td>
<td>7.99</td>
</tr>
<tr>
<td>Widened</td>
<td>4.45</td>
<td>12.00</td>
<td>8.23</td>
<td>8.78</td>
</tr>
<tr>
<td>All processes</td>
<td>4.37</td>
<td>9.43</td>
<td>7.80</td>
<td>7.50</td>
</tr>
</tbody>
</table>

Table 6.6: The root-mean-square error of prediction (RMSEP) from using the Mason model to predict the width results of the third listening test.

Fig. 6.8 shows the widths predicted by the Mason model plotted against the width results from the third listening test. Fig. 6.9 shows the same results with the addition of the confidence intervals for the listening test results, which again show how varied and inconsistent the listening test results are. Note that the line of best fit passes through all the 95% confidence intervals in all the graphs in Fig. 6.9, unlike the corresponding graphs for the modified Supper model (Figs. 6.2).
Figure 6.8: The results of using the Mason model to predict the source width results from the third listening test. The top left graph shows the results for all three listening positions, the top right graph shows the results from the centre listening position and the results from the left and right listening positions are shown in the bottom left and bottom right graphs respectively. In the top left graph the results for the centre, left and right listening positions are plotted in red, blue and green respectively. In the other graphs the stimuli created by either being played through a single loudspeaker, pair-wise panning or the widening algorithm have been plotted in red, blue and green respectively. In all the graphs the outliers have been labelled to show the listener position (centre (C), left (L) and right (R)) and the stimulus number (see Table 4.10). The line of best fit was plotted through the data in each graph.
Figure 6.9: The results of using the Mason model to predict the source width results from the third listening test, together with 95% confidence intervals. These graphs show the same results as the graphs in Fig. 6.8 with the addition of the 95% confidence intervals added. The confidence intervals are plotted as dotted lines for those combinations of listening position and stimulus whose width results from the listening test were judged not to be normal by the Shapiro-Wilk test (see Section E.3). The top left graph shows the results for all three listening positions, the top right graph shows the results from the centre listening position and the results from the left and right listening positions are shown in the bottom left and bottom right graphs respectively. In the top left graph the results for the centre, left and right listening positions are plotted in red, blue and green respectively. In the other graphs the stimuli created by either being played through a single loudspeaker, pair-wise panning or the widening algorithm have been plotted in red, blue and green respectively. In all the graphs the outliers have been labelled to show the listener position (centre (C), left (L) and right (R)) and the stimulus number (see Table 4.10). The line of best fit was plotted through the data in each graph.
6.3 Discussion

From the results in Sections 6.1.2 and 6.2.2 it can be seen that the modified Supper model and the Mason model perform similarly when attempting to predict the width results from the third listening test experiment. The RMSEP values are slightly lower for the modified Supper model, although the errors for both models are relatively low when the variability of the widths elicited from the listeners is considered (the largest standard deviation from the listening test results is 33°). When all the stimuli and listening positions are considered then the Mason model has a higher correlation than the modified Supper model. The correlations for both width models are much lower than those obtained for directional localisation using the modified Supper model (see Table 5.2). A possible reason for this is that directional localisation is a less complicated perceptual attribute than source width.

Rumsey [140] proposed at least three different types of perceived width. The first is the width of individual sources in the sound scene, which Rumsey notes is often related to the ease or difficulty with which the source can be located. The second type of width is ensemble width, which is the width of a number of sources that are grouped together cognitively, e.g. a string quartet or an orchestra. The third type of width proposed by Rumsey is environmental width, which is the width perceived due to reflections in the room. Rumsey also proposes a possible fourth type of width: the width of the overall scene. This will mainly be identical to the environmental width, although there is the possibility in artificially constructed scenes in reproduced audio that sources may be placed outside the implied environment.

Now, in the case of individual source width, a source will be perceived as wide if either the source is diffuse and hard to localise or else the left and right edges of the source are localisable and far apart. Determining the locatedness or diffuseness of a source requires first attempting to localise the source and then judging the ease or difficulty of this act of localisation. This is clearly a higher level cognitive process than merely localising the source. Similarly, localising both the left and right edges of an individual source and assessing the distance between them is also a higher level cognitive process than just localisation. This is also the case with ensemble width, where the listener has to localise the component sources at the left and right edges of the ensemble. Rumsey states that the environmental width is perceived using a cognitively separate information stream to the perception of the spatial qualities of the direct sound from the foreground sources. Again, this separation of the binaural signals into foreground and background streams (see Griesinger [66]) is cognitively much more complicated than simple localisation. The fact that all these forms of the perception of width are both more complicated and higher level processes than just localisation means that it is unsurprising that the width models discussed in this chapter performed worse at predicting the listening test results than the localisation model discussed in the previous chapter. The complexity of the perception of width means that it is a harder task for the subjects in listening tests [115]. This can be seen in Section 4.4.5, where the width results from the third listening test experiment were much more inconsistent than the localisation results. It also means that it is a much more difficult perception to model.

The boundaries between the different types of width perception described by Rumsey are not clear cut. For example, the early reflections due to the environment may also cause a source to become more diffuse,
thus increasing its individual source width. Another example is that a large object such as a car could be considered as either a single large source or else as an ensemble of smaller distinct sources.

It was implicit in the second and third listening test experiments that it was the individual source width that was being investigated rather than any of the other types of width described by Rumsey. Now, informal listening to the stimuli used in the third listening test showed that the wider sources were generally considered to be more diffuse than the others, rather than having well defined edges to the sound. The possible exception to this was stimulus 27, which consisted of a trumpet artificially widened and played through loudspeakers 75° apart. With this stimulus the sound source appeared to move to the right during the course of the trumpet's phrase, which started low in pitch and proceeded to ascend an octave. This will have been due to the nature of the algorithm used to artificially widen the signal: an all pass filter was used to create a copy of the original signal which was identical except for having a randomised phase shift which varied over frequency, and then the original signal and its filtered copy were both panned to different positions between a pair of loudspeakers. This has the consequence that different frequency bands in the widened signals will have their perceived location at different positions between the two loudspeakers. For broadband signals such as pink noise (e.g. stimulus 25 in the third listening test) this will result in the source becoming more diffuse and harder to localise. In the case of the widened trumpet stimulus, however, this resulted in the low notes being perceived in one location followed by the higher notes being perceived in another. Were these two extremes (the positions of the low notes and the high notes) perceived as the left and right edges of the source? This is likely given the context of the task assigned to the subjects in the listening test.

Recall that the method used to elicit the width information in the second and third experiments was for the listeners to identify the left and right edges of each stimulus. As has been discussed above, the majority of the wide stimuli were experienced more as being diffuse and hard to localise rather than having two widely spaced distinctly localisable edges to the sound, with stimulus 27 being the exception rather than the rule. Consequently, in the second and third listening tests the width results depended on the subjects being able to localise the left and right edges of a source which was already hard to localise in its own right. This suggests that the design of the user interface, with the decision to elicit widths by identifying the left and right edges of each stimulus, may have exasperated an already difficult task.

It is possible that a user interface based on a more graphical language such as that proposed by Ford et al. [51] may be more suited to the task of eliciting source width. In Fig. 6.10 can be seen a single object represented in this graphical language, where the circle containing the “P” represents the focal point of the sound, the rounded rectangle describes the shape of the sound and the cloud on the outside represents a feeling of space. For the purposes of this experiment, where the width of the stimuli is generally of the individual source width type, this graphical language could be modified by omitting the cloud representing the feeling of space and also adding shading to the rounded rectangle so the listener can use the language to differentiate between large objects with clearly defined edges and large diffuse objects, as shown in Fig. 6.11. The results of using such a graphical language in an experiment to gather width data would be harder to interpret when compared to the method used in the second and third listening tests described in Chapter 4, but it may have the advantage that it is closer to the perception of the listeners.
Figure 6.10: Example of a single sound source described using the graphical assessment language proposed by Ford et al. [51]. The circle containing the letter “P” represents the focal point of the sound, the rounded rectangle represents the shape of the sound and the cloud surrounding these represents a feeling of space. Figure adapted from [51].

Figure 6.11: Examples illustrating how the assessment language proposed by Ford et al. [51] may be modified for the purposes of experiments to gather width data. As in Fig. 6.10, the circle containing the letter “P” represents the focal point of the sound and the rounded rectangle represents the shape of the sound. The figure on the left represents a sound source where the sound is diffuse (shown by the light grey shading of the rounded rectangle). The figure on the right represents a sound source which has the same shape and size as the figure on the left, but the shape of the sound has clearly localisable boundaries (shown by the black shading of the rounded rectangle).

Another change to the user interface that may make the width results less inconsistent is to give the subjects the ability to navigate more easily between the different stimuli instead of preventing the listener from revisiting earlier stimuli. This would allow comparisons between the different stimuli, which may help the subjects to clarify their width perceptions.

There are a number of changes that could be made to the models described in this chapter that may lead to an improvement in their performance. The first possible improvement is changing the orientation of the modelled head to face the direction of the sound source. The previous validation of the Mason model only used stimuli located directly in front of the subjects [98]. In the case of the modified Supper model it was seen in Chapter 5 that the model was most accurate when localising sound sources positioned around 0° (recall that there were also issues with the unmodified Supper model localising sounds towards ±90°). The importance and relevance of localisation to the assessment of width has been discussed previously in this section. This suggests that using the Supper model in such a way to minimise the localisation errors (e.g. facing the artificial listener towards the source) may also improve the ability of the Supper model to predict width. In this regard the modified Supper model reflects the behaviour of human listeners, who are also
better at localising sounds in front (i.e. closer to 0°) than sounds at the sides (closer to ±90°). This is partly due to the fact that both human listeners and the modified Supper model use the differences between the left and right ear signals as their primary cues for localisation. Changes in the direction of the source relative to the listener have a greater effect on these interaural differences near 0° than they do near ±90°. In the same way that real listeners, given a localisation task, will first locate the sound source approximately and then turn their head towards the source in order to be able to localise it most accurately, it is likely that listeners behave similarly when assessing the width of a sound source, i.e. first localise the sound source and then turn the head to face in this direction. This suggests that not placing any restrictions on head movement during the width listening tests and also allowing the model to choose the orientation of the head based on an initial localisation may both better reflect what real listeners do in practice and so give more ecologically valid results.

The second change which could be made to both width models is to consider explicitly the cases where a source is localised at different positions in different critical bands (note that the full Mason model includes the functionality of localising sources, although this was not considered in the investigation of width prediction described in Section 6.2). This modification would be particularly relevant for sounds such as stimulus 27 from the third listening test, the trumpet whose location appeared to vary depending on pitch.

Another change to the models that may lead to an improvement in the prediction of width is to move them both towards a more Bayesian approach. This has already been discussed with regard to the modified Supper model in Section 5.6, and the resulting probability histograms output by the model may lead to measures of the distributions such as the interquartile range or the standard deviation becoming better predictions for perceived width. Currently in the Mason model, an IACC value is calculated for each critical band at each discrete time frame. Each of these IACC values is then transformed into a single angle of width. These are then combined across the critical bands to give a single distribution (this is described in Section 6.2.1: originally the width values were combined across critical bands in such a way that the resulting distributions could be visually interpreted when plotted (see Fig. 6.5) and this was changed so that the distributions more closely resembled the histograms output by the Supper model). This could be modified by instead calculating the probability of each angle of width given the IACC calculated in that critical band. This would give a probability histogram for each critical band. These could then be combined using Bayesian techniques to give a single distribution of the probabilities of the width angles given the combination of IACCs calculated in all the critical bands. This combined probability distribution would allow the most probable width angle to be identified, but in addition other statistical measures could also be used to infer more information about the width prediction. The probability distributions for each IACC value in each critical band could be determined in a number of ways. One way is to use width data gathered experimentally (such as the second and third listening test experiments described in Chapter 4) to generate look-up tables of probability distributions for each possible IACC value in each critical band. Another method is to approximate each probability distribution with a Gaussian distribution. This would still require that the parameters of the Gaussian distributions be estimated in a sensible manner. Again, the use of width data obtained from listening tests is one possibility for the generation of these parameters.
The final area of possible future work which will be discussed in this section has also been mentioned in Section 5.3.3 in the chapter on directional localisation. This is the fact that in this chapter only stationary sources have been considered. This has simplified the problem of modelling source width, allowing the results of the two models to be averaged over time in order to give more robust predictions. However, this method of averaging over time will not provide sensible results in the case of moving sources or when two different sources in two different locations play one after the other. Quite apart from cases like these, which are not uncommon, the method of averaging the results over time still makes a large assumption about the nature of perceived width in stationary sources, namely that the results extracted from the binaural signals contribute equally to the perception of width, regardless of their time location within the binaural signals. Both the models considered in this chapter do this to a certain extent, but only by weighting the results by the loudness of the signals, thus reducing the influence of the silences and near-silences in the binaural signals. In the literature Apparent Source Width (ASW) has been associated with spaciousness and early lateral reflections (typically the reflections in the first 80ms) [25]. This suggests that the part of the sound which provides the principal cues for the perception of source width is related to onsets of the sound (e.g. the start of notes in music or start of words in speech). This could be investigated by incorporating the onset detector from Supper's model, as has previously been discussed in Section 5.3.3 in the directional localisation chapter.

6.4 Summary

This chapter has described investigations into two computational models for predicting perceived source width. The two models were the modified Supper model (discussed in detail in Chapter 5) and Mason's width model [98]. The stimuli from the third listening test, described in Section 4.4, were input into the models and their results were compared against the width results from the listening test. The two models gave similar results, with the Mason model appearing to be slightly better at predicting the listening test results. This was followed by a discussion of possible changes to both the design of the width listening test and the design of the models that may improve the prediction of perceived source width and also other areas of potential future work.
Chapter 7

Envelopment

The majority of the literature regarding the perception of envelopment has been written from the perspective of concert hall acoustics [12, 24, 25]. In these the listener envelopment (LEV) is primarily related to the reverberation characteristics of the hall, in particular the reflections arriving 80ms or more after the direct sound. More recently the subject of envelopment in relation to reproduced audio has also been researched [14, 66, 140, 151]. One of the more striking differences of the treatment of envelopment between these two approaches is that with reproduced audio the perception of envelopment is not limited to the effect of late reflections and the implicit assumption that the original sound sources are located in front of the listeners on a stage, as is typical of real concert halls. In relation to reproduced audio, the treatment of envelopment is expanded to include the sense of envelopment from an individual source and the sense of envelopment resulting from being surrounded by a number of different sources [66, 140].

Griesinger [66] discusses how human listeners do not use the same method for perceiving these different types of envelopment. He proposes that listeners perceive a foreground stream consisting of the direct sound and a background stream consisting of the part of the sound corresponding to the reverberation of the original environment. He also states that when there is more than a single source, e.g. two different voices, then each source is assigned a separate foreground stream. Note that Griesinger's use of the terms "foreground stream" and "background stream" differs from their use in the speech recognition community. Griesinger uses the two terms to distinguish between two types of auditory stream: those corresponding to direct sound and those corresponding to the indirect, reflected sound. In contrast, in speech recognition the term "foreground stream" refers to the auditory stream containing the speech which is being translated into words, while the term "background stream" refers to the collection of all the other auditory streams which the listener is not giving their attention to. Note that when used in this sense, the background stream may include auditory streams corresponding to direct sound. One of the reasons for grouping all the unattended auditory streams into a single "background stream" in speech recognition is that studies by Cherry [33] and Brochard et al. [29] have shown that these unattended streams are undifferentiated by the listener.
Griesinger's argument informed Conetta's decision to consider separately two types of envelopment in his listening test experiments [38]. The first type is termed direct envelopment and is the sense of envelopment resulting from being surrounded by a number of different sources. Accordingly, the stimuli used by Conetta in his direct envelopment listening tests were created by panning anechoic speech recordings to different positions around the listener. The second type of envelopment is termed indirect envelopment and is the sense of envelopment arising from reverberation. The stimuli used by Conetta in his indirect envelopment listening tests were created by convolving anechoic speech recordings with the impulse responses from a very reverberant hall.

This chapter describes an investigation into using the output of the modified Supper model to predict envelopment. The results of Conetta's listening test experiments were used to calibrate and validate the envelopment prediction models, and, because of this, the same distinction between direct and indirect envelopment was also used. The investigation described in this chapter is divided into three main areas. The first area is the investigation of how the localisation of the individual sources that comprise more complicated stimuli can be used to predict direct envelopment. The second area is the investigation of how the modified Supper model can be used with complex, multi-source stimuli to generate metrics for the prediction of direct envelopment. This is repeated in the third area of investigation for indirect envelopment. In all three areas the metrics developed from the modified Supper model were used in regression models along with the metrics developed by Conetta [38]. The chapter concludes with a discussion of the results in the context of the literature.

7.1 The data set for direct envelopment

The data set for the perception of direct envelopment was collected in two listening test experiments designed, supervised and analysed by Conetta [38]. This section briefly describes these two listening tests.

The same eight Bang and Olufsens Beolab 3 active loudspeakers that were used in the second and third listening test experiments described in Chapter 4 were used for the direct envelopment listening tests. These were arranged in an equally spaced circular array with radius 2.2m, as shown in Fig. 7.1. The signals to the loudspeakers were controlled from the Max/MSP environment on an Apple Mac laptop computer via a Fireface digital audio interface, again as in the second and third listening test experiments in Chapter 4 (see Fig. 4.10). An acoustically transparent curtain was suspended inside the loudspeaker array to conceal the loudspeaker positions from the listening test subjects.

The majority of the stimuli in the direct envelopment listening tests were created using combinations of constant power pair-wise panned anechoic recordings of speech; the remainder were 'real' program material with varying degrees of envelopment. The stimuli used in the two direct envelopment listening tests are described in detail in Section II.1 in Appendix II. Conetta used a novel multi-stimulus test paradigm in his direct envelopment listening tests. This was based on the MUltiple Stimuli with Hidden Reference and Anchor method (MUSHRA) [132], but with the major difference of using two explicit anchors to define the test scale rather than using a hidden reference and hidden anchor. The two explicit anchors were positioned
Figure 7.1: The equipment layout used in the direct envelopment listening test. Eight loudspeakers were arranged in an equally spaced circular array with radius 2.2m. An acoustically transparent curtain was suspended inside the loudspeaker array to hide the positions of the loudspeakers from the listening test subjects.

Figure 7.2: Screen shot of the Conetta's Max/MSP user interface from his envelopment listening tests, using the MUltiple Stimuli with Two Explicit Anchors (MUSTEA) method.
near the top and bottom of the scale (85 and 15 on a 100 point scale). The user interface was created in Max/MSP and is shown in Fig. 7.2. On each page of the user interface the test subject selects the stimuli to listen to and grades their perceived envelopment using sliders. The buttons allowing the test subjects to hear the two explicit anchors can be seen marked A and B on the left-hand side of Fig. 7.2. Conetta’s novel test paradigm will be referred to as the MUltiple Stimuli with Two Explicit Anchors method (MUSTEA). The listening test consisted of fourteen pages on the user interface, including two familiarisation pages. Nineteen listeners participated in the first direct envelopment experiment and twenty listeners participated in the second experiment. All the subjects were members of the Institute of Sound Recording at the University of Surrey and had experience of critical assessment of reproduced audio.

7.2 Investigating the relationship between direct envelopment and directional localisation

Existing measures of listener envelopment include the lateral hall gain, LG (see Section 2.4.2), proposed by Bradley and Soulodre [25], which measures the relative level of the late lateral reflections (after 80ms). Also, Griesinger states that “the optimal sound direction for envelopment is 90°” [66]. These both suggest that the directional localisation of a sound (or at least the directional localisation of the late reflections of the sound) is an important factor in the perception of envelopment. Direct envelopment has been defined at the beginning of the chapter as the sense of envelopment arising from being surrounded by a number of (non-reverberant) sources, and hence the stimuli used in the MUSTEA listening tests described in Section 7.1 were mainly generated by panning dry sources to positions around the listener. All of this informed the decision to investigate how listeners localise the individual sources in these stimuli, as this may help to improve the understanding of the nature of perceived envelopment. Another reason for undertaking this investigation is that the modified Supper model has already been shown to give good predictions of directional localisation. If there is a link between directional localisation and envelopment, then, potentially, the modified Supper model could be used to predict envelopment as well as localisation.

Human listeners have the ability to focus their attention on a single talker when presented with multiple speech sources simultaneously. This is known as the “cocktail party effect” [18]. Cherry [33, 34] proposed a number of cues used by real listeners to separate different speech sources. These include the localisation of different voices, visual cues such as lip-reading, the different speech characteristics of voices such as mean pitch and mean speed and the transition probabilities, which enable listeners to predict word sequences [144]. The problem of perceptual grouping, i.e. separating different audio stimuli into perceptual streams corresponding to their components, is not confined to just speech. The task of solving this more general problem is referred to as audio scene analysis, and the research into this has been summarised by Bregman [28]. As in the more specific problem of the cocktail party effect, which only involves speech, a number of different cues are used by human listeners in the segregation of sounds. These include: spatial continuity, where different localisations correspond to different streams; temporal continuity, where continuous sounds are more likely to be grouped together as a single stream; visual cues; history, where knowledge of previous cues influences
the segregation; harmonics, where tones belonging to the set of harmonics of some fundamental frequency are more likely to be classed as a single stream.

In order to be able to investigate the relationship between direct envelopment and directional localisation it is necessary to predict the localisation of the different sources which comprise each stimulus. This requires segregating each stimulus into its individual components, which may be all sounding simultaneously. This is not part of the functionality of the modified Supper model, and indeed the problem of audio scene analysis is beyond the scope of the research described in this thesis.

For the purposes of the investigation described in this section the problem of segregating the stimuli was overcome in the following manner. First, the assumption was made that the source segregation performed by real listeners could be approximated by using a perfect segregation, i.e. the sources were assumed to be perceived in the same way whether they were heard separately or together. Secondly, the stimuli considered in the investigation were limited to those which were artificially constructed by panning anechoic recordings to different locations around the listener. This meant that, with access to the component signals used to create these stimuli and also the knowledge of exactly how the stimuli were constructed, it was possible to decompose each stimulus into its component sources. This is effectively performing a perfect segregation on the stimuli. Each individual source was then input separately to the modified Supper model, giving a set of localisation angles for the component sources in each stimulus.

7.2.1 The data set for the investigation of direct envelopment and directional localisation

The data set for the investigation of the relationship between directional localisation and direct envelopment was collected in a listening test experiment jointly designed with Rob Conetta, who ran the experiments and also analysed the data. This section briefly describes this listening test.

The same equipment setup was used as in the MUSTEA listening tests undertaken to investigate direct envelopment which are described in Section 7.1 and shown in Fig. 7.1. The only change was the addition of a circular scale with angles marked at 22.5° intervals starting at 0°. This scale was placed inside the acoustically transparent curtain and was intended to aid the test subjects with directional localisation of the perceived sources that comprise the stimuli. Sixteen stimuli were presented to the test subjects. The first eight of these were a subset of the stimuli used in the first MUSTEA direct envelopment listening test. The first-stimulus consisted of eight different voices equally spaced around the circular loudspeaker array and the final eight stimuli in the experiment consisted of the decomposed elements of this first stimulus. In other words, each of the final eight stimuli consisted of a single voice panned to a different position around the loudspeaker array. The stimuli used in this experiment are described in detail in Table II.1 in Appendix II. The user interface was designed and created by the author in Max/MSP, based on the user interface used in the listening tests described in Sections 4.3 and 4.4. The user interface is shown in Fig. 7.3, and allowed the test subjects to add an arrow for each perceived source in the stimulus and then position the arrows according to the perceived location of each source. Twenty listeners participated in the experiment,
7.2.2 Predicting the localisation of decomposed stimuli

The biggest difference between the decomposed stimuli from Conetta's direct envelopment experiments and the stimuli explored in Chapter 5 is that the location of the sources is no longer limited to the front hemisphere, i.e. Conetta's decomposed stimuli include sources behind the head. This presents a problem to the modified Supper model, as it is unable to differentiate between sources in front of the listener and sources behind the listener. For instance, the model is unable to differentiate between a source at 45° and a source at 135°. Indeed, this problem was acknowledged in the original Supper model, where the IID and ITD cues from the rear hemisphere were folded into the front hemisphere when creating the look-up tables (see Section 5.1.3).

Griesinger [66] states that there are two sets of cues used by listeners to differentiate between sounds originating from the front and sounds originating from the rear. The first set consists of spectral cues. The frequency response of the HRTF has notches around 8kHz for sounds located in front of the listener. These notches disappear as the source moves round to the side to ±90°. As the source continues to move round the listener, notches at 5kHz appear when the source reaches ±150°. These notches are the primary cues for differentiating between front and rear sources for sounds with high frequency content. The second set of cues used to disambiguate front and rear sources are caused by small head movements. As the direction of the head changes in relation to the source, the IID and ITD cues also change. The form that these changes...
take can be used to determine whether the source is located in front of or behind the listener. The modified Supper model does not use spectral cues so it cannot use the notches in the frequency spectrum of the binaural signals to disambiguate front and rear sources. However, in the context of the model framework, which includes modelling the transmission of the sound from the loudspeakers to the listener’s ears, the orientation of the head can be changed. This means that small head movements can be incorporated into the model to allow front/rear disambiguation.

Modelling head movements over the duration of a stimulus is complex, requiring a smooth transition of HRTFs as orientation of the head changes (see [71]). An alternative to modelling continuous head movement is to allow the artificial listener to hear the stimulus twice, first with the original head orientation and secondly with the head orientation rotated by small known amount. This gives two sets of results from the modified Supper model where the difference between these two sets of results can be used to determine whether the source is in front of or behind the listener. This considerably simplifies the computational task. This method was implemented with a shift of 5° to the left between successive occurrences of each stimulus. The angle of 5° was chosen with the intention that it would be small enough to be representative of a head movement typical or a real listener, yet also large enough for the differences between the two sets of results to be large enough to allow the differentiation of front and rear sources.

Figs. 7.4 and 7.5 illustrate this method. Fig. 7.4 shows the results from the modified Supper model for the ninth stimulus from Conetta’s direct envelopment localisation experiment. This stimulus consisted of an anechoic recording of a voice panned to -22.5° (i.e. in front of the listener). The top graph in Fig. 7.4 shows the AOT histograms for the two head positions, with the black and red lines corresponding to the 0° and -5° head orientations respectively. Note that both of these graphs are plotted relative to the modelled head and that the peak of the histogram moves to the right when the head orientation is shifted by 5° to the left. The bottom two graphs in Fig. 7.4 shows the same histograms plotted radially relative to the listening environment. As the modified Supper model is unable to differentiate front and back sources with a single head position, every angle θ output by the model either really is θ (i.e. the source is in front of the listener) or else is really 180° − θ (i.e. the source is behind the listener). The bottom left graph in Fig. 7.4 shows the hypothesis that the source is in front of the listener. In this case the two peaks agree with each other, lining up in the same direction when both plotted relative to the listening environment. This confirms the hypothesis that the source is located in front of the listener. Conversely, the bottom right graph shows the hypothesis that the source is behind the listener. In this case the two peaks diverge when the results are plotted relative to the listening environment. This shows that the hypothesis that the source is located behind the listener is false.

Fig. 7.5 shows corresponding results for the twelfth stimulus from Conetta’s direct envelopment localisation experiment. This stimulus consisted of an anechoic recording of a voice panned to -157.5° (i.e. behind the listener). Note in the top graph of Fig. 7.5 that this time the peak of the histogram moves to the left when the head orientation is shifted by 5° to the left. The bottom left plot in Fig. 7.5 shows the hypothesis that the source is in front of the listener. Here the two peaks diverge, disproving this hypothesis. Conversely, the two peaks coincide in the bottom right plot, confirming the hypothesis that the source is located behind the listener.
Together Figs. 7.4 and 7.5 confirm Griesinger's statement about the use of head movements and IID and ITD cues to disambiguate front and back sources. From these figures it can be seen that when the head orientation is rotated 5° to the left then if the source is located in front of the listener then the peak of the histogram moves to the right (see Fig. 7.4) and if the source is located behind the listener then the peak of the histogram moves to the left (see Fig. 7.5). A variant on this method that considered only sinusoidal signals was proposed by van Soest [18, 147].

Figs. 7.6 and 7.7 show the results of localising the decomposed sources for the stimuli from Conetta's direct envelopment experiment. The method described above was used to determine whether each source was located in the front or rear hemisphere. It can be seen from the plots in these two figures that this method was successful for all the decomposed sources in the first six and last eight stimuli. In the results for the first four stimuli it can be seen that the angles obtained using the peak-picking algorithm from the output of the modified Supper model tend to be pulled towards ±90° compared with the intended pan positions. Note that the test subjects found the sources close to ±90° harder to localise, as can be seen by the number of listener responses at ±90° for stimulus one in Fig. 7.6 and also stimuli eleven and fifteen in Fig. 7.7. Also note that the model was less successful at localising the sources in the seventh and eighth stimuli. These two stimuli both contained correlated signals being played out of multiple loudspeakers, and in both cases the majority of the listener responses were also not at the positions at which the signals were panned. In both cases the angle output by the model matched a peak in the listener responses.
Figure 7.4: The AOT histograms output by the modified Supper model for the ninth stimulus from Conetta's direct envelopment localisation experiment. The stimulus consisted of an anechoic recording of a voice panned to -22.5°. In the top graph, the black histogram corresponds to the modelled head orientated to 0° (facing straight ahead) and the red histogram corresponds to the modelled head orientated to -5° (facing slightly to the left). The bottom two plots show radial plots of the same AOT histograms. The bottom left plot shows the case where the source is assumed to be in the front hemisphere and the bottom right hand plot shows the case where the source is assumed to be in the rear hemisphere. Again, the black and red lines correspond to head orientations of 0° and -5° respectively. The blue lines in the bottom two plots show the intended position of the source, an anechoic recording of a voice panned to -22.5°.
Figure 7.5: The AOT histograms output by the modified Supper model for the twelfth stimulus from Conetta's direct envelopment localisation experiment. The stimulus consisted of an anechoic recording of a voice panned to -157.5°. In the top graph, the black histogram corresponds to the modelled head orientated to 0° (facing straight ahead) and the red histogram corresponds to the modelled head orientated to -5° (facing slightly to the left). The bottom two plots show radial plots of the same AOT histograms. The bottom left plot shows the case where the source is assumed to be in the front hemisphere and the bottom right hand plot shows the case where the source is assumed to be in the rear hemisphere. Again, the black and red lines correspond to head orientations of 0° and -5° respectively. The blue lines in the bottom two plots show the intended position of the source, an anechoic recording of a voice panned to -157.5°.
Figure 7.6: The results of localising the decomposed sources for the first eight stimuli from Conetta's direct envelopment localisation experiment. The top row contains the results for stimuli numbers one to three, the middle row contains the results for stimuli numbers four to six and the bottom row contains the results for stimuli numbers seven and eight. The grey histograms show the normalised results from the listening test, scaled so the area of each bar is proportional to its value. The blue lines show the panned positions of the sources and the red lines show the angles output by the modified Supper model.
Figure 7.7: The results of localising the decomposed sources for the last eight stimuli from Conetta’s direct envelopment localisation experiment. The top row contains the results for stimuli numbers nine to eleven, the middle row contains the results for stimuli numbers twelve to fourteen and the bottom row contains the results for stimuli numbers fifteen and sixteen. These eight stimuli are the decomposed sources for stimulus number one. The grey histograms show the normalised results from the listening test, scaled so the area of each bar is proportional to its value. The blue lines show the panned positions of the sources and the red lines show the angles output by the modified Supper model.
7.2.3 Metrics for the prediction of envelopment based on decomposed stimuli

The previous section showed that once the stimuli have been decomposed into their individual elements then the modified Supper model seems to be able to predict the results of the directional localisation envelopment experiment. Although half the stimuli in the directional localisation envelopment experiment were also used in Conetta's first MUSTEA direct envelopment listening test, the test subjects were not asked to judge the envelopment, instead they were asked to identify the directions of the sources in the stimuli. Now, the aim was to investigate the relationship between directional localisation and direct envelopment. Using the modified Supper model on the decomposed stimuli gives a collection of localisation angles for each stimulus, while the results from Conetta's MUSTEA listening tests are in the form of a single value for the perceived envelopment of each stimulus. This means that a method of transforming a collection of angles into a single value for the prediction of envelopment was required. This section describes two measures created from the output of the modified Supper model together with an investigation into their performance as predictors of direct envelopment.

The first of these measures was calculated using the following method. First a circle was drawn, centred on the listening position. Secondly, a line was drawn from the centre of the circle to its circumference for each localisation angle output by the modified Supper model (this is illustrated by the red lines in Fig. 7.8). The set of points where these radial lines cross the circumference is then determined. The convex hull of this set is the smallest polygon containing all its points [126]. In the case of this set of points this is the polygon formed by joining each point in the set to its neighbours on the circumference, as illustrated by the green lines on the two graphs in Fig. 7.8. Finally, the measure is calculated as the ratio of the area of the convex hull to the area of the enclosing circle. This will be referred to as the hull measure.

Two of the most important criteria for an effective measure of direct envelopment are that it must have a low value when a stimulus is not enveloping and have a high value when a stimulus is enveloping. When a stimulus consists of different sources positioned around the listener, as illustrated in the left hand plot in Fig. 7.8, then the hull metric will be close to one. Conversely, a single source in front of the listener is judged to have a low value of direct envelopment, and the convex hull will consist of a single point, giving the hull metric a value of zero.

The second of the measures created from the output of the modified Supper model was based on the observation of Conetta's direct envelopment listening test results that listeners find sources positioned close to ±90° (i.e. to the sides of the head) to be more enveloping than sources positioned close to 0° (i.e. in front of the head). This is supported by the literature on the importance of lateral reflections to the sense of envelopment in concert hall acoustics [12, 24, 25]. This measure is calculated from the set of angles output by the modified Supper model from the decomposed stimuli. The measure consists of the absolute value of the angle from this set which is closest to 90°. This measure will be referred to as the c90 metric. When a stimulus consists of different sources positioned around the listener, then the sources positioned to the sides of the head will ensure that the c90 value has a high value. Conversely, a single source in front of the listener, which has a low value of direct envelopment, will be localised by the model close to 0°, giving a low
Figure 7.8: The two graphs illustrate the calculation of the hull metric from the set of localisation angles output by the modified Supper model acting on a decomposed stimulus. The left- and right-hand graphs correspond to the first and third stimuli respectively from Conetta’s direct envelopment localisation experiment. The grey histograms show the normalised results from the listening test, scaled so the area of each bar is proportional to its value. The blue lines show the panned positions of the sources and the red lines show the angles output by the modified Supper model. The green lines show the boundary of the convex hull: the hull metric is calculated as the ratio of the area of the convex hull to the area of the enclosing circle.

c90 value.

The results of using these two measures to each predict the results from Conetta’s first MUSTEA direct envelopment listening test experiment are shown in Fig. 7.9. Only the results corresponding to the first twenty-five (i.e. the artificially constructed) stimuli were included. Attempting to decompose real program material where there is no access to the individual components that comprise the stimulus is beyond the scope of this project. The values of the hull and c90 metrics applied to these twenty-five stimuli were fitted using linear regression to the listening test results. For both metrics it can be seen that on their own they are best at differentiating the stimuli at the upper end of the direct envelopment scale. Both metrics are unable to differentiate between stimuli with low values of direct envelopment. Despite this, both metrics show reasonably high R values and low RMSEP values.

The introduction at the beginning of this chapter included a discussion of the fact that envelopment is a more complicated perceptual attribute than both directional localisation and source width. This suggests that a relatively high-level cognitive process is used to fuse a number of lower-level psychoacoustic cues to create the perception of envelopment. Because of this it is unlikely that a single measure based on directional localisation will be able to predict the behaviour of all the results seen in Conetta’s direct envelopment listening test experiments. Indeed, Conetta used five different metrics based on acoustic and...
psychoacoustic measurements to build a regression model to predict the results from his listening tests: IACCO, IACCO*IACC90, TotEnergy, EntropyL and CardKLT. The IACCO is calculated from the binaural signals when the listener is orientated towards 0° (i.e. facing forwards). Mason’s width model [98] is used to calculate twenty-two time series of IACC values, corresponding to twenty-two critical bands. The minimum IACC value is found for each critical band and the final IACCO measure is calculated by taking the mean of these twenty-two minimum values. The IACC90 is calculated in an identical manner except that the binaural signals are generated with the listener orientated towards 90° (i.e. to the right). The IACCO*IACC90 metric is simply the product of the IACCO and the IACC90. The TotEnergy metric is calculated as the root mean square of the signal captured by an omni microphone positioned at the centre of where the listener’s head would be. The EntropyL metric is calculated as the information entropy [144] of the left ear binaural signal when the listener is orientated towards 0°. The CardKLT is the largest eigenvalue calculated by decomposing the signals from four cardioid microphones (positioned at the centre of where the listener’s head would be and facing 0°, ±90° and 180°) into their principal components.

It was decided to investigate how the hull and c90 metrics performed in the context of a linear regression model which also used Conetta’s five existing metrics. Table 7.1 shows a summary of the results of this investigation. The rows of the table contain, from top to bottom, results for: the feature set containing all five of Conetta’s metrics, the two best performing feature sets containing containing four metrics, the two best performing feature sets containing three metrics, and the best performing feature set containing only two metrics. The performance of each feature set was judged on the R² and RMSEP values. The best
performing feature sets were found to have both the highest $R^2$ values and the lowest RMSEP of the feature sets being considered. Note that the hull metric does belong to any of the feature sets contained in the table.

In the table it can be seen that the two feature sets with four metrics perform almost as well as the feature set with all five of Conetta's metrics. The performance is still very good when the feature sets are reduced to three metrics. From the table it appears that the feature sets containing the c90 metric perform similarly to the feature sets containing only Conetta's metrics. The final row in Table 7.1 shows that most of the variance in the listening test results for the artificial stimuli can be explained using only two of Conetta's metrics: IACC0*IACC90 and TotEnergy.

As in the previous results calculated using the c90 and hull metrics, only the twenty-five artificially constructed stimuli from Conetta's first MUSTEA direct envelopment listening test experiment were used. This is significant when interpreting the results for two reasons. The first is that the artificial stimuli in the first direct envelopment experiment are very symmetrical and are very similar in the way that they are constructed. This means that some variations in the properties of the stimuli may not be present. Hence, predicting just the artificial stimuli may be easier than predicting all thirty stimuli, giving better results than if the models were used on a more varied set of stimuli. The second reason is that Conetta chose his metrics when using all thirty of the stimuli, including five real program items, which by their nature are more representative of real world stimuli than the twenty-five contrived stimuli. This means that some of Conetta's metrics may not appear to contribute much to the performance of the regression models when just the artificial stimuli are used.

Indeed, this can be seen when comparing the results in the first and third rows of Table 7.1. Both of these feature sets comprise only metrics taken from Conetta, the only difference being that EntropyL is omitted.
from the feature set for the results in the third row. This only resulted in a slight increase to the RMSEP value. One possible explanation for this is that all the stimuli were loudness equalised, so the total energy (TotEnergy) will only vary across the artificial stimuli depending on the amount of voices present in each stimulus, which is closely correlated to the entropy (EntropyL) of each artificial stimulus. The correlation between EntropyL and TotEnergy is likely to decrease when the five real program items are included, as the sources in these five stimuli have more varied characteristics than the anechoic voices used in the creation of the artificial stimuli.

Fig. 7.10 shows the results of the two feature sets containing four metrics from Table 7.1. The only difference between these two feature sets is that the c90 metric in the feature set for the top two plots is replaced by the IACCO metric in the feature set for the bottom two plots. The graphs on the left, showing the mean listening test results plotted against the predicted envelopment, are very similar in both graphs. The graphs on the right show the standardised coefficients of the metrics for the first and second principal components in the regression models. These show that, while all the metrics contribute to the first principal component, the major contributor to the second principal component is TotEnergy for both feature sets. In both cases the coefficients for TotEnergy are close to being orthogonal to the other three metrics.
Figure 7.10: These graphs show the results of linear regression models using the two feature sets from Table 7.1 containing four metrics fitted to the results of Conetta’s first direct envelopment listening test experiment. Only the results corresponding to the first twenty-five (i.e. the artificially constructed) stimuli were included. The graphs in the top row used the feature set containing the IACC0*IACC90, TotEnergy, CardKLT and c90 metrics, while the graphs in the bottom row used the feature set containing the IACC0, IACC0*IACC90, TotEnergy and CardKLT metrics. The graphs in the left column show the mean listening test results plotted against the predicted envelopment. The graphs in the right column show how each of the metrics contribute to the two largest principal components for each of the regression models.
7.3 Predicting direct envelopment with complete stimuli

The previous section showed how metrics based on the modified Supper model using decomposed stimuli were able to predict direct direct envelopment when used in conjunction with some of the metrics developed by Conetta. The major disadvantage of using this approach is that each stimulus needs to be decomposed into its constituent sources. As discussed in the previous section, to be able to do this blind (i.e. with no prior knowledge of the component sources) is a relatively complicated task and is the subject of ongoing research [28, 137]. This is beyond the scope of the research described in this thesis. Without this, however, the stimuli that can be used with the decomposed-signals approach is limited to those stimuli where the component sources are already available. This was seen in the previous section where the twenty-five artificially constructed stimuli from Conetta's first direct envelopment listening test experiment were used in the regression models, but it was not possible to include the five stimuli consisting of real program material. Therefore, in order to investigate the prediction of direct envelopment for more general stimuli requires metrics that take complete, non-decomposed signals as their input. This section describes the development of metrics using the modified Supper model that satisfy this condition.

Of the metrics developed from the modified Supper model described in the previous section, the c90 metric was seen to have the better performance of the two when combined with Conetta's metrics. This suggests that a metric based on the same idea (i.e. finding the source closest to ±90°) but using complete, non-decomposed signals as its input may also prove to be a good predictor of direct envelopment. Fig. 7.11 shows the averaged over time (AOT) histogram output by the modified Supper model for stimulus 8 from Conetta's first MUSTEA direct envelopment listening test experiment. This stimulus consisted of eight different speech signals, constant power panned so there are two different voices in each of the following positions: -45°, 45°, -135° and 135°. The most notable feature of this histogram is that the maximum peak is located at 0°, which does not correspond to any of the positions of the panned voices. It can also be seen that the histogram has smaller maxima (e.g. near -60°) in addition to the main peak. This behaviour is common when there are multiple sources at different locations present inn the stimulus: the AOT histogram is more spread, there are more smaller peaks and the maximum peak is most often located at 0° regardless of the positions of the individual sources, particularly when the component sources have left-right symmetry about the listener.

This means that the method of averaging the histograms output by the modified Supper model is not suitable for a metric based on identifying the component source closest to ±90°. However, from inspecting the series of histograms output by the model it is apparent that there are time frames where the model is able to identify the locations of one of the component sources (which one of the component sources depends on the time frame). Furthermore, the time frames where the model is able to locate a component source also correspond to the histograms with large peaks. This is because each histogram shows the degree of likelihood that the signal is located at each angle given the IID and ITD cues. Therefore, when the IID and ITD cues are in agreement about the location of the signal source in a time frame then the likelihood of the signal being in the corresponding location in that time frame is high, resulting in a high maximum peak in the histogram. This will tend to be when one of the component sources is louder than the others. From this it
Figure 7.11: The averaged over time (AOT) histogram for the combined (tid*tid) output of the modified Supper model for stimulus 8 from Conetta's first MUSTEA direct envelopment listening test experiment. This stimulus consisted of eight different speech signals, constant power panned so that each of the following positions contained two different voices: -45°, 45°, 135° and -135°. Note that the maximum peak is located at 0° and does not correspond to any of these positions.

follows that a better indication of the locations of individual sources can be obtained by considering only those histograms with large values for their maximum peaks. This can be achieved by applying a threshold to the histograms, so a subset of the complete time series of histograms is generated by selecting only those histograms whose maximum peaks lie above this threshold. Angles for this subset of output histograms can then be calculated using the peak-picking algorithm, and from this set of angles the angle closest to 90° can be determined as the final value of the metric.

More formally, let \( h_{t,\theta} \) be the value of the output histogram at time \( t \) and at the angle \( \theta \), where \( \theta \) is an integer in the range -90° to 90°. This gives the output histogram for time frame \( t \) as

\[
H_t = \begin{bmatrix}
  h_{t,90} \\
  h_{t,89} \\
  h_{t,88} \\
  \vdots \\
  h_{t,-88} \\
  h_{t,-89} \\
  h_{t,-90}
\end{bmatrix}
\]  \quad (7.1)

the peak value of this histogram as

\[
p_t = \max_{\theta} (h_{t,\theta})
\]  \quad (7.2)

and the corresponding angle (using the peak-picking method) as

\[
\phi_t = \arg \max_{\theta} (h_{t,\theta})
\]  \quad (7.3)
Let $T$ be the set of all values of $t$. Let $T_\beta$ be the subset of $T$ such that

$$p_t \geq \beta \quad \forall t \in T_\beta$$

(7.4)

Here $T_\beta$ is the set of time frames such that the peak value of all the histograms for these time frames are above the threshold $\beta$. Now, the values of $\phi_t$ are integers in the range $-90^\circ$ to $90^\circ$, so the final value of the metric is calculated as the maximum value of $|\phi_t|$ for values of $t$ in the set $T_\beta$:

$$c90_\beta = \max_{t \in T_\beta} |\phi_t| = \max_{t \in T_\beta} \left| \arg \max_{\theta} (h_{t, \theta}) \right|$$

(7.5)

Six different values of $\beta$ were used as the threshold, 5, 10, 20, 30, 50 and 100, giving six different $c90$ metrics: $c90_5$, $c90_{10}$, $c90_{20}$, $c90_{30}$, $c90_{50}$ and $c90_{100}$.

The second set of metrics derived from the output of the modified Supper model have already been described in Section 6.1.1. These are based on the standard deviation of the AOT histograms. As in Section 6.1.1, four different metrics were used, differing in the histograms used to generate the AOT histogram from which the standard deviation was calculated. These are shown in Table 7.2.

<table>
<thead>
<tr>
<th>Metric name</th>
<th>Model results used</th>
</tr>
</thead>
<tbody>
<tr>
<td>sd_iid</td>
<td>IID</td>
</tr>
<tr>
<td>sd_itd</td>
<td>ITD</td>
</tr>
<tr>
<td>sd_plus</td>
<td>IID + ITD</td>
</tr>
<tr>
<td>sd_times</td>
<td>IID * ITD</td>
</tr>
</tbody>
</table>

Table 7.2: The table shows the four metrics calculated by taking the standard deviation from an AOT histogram and the model results from which each AOT histogram was generated.

These two sets of metrics were then combined with Conetta’s five metrics using linear regression to investigate how these metrics are able to predict direct envelopment, in a similar manner to that used in the previous section. Matlab was used to loop over all the different combinations of five or fewer metrics and the $R^2$ and RMSEP values for corresponding linear regression models were calculated for each of these combinations. In addition, $R^2$ and RMSEP values were calculated for a leave-one-out cross-validation [156, 56] performed on each regression model. Table 7.3 contains a summary of the results of this process.

The table includes the three best performing feature sets with five metrics, the three best performing feature sets with four metrics, the four best performing feature sets with three metrics, the best performing feature set with two metrics and also the best performing single metric. The performance of each feature set was judged on the calculated $R^2$ and RMSEP values, both for the regression model trained on all the data and also for the cross-validation. Those feature sets whose linear regression models had individual Variance Inflation Factor (VIF) values greater than 10 or average VIF values greater than 6 were judged to have a problem with multicollinearity and were omitted from the table (see Table I in Appendix I). Note that in Table 7.3 the feature sets that had the best $R^2$ and RMSEP scores for the complete regression model also had the best cross-validation scores.

CHAPTER 7. ENVELOPMENT | 225
### Table 7.3: The results of applying linear regression to the data from Conetta's two MUSTEA direct environment listening test experiments. The second column contains the names of the metrics used in the regression, each followed by its standardised coefficient. The third and fourth columns contain the $R^2$ and RMSEP values of the regression and the fifth and sixth columns contain the $R^2$ and RMSEP values for the cross-validation.

<table>
<thead>
<tr>
<th>Number of metrics</th>
<th>Metrics and standardised coefficients</th>
<th>All data</th>
<th>X-Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R^2$</td>
<td>RMSEP</td>
</tr>
<tr>
<td>5</td>
<td>IACCO (-0.60), EntropyL (0.42), TotEnergy (0.25), c90.30 (0.36), sd_iid (-0.31)</td>
<td>0.90</td>
<td>7.18</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.61), EntropyL (0.45), TotEnergy (0.31), CardKLT (-0.27), sd_iid (-0.11)</td>
<td>0.89</td>
<td>7.42</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.46), EntropyL (0.48), TotEnergy (0.32), CardKLT (-0.23), IACCO*IACC90 (-0.20)</td>
<td>0.89</td>
<td>7.45</td>
</tr>
<tr>
<td>4</td>
<td>IACCO (-0.61), EntropyL (0.46), TotEnergy (0.30), CardKLT (-0.26)</td>
<td>0.88</td>
<td>7.78</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.56), EntropyL (0.50), TotEnergy (0.29), IACCO*IACC90 (-0.27)</td>
<td>0.86</td>
<td>8.23</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.75), EntropyL (0.49), TotEnergy (0.28), sd_times (-0.12)</td>
<td>0.86</td>
<td>8.46</td>
</tr>
<tr>
<td>3</td>
<td>IACCO (-0.79), EntropyL (0.49), TotEnergy (0.26)</td>
<td>0.84</td>
<td>8.80</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.66), EntropyL (0.51), CardKLT (-0.17)</td>
<td>0.79</td>
<td>10.11</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.77), EntropyL (0.52), sd_iid (0.12)</td>
<td>0.79</td>
<td>10.13</td>
</tr>
<tr>
<td></td>
<td>IACCO (-0.70), EntropyL (0.51), c90.30 (0.13)</td>
<td>0.79</td>
<td>10.16</td>
</tr>
<tr>
<td>2</td>
<td>IACCO (-0.77), EntropyL (0.52)</td>
<td>0.78</td>
<td>10.46</td>
</tr>
<tr>
<td>1</td>
<td>IACCO (-0.71)</td>
<td>0.51</td>
<td>15.63</td>
</tr>
</tbody>
</table>
It can be seen that the IACC0 metric is present in every feature set in the table. Indeed, the IACC0 accounts for over half of the variance seen in the listening test results. Similarly, the combination of IACC0 and EntropyL is included in every feature set with two or more metrics. The combination of IACC0, EntropyL and TotEnergy is included in the best performing feature set with three metrics and all the features with four or more metrics. This combination accounts for 84% of the variance in the listening test results; the best performing feature set with five metrics only managed to explain another 6% of the variance and lower the RMSEP from 8.80 to 7.18 (less than 1.7%, as the direct envelopment was measured on a hundred-point scale). The best performing feature set with five metrics included two of the metrics derived from the modified Supper model: the c90.30 metric and the sd_iid metric. Metrics derived from the modified Supper model also appear in four other feature sets in Table 7.3, including the second best performing feature set with five metrics.

In addition to calculating linear regression models using all sixty results combined from both of Conetta's MUSTEA direct envelopment listening test experiments, the feature sets were also trained on the results from the first experiment and validated on the results from the second experiment and vice versa. In both of these cases the IACC0, EntropyL and TotEnergy metrics were found to be present in all the best performing feature sets containing three or more metrics. This is similar to the results seen in Table 7.3 where the models were trained on the data from both experiments. The feature set containing only the IACC0, EntropyL and TotEnergy metrics was also found to give low RMSEP values for the validation in both cases (9.78 when trained on the results from the first experiment and validated on the results from the second and 10.49 when trained on the second and checked against the first). The best performing feature sets were different in both cases. In general the validation R² and RMSEP values were much better when the regression models were trained on the results from the second experiment and validated on the results from the first experiment. This is at least partly due to the fact that the stimuli for the second experiment included a wider variety of signals than the first experiment. For instance, stimulus thirteen from the second experiment consists of four different voices panned to -45° and four different voices panned to 135° (see Table 11.2), which has much less left-right symmetry than any of the stimuli from the first experiment.

7.4 The data set for indirect envelopment

The data set for the perception of indirect envelopment was collected in a listening test experiments designed, supervised and analysed by Rob Conetta. This section briefly describes this listening test. The equipment setup and the user interface were identical to those used in the MUSTEA direct envelopment listening tests described in Section 7.1. The stimuli for this experiment were created using impulse responses for a large reverberant hall calculated in CATT-Acoustics, described in detail in Appendix II in Section II.2 and Table II.4. Twenty listeners participated in the experiment, all members of the Institute of Sound Recording at the University of Surrey.
7.5 Predicting indirect envelopment with the output of the modified Supper model

Conetta found that the same five metrics that he used to predict direct envelopment also gave good predictions of the results from his indirect envelopment experiment when combined in a linear regression model. This suggested that the method and metrics described in the previous section could also be used for indirect envelopment. Again, Matlab was used to loop over all the different combinations of five or less metrics and $R^2$ and RMSEP values calculated for these regression models and also their leave-one-out cross-validation. Table 7.4 contains a summary of these results.

The first two rows of the table contain the two feature sets with the best performing linear regression (i.e. the highest $R^2$ and RMSEP values for all the data points). The third and fourth rows contain the feature sets with five metrics with the best cross-validation scores. The fifth row contains Conetta’s original feature set. The sixth to ninth rows contain a summary of the results for the feature sets with four metrics. The feature sets in the sixth and seventh rows had the best performing regression using all the data points while the feature sets in the eighth and ninth rows had the best cross-validation scores. The tenth to thirteenth rows contain a summary of the results for the feature sets with four metrics. The feature sets in the tenth and eleventh rows had the best performing regression using all the data points while the feature sets in the twelfth and thirteenth rows had the best cross-validation scores. The fourteenth row contains the best performing feature set containing two metrics. This feature set had the best scores using all the data and also the best scores for the cross-validation. The fifteenth row contains the best single metric. Those feature sets whose linear regression models had individual Variance Inflation Factor (VIF) values greater than 10 or average VIF values greater than 6 were judged to have a problem with multicollinearity and were omitted from the table (see Table I in Appendix I).

As in Table 7.3, the IACC0 metric is present in every feature set in the table. However, whereas the IACC0 metric accounted for 51% of the variance in the direct envelopment listening test results, for indirect envelopment the IACC0 metric alone accounts for 72% of the variance in the listening test results. When the IACC0 is combined with one of the metrics derived from the modified Supper model, the sd.ijd, then they are able to explain 84% of the variance in the listening test results. This combination, with only two metrics, performs better than Conetta’s five-metric feature set for the $R^2$, RMSEP and cross-validation. Note that the c90.30 metric only appears once in Table 7.4, and that this is the only c90 metric in the table.

The two best performing feature sets with five metrics are identical except for one having TotEnergy and the other having EntropyL. This occurs for several other pairs of feature sets where the they both perform similarly. The two feature sets with four metrics with the best cross-validation scores are identical except for one having TotEnergy and the other EntropyL. This is also the case with the two feature sets with three metrics with the best scores for the regression and also the two feature sets with three metrics with the best cross-validation scores. The fact that the feature sets in each of these pairs perform similarly to each other and also that the TotEnergy and EntropyL metrics do not appear in any of the feature sets selected for Table 7.4 together suggest that the TotEnergy and EntropyL metrics have highly correlated values for the
<table>
<thead>
<tr>
<th>Number of metrics</th>
<th>Metrics and standardised coefficients</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$R^2$</td>
<td>RMSEP</td>
</tr>
<tr>
<td>5</td>
<td>IACC0 (-0.53), IACC0*IACC90 (-0.35), sd_itd (-0.28), TotEnergy (0.28), CardKLT (-0.24)</td>
<td>0.89</td>
<td>8.47</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.73), sd_iid (-0.26), CardKLT (-0.23), EntropyL (0.21), c90.30 (-0.12)</td>
<td>0.88</td>
<td>8.58</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.60), TotEnergy (0.37), IACC0*IACC90 (-0.24), CardKLT (-0.23), EntropyL (-0.07)</td>
<td>0.82</td>
<td>10.78</td>
</tr>
<tr>
<td>4</td>
<td>IACC0 (-0.71), sd_iid (-0.31), EntropyL (0.19), CardKLT (-0.17)</td>
<td>0.88</td>
<td>8.81</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.67), sd_iid (-0.30), IACC0*IACC90 (-0.27), TotEnergy (0.23)</td>
<td>0.88</td>
<td>8.89</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.77), sd_iid (-0.27), TotEnergy (0.19), CardKLT (-0.18)</td>
<td>0.87</td>
<td>8.95</td>
</tr>
<tr>
<td>3</td>
<td>IACC0 (-0.83), sd_iid (-0.34), EntropyL (0.17)</td>
<td>0.87</td>
<td>9.23</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.89), sd_iid (-0.31), TotEnergy (0.17)</td>
<td>0.86</td>
<td>9.38</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.90), sd_plus (-0.32), EntropyL (0.18)</td>
<td>0.85</td>
<td>9.68</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.95), sd_plus (-0.28), TotEnergy (0.18)</td>
<td>0.85</td>
<td>9.83</td>
</tr>
<tr>
<td>2</td>
<td>IACC0 (-0.82), sd_iid (-0.34)</td>
<td>0.84</td>
<td>10.18</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.89), sd_plus (-0.31)</td>
<td>0.82</td>
<td>10.71</td>
</tr>
<tr>
<td></td>
<td>IACC0 (-0.91), sd_itd (-0.27)</td>
<td>0.79</td>
<td>11.51</td>
</tr>
<tr>
<td>1</td>
<td>IACC0 (-0.85)</td>
<td>0.72</td>
<td>13.27</td>
</tr>
</tbody>
</table>

Table 7.4: The results of applying linear regression to the data from Conetta’s MUSTEA indirect envelopment listening test experiment. The second column contains the names of the metrics used in the regression, each followed by its standardised coefficient. The third and fourth columns contain the $R^2$ and RMSEP values of the regression and the fifth and sixth columns contain the $R^2$ and RMSEP values for the cross-validation.
7.6 Discussion

Regression models have been used previously in research into predicting spatial aspects of reproduced audio. George [57] was primarily concerned with how various degradations affect the perceived quality (including the spatial aspects) of sound reproduced using the FCS loudspeaker format at the sweet spot. A number of the metrics he used in his regression model took the five loudspeaker signals directly as their input. In contrast, Conetta created a more generalisable model of envelopment which was not limited to a single reproduction format and also was not just limited to the sweet spot. Consequently, none of the metrics in Conetta’s feature set take the signals fed to the loudspeakers directly as their input. Instead, all of Conetta’s metrics take either microphone signals or binaural signals as their input. All of these input signals can either be recorded directly in a real reproduced sound-field (using a dummy head in the case of the binaural signals) or else can be simulated using a computational model. As all of these input signals are measurements made on the sound-field, the use of these metrics is not limited to either a single reproduction system or to a single listener position in the reproduced sound-field.

The metrics which are based on the output of the modified Supper model and whose development is described in this chapter have an even more limited set of input signals: the binaural signals corresponding to the orientation of the listener. The design of Supper’s localisation model is very influenced by psychoacoustics: as mentioned before, some of the stages explicitly model processes in human listening, including the use of binaural signals and the separation of these signals into critical bands. This approach is continued into the metrics based on the Supper model described in this chapter, where the inputs to the metrics have been limited to the same cues that would be experienced by a human listener. This contrasts with the development of the metrics in Conetta’s feature set, where any measurements of the sound-field made at the listening position were considered. However, as discussed before, envelopment is a higher level perceptual attribute than both directional localisation and source width. The development of the metrics described in this chapter consisted of interpreting the output of the Supper model. This is effectively taking the low-level perception of direction and then approximating the higher-level cognitive process of interpreting these localisations. As the metrics based on the modified Supper model use the same cues and some of the same mechanisms of interpreting these cues as human listeners, the intention is that they will more accurately model the perceptions of real listeners than some of the more abstract metrics based on sound-field measurements, such as the CardKLT metric. This contrasts with the freer approach used by both Conetta and George, where metrics that did not use a low-level physiological approach or binaural signals were also considered. The main criterion for their choice of metrics being that each metric reflect some aspect of the behaviour observed in the results of the listening tests.
From Table 7.3 it can be seen that the three metrics that appear to make the most important contributions to the regression models for direct envelopment are the IACCO, EntropyL and TotEnergy. These three metrics are members of all the best performing feature sets containing three or more metrics. The feature set comprising just the IACCO and EntropyL metrics account for 78% of the variance in the direct envelopment listening test results, and both of these metrics use as their input the the binaural signals corresponding to the listener orientation (i.e. the binaural signals corresponding to the direction in which the listener is facing, unlike the IACC90 metric, which uses binaural signals corresponding to a head at 90° to the direction in which the test subjects were facing). The third metric, TotEnergy, uses the signal from an omni-directional microphone as its input. It is likely that the values of the TotEnergy metric would not change significantly if the total energy was calculated from one or both components of the components of the binaural signals corresponding to the orientation of the listener. Note that for direct envelopment, the best performing feature set with five metrics consists of four metrics whose inputs are the binaural signals corresponding to the orientation of the listener, with the fifth metric being TotEnergy. This suggests that limiting the metrics in the regression model to those using the same cues as real listeners may actually better reflect the perception of real listeners. This feature set also includes one of each of the two kinds of metric developed from the output of the modified Supper model, namely the e90.30 and the sd_iid metrics. This lends support to the hypothesis that the perception of the listeners can be best modelled by using the same cues as real listeners and also physiologically based low-level processing.

Griesinger [66] identifies three types of spatial impression, of which he considers only two to be related to envelopment. The first of these is the continuous spatial impression (CSI), resulting from the interference of lateral reflected energy with a continuous source such as pink noise. The other two types of impression need the context of discernible audio events, such as individual notes in music or phones from speech. Early spatial impression (ESI) results from the lateral reflected energy within the first 50ms of the end of the audio event, and background spatial impression (BSI) results from the lateral energy after 50ms from the end of the audio event. Griesinger states that, unlike the BSI, the ESI becomes fused with the direct sound and so does not give rise to a sense of envelopment. It is important to note that BSI is partly dependent on being identified as being not the foreground, i.e. the perception is dependent on its relationship with the direct sound. Fig. 7.12 shows the three types of spatial impression.

The stimuli used in the indirect envelopment listening test were designed so that they were typical of the reverberant part of reproduced sound, i.e. the portion of the sound where the late reflections dominate. The stimuli used in the indirect envelopment experiment were so reverberant that the direct sound (the original speech) was difficult to discern. This means that, using Griesinger's classification, the spatial impression would have been somewhere between CSI and BSI. As the direct sound is clearly discernible in the majority of recorded program material, it may be argued that BSI is more important than CSI in the context of judging the quality of reproduced sound. If this is the case then it would be beneficial to repeat the indirect listening test using stimuli where the direct sound was easier to distinguish for the listeners (and any metrics), as this would mean that any cues which depend on the relationship between the direct sound and the reverberant part of the sound would also be present.
7.7 Summary

This chapter has described the development of metrics for the prediction of envelopment from the output of the modified Supper model. The results from listening tests designed and conducted by Conetta were used to investigate the performance of these metrics. Two types of envelopment were considered: the sense of envelopment from being surrounded by different sources (direct envelopment) and the sense of envelopment resulting from the reflected sound in a reverberant environment (indirect envelopment). An investigation was conducted into the relationship between directional localisation and direct envelopment. This was conducted by first considering each component source in a stimulus separately, using the modified Supper model on these signals and then combining the resulting angles to create a metric for the prediction of direct envelopment. As envelopment is a multi-dimensional perceptual attribute, including cues that are not just associated with localisation, this was followed by investigating how the metrics based on the modified Supper model performed in a regression model which also included metrics developed by Conetta. Further metrics which did not require the stimuli to be decomposed into their component sources were then developed from the output of the modified Supper model. These more generalisable metrics were then used along with Conetta's metrics in regression models for both the direct and indirect envelopment listening test results. This was followed by a discussion of these results in relation to the literature and also a discussion of the possible changes to the metrics which may improve the results.
Chapter 8

Conclusion

This conclusion presents a summary and the main conclusions for each chapter. This is followed by a discussion of how the research contained in the thesis satisfies the research question proposed in Chapter 1, namely how to realise an artificial listener model capable of predicting a number of perceived spatial attributes at arbitrary locations in the listening area. Following this is a summary and brief discussion of the novel contributions of the thesis. Finally is a discussion of a number of ways in which the performance of the artificial listener model can be improved and its functionality can be expanded.

8.1 Summary of Chapter 1: Introduction

It is often necessary to evaluate the performance of reproduced sound, for instance, when designing audio equipment or operating a broadcast network. A common approach for this is to use subjective listening tests. However, as these require a lot of resources, there is increasing interest in computational models which are able to predict the attributes of the reproduced sound perceived by the listeners. Models which predict the timbral and temporal aspects of reproduced sound have already been developed and established (the PESQ [134] and PEAQ [160] models have been adopted by the International Telecommunications Union). However, the increased use of multi-channel reproduction systems has led to an increased need for models which predict the perceived spatial aspects of reproduced sound.

When evaluating the spatial performance of reproduced sound, it is necessary to differentiate between the physical attributes of the sound and the perceived attributes of the sound, as these are often not identical. The perceived spatial attributes of sound have been divided into localisation, the primary spatial attribute, and those spatial attributes which are not localisation, the secondary spatial attributes, which include source width and envelopment.
A number of binaural models (i.e. models which use as their input the pressure signals at the ears of the listener) have been developed to predict the perceived spatial attributes of reproduced sound. However, previous research has concentrated on the sweet spot at the centre of the listening area. The research that has investigated spatial attributes away from the sweet spot has been confined to moving the listener in a single dimension, parallel to the axis between the two ears of the listener. As multi-channel audio reproduction systems are often used with multiple listeners, and hence multiple listening positions, there is a need to evaluate the perceived spatial attributes of reproduced sound at arbitrary locations in the listening area.

8.1.1 Conclusions from Chapter 1

The aim of this research project was to develop a model which can be positioned anywhere in a reproduced sound-field and give predictions of three perceived spatial attributes: directional localisation, source width and envelopment. The model was required to use binaural signals and a physiologically based approach whenever practical, in order that the behaviour of the model should reflect that of real listeners. Two stages of acoustic modelling were required in the project. One of these was to model the sound from the loudspeakers to the ears of the listener, as this allows the model to calculate perceived spatial attributes at any location in the listening area. The other stage involving acoustic modelling was in the simulation of the microphone capture of original sources. This was required in order that a number of different signals for the reproduction systems could be investigated. The final two specifications for the project were that it should be highly modular, in order to give the maximum flexibility, and that the predictions of the different spatial attributes should be validated using formal listening tests.

8.2 Summary of Chapter 2: Spatial perception

Theile [159] divided the spatial sound stimuli that a listener can perceive into three types. These are direct sound, which allows the listener to localise the sound, the indirect (reflected) sound, which gives the spatial impression of the sound, and the environmental (non-reflected) sound, such as the noise of an audience during a concert. The principal perceived spatial attribute associated with direct sound is source localisation, which Supper [158] has termed the primary spatial attribute, which has been regarded as the main purpose of spatial hearing. The perceived spatial attributes which are not localisation are principally associated with indirect sound, and Supper terms these the secondary spatial attributes. Historically, the research into perceived spatial attributes has been undertaken by three different research communities. The first of these consisted of psychophysicists, who were primarily concerned with how the human auditory system worked, while the other two groups were concerned with the perceived spatial attributes in the context of either the acoustics of concert halls or, alternatively, assessing the performance of audio equipment. However, these three research communities have become much more closely linked, as techniques from each field have been employed by the others.
The filter model developed by Fog and Pederson [50] distinguished between objective perceptions and subjective preference. The model consists of two filters, the first of which is a mapping from the physical stimulus to the objective perception. In the context of the perception of sound, this first filter depends on the thresholds and masking effects of the human auditory system. The second filter is a mapping from the objective perception to the subjective preference, which is dependent on the background experience and emotions of the listener. The treatment of Basic Audio Quality by Bech and Zacharov [9] implies that some higher level cognitive processing may still be objective. One consequence of this is that the boundaries between objective perceptions and subjective preference may not be as clearly defined as suggested by the Filter model.

Three different cues are used by the human auditory system in order to determine the directional localisation, the primary spatial attribute. Below 700Hz the main cues are the interaural time differences (ITDs), between 700Hz and 5kHz the most important cues are the interaural intensity differences (IIDs), and spectral cues are the primary cues above 5kHz. Of these three cues, generally only the ITDs and IIDs have been used in binaural models of localisation in the literature. Many of these localisation models have a similar architecture, based on the human auditory system. This firstly consists of the binaural signals being separated into critical bands, with the resulting signals then being half-wave rectified and low-pass filtered. These signals are then used to calculate the IIDs and ITDs for each critical band, with the ITDs derived from the cross-correlation of the left- and right-ear signals in each critical band. These IIDs and ITDs are then converted to azimuths using a database of IIDs and ITDs calculated for known source angles.

Determining the directional localisation is further complicated when reflections are present, with the human auditory system masking the effect of reflections between 1ms and 5ms of the onset of the direct sound, known as the precedence effect. Two approaches have been used in the literature to model this behaviour. The first approach is to modify the calculation of the interaural cross-correlation by introducing inhibitors on the delay lines, while the second approach is to calculate the onsets of the direct sound and determine the azimuth only from the portion of the sound immediately following an onset.

Two secondary spatial attributes have been identified in concert hall acoustics. The first of these is the broadening of a source associated with the early (before 80ms) lateral reflections, termed the Apparent Source Width (ASW). The other secondary spatial attribute is the listener envelopment (LEV) associated with the late (after 80ms) lateral reflections. Measures for ASW and LEV have been calculated as the ratio of the energy of the lateral energy (due to the reflections) to the energy of the total sound over the appropriate time period (the first 80ms for ASW and after the first 80ms for LEV). As the lateral reflections cause the left- and right-ear signals to become decorrelated, alternative measures for ASW and LEV have been based on the calculation of the interaural cross-correlation (IACC) over the appropriate time period.

All of the measures for secondary spatial attributes developed for use in concert hall acoustics are based on an impulse response from the concert hall. As the characteristics of the source signal have been shown to have an effect on the perceived spatial attributes, measures which use continuous signals were developed, for example, by Griesinger [65] and Mason [97]. As the interpretation of these measures was dependent on the whether the binaural signals included the direct part of the sound or only the reflections, Griesinger proposed automatically detecting both the onsets and the offsets of the direct sound in order to segregate
the different portions of the sound.

The scene-based paradigm proposed by Rumsey was developed specifically for reproduced sound and the evaluation of its spatial attributes. This means that spatial attributes which only typically arise in the context of reproduced audio are considered in addition to those spatial attributes closely related to the ASW and LEV attributes used in concert hall acoustics. However, even those models whose development has been influenced by Rumsey's scene-based paradigm, such as Mason's source width model [98] and Conetta's envelopment model [38], have concentrated on similar secondary spatial attributes to those developed for concert hall acoustics.

8.2.1 Conclusions from Chapter 2

This chapter contained an overview of the literature on the perceived spatial attributes of sound, and hence provided a basis for the justification of the decisions taken in the thesis.

The discussion of Fog and Pederson's Filter model [50] showed how different levels of cognition are used for different perceived attributes. This also showed that those perceptual attributes which involve a high level of cognition are also likely to be influenced by the listener's emotions and experience, leading to many of these attributes being more subjective than perceptual attributes involving only a low level of cognition. Some perceived spatial attributes involve consciously combining a number of simpler, lower level attributes. This is reflected in the research by George [57], Conetta [38] and Choi et al. [35], where a number of low level measures where combined using either linear regression or a neural network to predict higher level spatial attributes. As Conetta's work was used as the basis for the investigation into envelopment in Chapter 7, this approach has also to some extent been adopted in this thesis.

This thesis also used the division of the perceived spatial attributes into primary spatial attributes (i.e. localisation) and secondary attributes (i.e. all spatial attributes other than localisation). This is shown both in the relative proportions of the thesis given to each of the three spatial attributes considered and also by the fact that the modified Supper model which was developed to predict directional localisation was also used extensively in the investigations of source width and envelopment.

Models based on the binaural signals at the ears of the listener have been developed for all the perceived spatial attributes discussed in this chapter. In fact, only binaural models have been developed for the prediction of perceived directional localisation. Furthermore, all the binaural models in the literature have had some basis in the physiology of the human auditory system. Hence, this was also the approach adopted in this thesis.

As the characteristics of the audio signals affect the perceived spatial attributes of reproduced sound, the models developed in this thesis were all able to predict the perceived spatial attributes of arbitrary signals. The use of arbitrary signals in models for the prediction of perceived spatial attributes led Griesinger to propose detecting the onsets and offsets of the signals to ensure that only the relevant parts of the signals are used in the calculation of different attributes. Indeed, onset detection has been employed by some
researchers to model the precedence effect. However, the use of onset detection is by no means used by all
the models for predicting perceived spatial attributes in the literature. Hence the decision was made not to
include this in the thesis, but instead to investigate the capabilities and limitations of models which do not
include onset detection.

Finally, even though Rumsey has identified more secondary spatial attributes than just width and envelop-
ment, these two attributes have been the subject of the most research, both in the context of concert hall
acoustics and in the context of reproduced audio. For this reason, the three perceived spatial attributes
investigated in this thesis were directional localisation, source width and envelopment.

8.3 Summary of Chapter 3: Framework of the “source to percept”
model

Compression algorithms based on perceptual models have demonstrated how audio signals can be altered
without affecting the perceived attributes of the reproduced sound. These show the importance of the
perception of the listener in assessing the performance of an audio reproduction system. This motivated the
development of a framework for modelling the perception of reproduced audio, described in Chapter 3. This
began with the identification of nine different stages involved in audio reproduction:

I. Finding or creating the original sound-field
   (e.g. musicians playing a string quartet in a concert hall)

II. Capturing the sound-field
    (e.g. using a microphone array)

III. Processing and/or encoding of the recorded signals
    (e.g. converting the analogue signals captured by the
     microphones into digital signals and writing onto a CD)

IV. Transmission
    (e.g. broadcasting radio signals)

V. Decoding the transmitted signals
   (e.g. converting the data on a CD into analogue signals
    which can be sent to the loudspeakers)

VI. Playing the decoded signals through the reproduction system
    (e.g. playing the loudspeaker signals back through a FCS
     loudspeaker setup)

VII. Capturing the binaural signals
     (e.g. using a dummy head with microphones in the ears
      to record binaural signals for a given listener position)
VIII. Translation to the perceptual domain
(e.g. using a binaural model to predict perceived spatial attributes)

IX. Communicating the perceived spatial attributes
(e.g. the use of a computer interface by a listener in a formal listening test in order to record their perceptions)

The evaluation of the perceived spatial attributes of reproduced audio will concentrate on steps V to IX. However, it is also necessary to include steps I to IV in the model framework, as these determine the signals fed to the reproduction system, which then affect the perceived spatial attributes stimulated in the listener. Stages I and II can be modelled as a linear invariant system, where the pressure at each microphone is calculated as the sum of the pressures due to each sound source. The transfer functions from each source to each microphone were modelled directly in Matlab for the case where the recording environment was anechoic and using either the CATT-Acoustics or ODEON acoustics modelling software for the case where the recording environment was reflective.

Similarly, stages VI and VII can also be modelled as a linear invariant system, with Head Related Transfer Functions (HRTFs) between the each loudspeaker and the ears of the listener. For the case where the reproduction environment was anechoic, the HRTFs were calculated by interpolating the Gardner-Martin HRTF database [53]. For the case where the reproduction environment was reflective, the HRTFs were calculated using either the CATT-Acoustics or ODEON acoustics modelling software.

The last sections in the chapter described the coordinate system used in the thesis and the reproduction systems which were investigated, including Two Channel Stereo (TCS), Five Channel Stereo (FCS) and Wave Field Synthesis (WFS). The discussion of WFS included the reproduction of one of the pressure plots included in Daniel et al.'s 2003 paper [43]. Indeed, all the pressure plots in this paper were successfully reproduced using the Matlab implementation of the model framework described in Chapter 3, which can be seen as an informal validation of the acoustic modelling used in this project.

8.3.1 Conclusions from Chapter 3

The different stages in the audio reproduction chain were identified in order that this could be used as the framework for developing a model for predicting different perceived spatial attributes. From this it was determined that two stages of acoustic modelling were required, one for the capture environment and one for the reproduction environment. It was found that the most practical method of performing the acoustic modelling when the environment was anechoic was to implement it directly in Matlab. This was informally verified, along with the Matlab implementation of the 32-channel WFS reproduction system, by reproducing the pressure plots in Daniel et al.'s 2003 paper [43]. Furthermore, as modelling the reproduction of steady-state sine tones through the WFS system involved stages I through to VI in the model framework, this also demonstrated that these stages of the framework could be successfully used to model reproduced audio.
8.4 Summary of Chapter 4: Listening test experiments

This chapter describes the three listening test experiments that were undertaken in order to provide a data set with which to validate the models for the prediction of directional localisation and source width at different listener positions in the listening area. The first listening test experiment obtained results for a single perceived spatial attribute, namely directional localisation, while the second and third listening test experiments obtained results for directional localisation and source width. Listening tests were not required for envelopment, as a number of listening test experiments investigating this spatial attribute had already been undertaken by Conetta at the Institute of Sound Recording at the University of Surrey, who had generously allowed the use of his results [38, 39].

The first listening test experiment was concerned only with obtaining directional localisation results from the test subjects. Seven loudspeakers were used, five in the standard FCS layout [133] and two additional loudspeakers to provide reference stimuli at known locations, all of which were hidden from the test subjects' view by an acoustically transparent curtain. The directions of each stimulus were elicited from the test subjects using a map-based user interface on an Apple Mac computer, with the aid of a numbered scale, for three different listener positions. Each stimulus in the first listening test experiment was created either by simulating the capture of a sound-field using a microphone array or else by routing the original signal directly to a single loudspeaker. The original signals used in the experiment consisted of sine tones and pink noise generated in Matlab and anechoic speech and music recordings. The most noticeable aspect of the results was the difficulty experienced by the listeners in localising the sine tones, especially compared to the localisation of the other stimuli. This is consistent with the literature [70] and led to the decision to omit the sine tone stimuli from the subsequent listening test experiments.

While the first listening test experiment was only concerned with the direction of the sources perceived by the test subjects, in the second and third listening test experiments the test subjects were required to evaluate both the direction and the width of each stimulus. As in the first listening test experiment, a map-based user interface was used to elicit the perceived spatial attributes from the test subjects. The stimuli for the second and third listening test experiments were created in one of three ways: either routing the original signal directly to a single loudspeaker, using pair-wise constant power panning or using a signal processing algorithm to deliberately widen the source image. The original intention was to have just two listening test experiments, but the third experiment was required for two reasons. Firstly, technical issues with the loudspeakers reduced the amount of time available to conduct the experiment, so only a single listening position was used. Secondly, the algorithm used to widen some of the stimuli in the second experiment did not have a noticeable effect on the source width, so an alternative algorithm was used for the third experiment.

A Wilcoxon signed-rank test showed that there were significant differences between the localisation results from the second and third listening test experiments at the 5% level. Further investigation showed that these changes were mainly improvements in the accuracy of the localisation results, which could be attributed to the changes made to the experimental procedure between the second and third experiments. A similar
test showed that the differences between the source width results from the second and third listening test experiments were not significant at the 5% level. Finally, although a comparison of the source width results for the two experiments showed that there was a noticeable increase in the range of source width values, this was shown not to be significant at the 5% level.

8.4.1 Conclusions from Chapter 4

The experiments described in this chapter successfully created a database of listening test results which were subsequently used to validate the spatial attribute models investigated in this thesis.

The source width results from the second listening test showed that the artificially widened stimuli were judged by the listeners to be not much wider than the panned stimuli. This demonstrates that generating stimuli with a range of sources is not straightforward. Indeed, one of the reasons for including the third listening test experiment was to use a different widening algorithm to generate wider stimuli. The results from the third listening test experiment showed that there was much more variation in the width results than in the localisation results. While this could be partly due to the fact that assessing source width is a more complicated task than assessing source location, it may also be due to the design of the interface used to elicit the source width information from the test subjects, which required the subjects to identify the left and right edges of the stimuli. Rumsey [140] has noted that source width is related to the diffuseness of a source, in which case the edges of the sound will be hard to identify, which may explain the greater variability seen in the source width results.

8.5 Summary of Chapter 5: Directional localisation

The first part of this chapter describes how Supper’s localisation model was incorporated into the framework described in Chapter 3. In addition to the changes which were needed in order to achieve this, a number of modifications were made to the localisation algorithm to improve its performance. The majority of these changes were applied to the generation of the look-up tables used to convert the IID and ITD values to likelihood histograms for the different angles. These changes to the look-up tables included the removal of vertical gaps, the reduction in the end effects at 0° and ±90°, both horizontal and vertical normalisation, and the use of only HRIRs from the front hemisphere. The other changes to the localisation model included fixing the indexing into the binaural streams for the calculation of the correlograms and changing the method of combining the IID and ITD histograms from addition to multiplication.

The modified Supper localisation model was validated against the directional localisation results from the three listening test experiments. There were four groups of stimuli which the model had difficulty localising; these were the sine tones, stimuli with misallocated channels and the stimuli created using a reflective environment, all from the first listening test experiment, and the artificially widened stimuli from the third listening test experiment. The model was able to predict accurately all the other stimuli, having a high
correlation, with an $R^2$ value of 0.96, and a low root-mean-square error of prediction (RMSEP) value of 6°.

As the model had trouble localising the stimuli from the first listening test experiment that had been created using a reflective environment, an investigation was conducted into the effects of reflections on the performance of the model. This investigation was divided into two parts. The first part considered the reflections due to the loudspeaker signals being created by simulating a microphone array in a reflective environment. The second part considered the reflections due to a reflective reproduction environment. In both cases the localisation azimuths predicted by the model tended to 0° as the amount of reflections increased (in both cases more than half of the stimuli considered were localised by the model at 0° when the reverberation time of the room models was increased to around 0.2 seconds). The results from the first listening test experiment also showed that the angles tended towards 0° as the amount of reflections increased. However, this was much less pronounced for real listeners compared with the model: the reflective stimuli in the first listening test experiment were created using a room model with a reverberation time of 0.85 seconds and only 14% of the stimuli were localised by the test subjects at 0°.

The model was then used to compare the performance of different reproduction systems to reproduce source locations accurately. This comparison was first undertaken only at the sweet spot at the centre of the listening area, but this was then followed by systematically sampling the listening area to assess the performance of the reproduction systems across a wider area. In both cases the interpretation of the results was complicated by errors in the model: the model consistently returns angles slightly closer to ±90° than the true angles, and the model also sometimes returns spurious angles of 0° when sources are situated close to ±90°. When these factors were taken into consideration, it was found that the 32-channel WFS system was able to produce the most accurately placed sources, both across the listening area and when only the sweet spot was considered.

8.5.1 Conclusions from Chapter 5

Supper's directional localisation model was successfully incorporated into the model framework described in Chapter 3. Furthermore, the changes made to the model were shown to improve its ability to predict the localisation results from the three listening test experiments. The model experienced problems when localising binaural signals which included reflections from either a reflective recording environment or a reflective reproduction environment. These problems may be alleviated by incorporating Supper's onset detector so that only the portion of the sound after the onset and before the early reflections is used to calculate the directional localisation. This may also help with the other groups of stimuli from the listening test experiments that the model had difficulty localising, such as the artificially widened stimuli from the third listening test experiment.

Finally, it was demonstrated how the model could be used to compare how different reproduction systems were able to reproduce source locations accurately, both at the sweet spot and across the listening area. This also helped to identify systematic errors in the model: one the areas of future work for the research project is to address these systematic errors, which will allow different reproduction systems to be compared much more easily.
8.6 Summary of Chapter 6: Source width

Most of the research into measures of source width has been in the field of concert hall acoustics, where the aim has been to characterise the properties of the hall itself. Hence, the majority of the measures of source width are based on an analysis of the impulse response of the hall. Consequently, these measures are not suitable for measuring the source width in reproduced audio, as the nature of the stimulus (e.g. the rate of speech, or the types of musical instrument) affect the perceived spatial impression [2, 68, 102]. Of the models that calculate secondary spatial characteristics, two have the ability to explicitly calculate source width: Mason's source width model [98] and Supper's spatial analyser [158]. Chapter 6 describes how these two models were incorporated into the framework described in Chapter 3 and subsequently used to predict the source widths for the stimuli used in the third listening test.

The Supper model was incorporated into the framework in the same way as for the prediction of directional localisation, the only difference being how the histograms output by the model are interpreted. Two methods of interpreting the histograms were used to generate measures of source width, both using standard measures of dispersion: the first was the standard deviation of the histograms (the SD method), and the second was the interquartile range of the histograms (the IQR method). These measures were tried with four different histograms from the output of the modified Supper model: the histogram from just the IID cues, the histogram from just the ITD cues, the histogram where the cues have been combined by taking the sum (IID+ITD) and the histogram where the cues have been combined by taking the product (IID*ITD). The product histogram (IID*ITD) was found to have the best performance in terms of predicting the source width results from the third listening test. With this histogram, the SD method had a correlation of 0.35 and a root mean square error of prediction (RMSEP) value of 7.23° and the IQR method had a correlation of 0.23 and a RMSEP value of 8.89°.

The Mason source width model separates the input binaural signals into critical bands, calculates the IACC for each critical band and then converts these IACC values into angles of source width. These angles are then combined across the critical bands by creating a cumulative frequency histogram, with the contribution from each critical band weighted by loudness. Finally, this cumulative frequency histogram is displayed as an intensity plot, giving an intuitive visual indication of the source width of the stimulus. In order to incorporate the model into the framework described in Chapter 3, the method of combining the angles across the critical bands was altered so that a regular frequency histogram was used instead of a cumulative frequency histogram. The spanned angle corresponding to the maximum value of this histogram was then calculated as a single value for the source width. When used to predict the source width results from the third listening test, this model gave a correlation of 0.49 and a RMSEP value of 7.5°.
8.6.1 Conclusions from Chapter 6

The Mason source width model performed better than the source width models based on the output of the modified Supper model. However, the performance of both the models was much worse at predicting source width than the performance of the modified Supper model at predicting directional localisation. One of the reasons for this may be that source width is a more complicated perceptual attribute than directional localisation.

The results for the validation of the Mason source width model were also much worse than those obtained by Mason [99]: Mason obtained a correlation of 0.77, which is much higher than the correlation of 0.49 obtained in this thesis. However, Mason also obtained a RMSEP value of 14.3°, which is higher than the RMSEP value of 7.5° obtained in this thesis. Part of the difference in the validation results may be explained by the difficulty experienced in generating stimuli with large perceived source widths for the second and third listening test experiments over loudspeakers. Mason presented the stimuli for his validation listening test experiment over headphones and obtained a much larger range of source widths from his test subjects (Mason's largest mean source width from his listening test was 120°, compared with 38° for the third listening test experiment in this thesis).

Mason also obtained much more consistent source width results from his validation listening test. There were a number of differences between the design of Mason's source width listening test experiment and the source width experiments undertaken for this thesis. While a visual scale was used in both cases, Mason presented his stimuli over headphones and had multiple stimuli on each page, allowing the test subjects to freely switch between the stimuli, and the stimuli were always positioned directly in front of the listener. In contrast, in the second and third listening test experiments described in Chapter 4, the stimuli were presented over loudspeakers, the test subjects were not able to revisit stimuli once they had submitted their answers and the stimuli were positioned at a variety of angles, not just directly in front of the listener. Another possible reason for the relatively inconsistent source width results from the second and third listening test experiments is that the method used in the user interface relied on identifying the two edges of the sound. As source width is closely related to the diffuseness of the sound source, this may not be the best method of eliciting perceived source width information from the test subjects.

A number of possible approaches to improving the performance of the source width models were identified. The first of these is to turn the modelled head so that it faces the direction of the perceived sound source. The second is to improving the source width model is to adopt a more probabilistic approach in both the Supper and the Mason models, so that instead of returning a single width for each critical band, a conditional probability distribution is output. The third approach to improving the performance of the source width models is to incorporate Supper's onset detector in order to identify the portions of each stimulus corresponding to the early reflections and use only these portions to calculate the source width.
8.7 Summary of Chapter 7: Envelopment

While much of the literature regarding the perception of envelopment has been in the field of concert hall acoustics, the last fifteen years have also seen the perception of envelopment being researched in the context of reproduced audio. Whereas, in concert hall acoustics, the sense of envelopment is closely related to the reflective properties of the hall, this is not necessarily the case with reproduced audio. For instance, envelopment with reproduced audio can be related to an individual source, or can result from being surround by a number of different sources. For this reason, Conetta [38] divided his study of envelopment into direct envelopment, resulting from being surrounded by sound sources, and indirect envelopment, resulting from the reflections of either the recording or reproduction environments. The same division of envelopment was also adopted in this thesis, partly in order to be able to use the data from Conetta's listening test experiments.

Two different approaches were used in the development of metrics that could be used to predict envelopment. The first approach was only used with Conetta's direct envelopment stimuli and data. This consisted of first decomposing each stimulus into its component sources and then passing each component source one at a time through the modified Supper model. This approach was adopted for two reasons: firstly, it could provide an insight into the relationship between perceived localisation and perceived envelopment; secondly, the use of the modified Supper model to predict the localisation of single sources has already been validated (see Section 5.2). Metrics were then developed that combined the resulting set of localisation angles into a single value for the prediction of direct envelopment. Two different metrics were developed that used the decomposed stimuli: one based on the area within a convex hull determined by the localisation angles, and the other based on the localisation angle closest to ±90°. Finally, the newly developed metrics were used in conjunction with the metrics developed by Conetta in a linear regression model to predict the results from Conetta's direct envelopment listening test. The metric sets containing both metrics derived from the modified Supper model and Conetta's metrics were found to have similar performance to the metric sets containing just Conetta's metrics. Conetta's metric set containing five metrics had an R² value of 0.89 and a root-mean-square error of prediction (RMSEP) value of 7.45%. The best performing metric set containing five metrics chosen from both those developed by Conetta and those developed from the modified Supper model had an R² value of 0.90 and a RMSEP value of 7.18%. This shows a small improvement on Conetta's five metric set when trying to predict the results of the direct envelopment experiments.

One of the limitations of the first approach is the difficulty of decomposing each stimulus into its component sources without prior knowledge of the individual source signals. This task is beyond the scope of the research for the thesis. Consequently, the second approach consisted of inputting the complete (non-decomposed) stimuli through the modified Supper model and developing metrics based on the histograms output by the model. Two different families of metric for interpreting the histograms output by the modified Supper model were developed. In the first metric family, the metric selects a subset of the histograms output by the Supper model by only selecting those time frames where the histogram peak is above a given threshold. This ensures that only the time-frames where the model is able to localise a source are used. The set of angles corresponding to the maximum values of the histograms are then calculated, and the final value of the metric is the angle from this set that is closest to ±90°. The second metric family consists of calculating the
standard deviations of the histograms output by the modified Supper model. These two families of metrics were again used in conjunction with Conetta's metrics in linear regression models to predict the results from Conetta's direct envelopment listening test. The same metrics were then used in linear regression models to predict the results from Conetta's indirect envelopment listening test. The metric sets containing both metrics derived from the modified Supper model and metrics developed by Conetta were found to perform similarly to the metric sets containing just Conetta's metrics. Conetta's metric set containing five metrics had an R² value of 0.82 and a RMSEP value of 10.78%. The best performing metric set containing five metrics chosen from both the metrics developed by Conetta and the metrics derived from the modified Supper model had an R² value of 0.89 and a RMSEP value of 8.57%. Again, this shows a small improvement on Conetta's five metric set when trying to predict the results of the direct envelopment experiments.

8.7.1 Conclusions from Chapter 7

This chapter has shown that the output of the modified Supper model can be used to generate metrics which can successfully predict envelopment when combined with other metrics in a linear regression model. Moreover, the inclusion of the metrics derived from the modified Supper model was shown to have better performance when compared to Conetta's original set of metrics used for the regression. This was shown to be true for both types of envelopment studied by Conetta. The metrics developed from the modified Supper model used a more physiologically based approach than had been employed by Conetta for some of his metrics, including the use of the same binaural signals as experienced by the listener. The improvement in the envelopment predictions when the metrics derived from the Supper model were included supports the hypothesis that the listeners' perception of envelopment can be best modelled by using the same cues as the listeners together with physiologically-based low-level processing.

Finally, Griesinger [66] discussed how there are three different types of perceived spatial impression depending on the relationship between the direct sound and the reflected sound, with two of these leading to a sense of envelopment. Conetta's listening tests for indirect envelopment used stimuli where the reflected sound was dominant, so the cues due to the temporal relationship between the direct and indirect sound were non-typical of most real program material. Also, none of the metrics developed either by Conetta or from the modified Supper model took into account the fact that different periods of the stimuli will affect the perceived envelopment in different ways, depending on whether the direct and/or indirect sounds were present. This could be remedied by using the solution proposed by Griesinger, of detecting both the onsets and offsets of the direct sound in order to segregate each stimuli into the regions where (i) only the direct sound is present (ii) the direct sound and the early reflections are present together and (iii) only the reflections are present. Supper [158] has already developed an onset detector for use with his localisation model, and it is possible that this could also provide the basis for an offset detector. This would allow predictions of envelopment to be made for stimuli which have a more typical relationship between the direct and indirect sound.
8.8 Conclusion and contributions to the field

This section describes how the research in the thesis has answered the research question and satisfied the project requirements contained in Section 1.3 in the Introduction Chapter. Following this is summary of how the research in the thesis has contributed to the relevant research fields.

8.8.1 Modelling the perceived spatial attributes of reproduced sound

The aim of the project described in this thesis was to develop an artificial listener with the ability to predict different perceived spatial attributes at different locations in the listening area of an audio reproduction system. Hence, the research question for the thesis, namely how can such an artificial listener be realised, has been answered by developing models for predicting the spatial attributes using the model framework described in Chapter 3. The following list recapitulates the requirements of the model, together with a brief discussion of how each requirement has been satisfied:

- Three different perceived spatial attributes were required to be investigated: directional localisation, source width and envelopment.

This requirement has been satisfied: Chapters 5 to 7 report how the three spatial attributes have been investigated using the model framework described in Chapter 3. However, the predictions of source width have a relatively low correlation with the source width results from the formal listening tests. Similarly, while the model was generally able to predict directional localisation with a high degree of accuracy, the validation with the listening test results showed that there are some groups of stimuli whose directional localisations are not well predicted by the model. Therefore, the models of both directional localisation and source width require further work, and changes to the two models which may improve their performance have been identified.

- Binaural signals were required to be explicitly modelled, allowing existing binaural models to be used as a basis for the model.

This requirement has been satisfied: the model framework described in Chapter 3 includes the acoustical modelling of the reproduced sound from the loudspeakers to the listener's ears. The calculated binaural signals are then input to binaural models to calculate the perceived spatial attributes. Furthermore, the thesis describes how two existing binaural models, namely those of Supper [158] and Mason [98], were incorporated into the model framework and used to generate predictions of all three spatial attributes investigated.
The model was required to predict the spatial attributes at different locations in the listening area.

This requirement has been satisfied: as stated previously, the model framework described in Chapter 3 includes the acoustical modelling of the reproduced sound from the loudspeakers to the listener's ears. This allows the artificial listener to be placed at arbitrary locations in the listening area and then calculate predictions of the spatial attributes at these locations. However, this was only systematically investigated for directional localisation. Three different listener positions were used in the validation of both directional localisation and source width, but the ability of the source width model to predict the results from the listening tests was not judged good enough to justify an investigation into perceived source width across the listening area. The investigation into predicting perceived envelopment was reliant on the Conetta’s listening test results [38], which was confined to a single listener position at the sweet spot in the centre of the listening area. Hence, the model for predicting envelopment was only validated at the sweet spot and it is not currently known how the model generalises to other listener positions. Therefore, although predictions of all three spatial attributes can be calculated at arbitrary positions in the listening area, only the predictions for directional localisation have been successfully validated at different listening positions.

The model was required to include acoustical modelling, both in the recording and reproduction environments, in order to be enable the comparison of the spatial attributes of different audio reproduction systems.

This requirement has been satisfied. The acoustical modelling of the recording environment was used to generate the majority of the stimuli in the first listening test experiment. An anechoic recording environment was also used to model the microphone techniques for TCS and FCS in the comparison of the ability of different reproduction systems to reproduce source locations accurately. The acoustical modelling of the reproduction environment was used to calculate the binaural signals for all three spatial attributes considered in the thesis. The effects of reflections in both the recording and reproduction environments on the behaviour of the localisation model was also investigated in Chapter 5.

The model was required to be as modular as possible in order to maintain flexibility.

This requirement has been satisfied. The two acoustic modelling sections in the model framework both use the WAV [47] file format for the input and output audio data. Similarly, the binaural models used in the thesis also use the WAV file format for the input audio data. As the acoustical modelling sections and the binaural models are all both self-contained and use standard file formats for data input and output, the different sections of the model can also be used separately or combined with other models in a straightforward manner. This was demonstrated to some extent by the investigation into envelopment reported in Chapter 7, where the stimuli created by Conetta [38] were input to only the latter half of the model framework (stages V to IX from Fig. 3.1).
8.8.2 Listening tests

A novel user interface for eliciting localisation and width perceptions from listeners at different positions within the listening area was developed and then employed in formal listening test experiments. The user interface was found to lead to reliable results for directional localisation. However, there was much greater variability in the results for source width obtained using the user interface. There are a number of possible explanations for this. One possible explanation is that the method of eliciting source width by identifying the left and right edges of the source is not well suited for reproduced sound. Another possible explanation is that the methods used to generate stimuli with a range of source widths were not suitable for listening tests where the stimuli were presented over loudspeakers. A third possible explanation is simply that assessing source width is a more complicated task than assessing directional localisation. Therefore, further investigation is required in order to be able to assess the suitability of the user interface for eliciting perceived source width for reproduced sound.

8.8.3 Primary spatial attribute: localisation

Supper's directional localisation algorithm [158] was modified in a number of novel ways to improve its performance. The majority of these modifications were made to the algorithm used to generate the look-up tables used to convert the calculated IID and ITD cues to localisation histograms. These included: the horizontal and vertical normalisation of the look-up tables, the removal of the mapping of the IIRIR results from the rear hemisphere to the front hemisphere, steps taken to minimise the effects due to the edges of the look-up tables, and the removal of vertical gaps in the look-up tables. Other modifications to the Supper's localisation algorithm included changing the method of combining the output histograms for the different cues (IIDs and ITDs) from addition to multiplication.

The implementation of the changes to Supper's directional localisation algorithm resulted in an improvement in the performance of the localisation algorithm. This was shown by the validation for multiple listening positions using ecologically valid stimuli (including wide-band noise, speech and musical instruments) and the results from the formal listening tests. While Macpherson [89] and Rose et al. [136] have both used binaural models to investigate directional localisation at multiple listening positions, these have both been limited to investigation along a single dimension and have also both been undertaken for a single reproduction system (TCS in Macpherson's case and transaural stereo for Rose et al.). In addition, Macpherson only used an informal listening test to validate his investigation.

A systematic exploration of the effects of the effects of reflections (both from the recording environment and from the reproduction environment) on the behaviour of the directional localisation model was undertaken. This showed that the model was unable to correctly localise sources when the reverberation time was above 0.2 seconds and also that its behaviour appeared to be an exaggerated version of the behaviour shown by real listeners. A number of possible changes to the model that may improve its performance when localising reflective signals were identified, principally the incorporation of Supper's onset detector [158] in order to take the precedence effect into account.
The modified Supper model together with the model framework described in Chapter 3 were used to compare how a number of different reproduction systems were able to reproduce source locations accurately. Similar investigations have been undertaken in the past, for instance Pulkki and Ilirvonen [128], but these have been confined to the sweet spot at the centre of the listening area. While the initial investigation comparing different reproduction systems in this thesis did concentrate only on the sweet spot, this was followed by a comparison of the reproduction of source locations for different reproduction systems across the listening area. This was achieved by using the model to sample the listening area on a grid with 10cm spacing. The large amount of computation involved in this necessitated identifying the computational bottlenecks in the Matlab code, rewriting this in ANSI C [77] and integrating the compiled code back into the Matlab model. Thus, the code for most of Supper's model was rewritten in order to reduce the computation time so that the approach of spatially sampling the listening area was practical. This reduced the computation time approximately by a factor of ten and allowed the localisation azimuths for a circular area with a radius of 1.5m to be sampled at points 10cm apart to be calculated in approximately one and a half hours. This allowed the creation of maps illustrating how reproduced source locations vary across the listening area. While Härma et al. [69] created maps showing how the timbral quality varies across the listening area for reproduced sound, no previous research has shown maps of spatial quality across the listening area for reproduced sound.

Finally, a method of differentiating between front and back sources was implemented. It has been noted in the literature (for example, Griesinger [66]) that head movements are one of the main cues for distinguishing between front and back sources. The model framework described in Chapter 3 includes the integration of both binaural models for calculating predictions of perceived spatial attributes and the acoustics modelling from the loudspeakers to the listener's ears (including the positions of both the loudspeakers and the listener). Hence, small head movements were incorporated into the model framework in a relatively straightforward manner. The changes in the localisation cues due to these small head movements were then used to differentiate between front and back sources. This was shown to work successfully with the stimuli from Conetta's localisation and envelopment listening test experiment [39].

8.8.4 Secondary spatial attributes: source width and envelopment

The performance of two different models for predicting source width was investigated using the source width results from the formal listening test experiments. Note that while Mason had already assessed the ability of his source width model using listening tests where the stimuli were presented to the test subjects over headphones [99], the listening tests used to validate the models in this thesis differed from this in that the stimuli were presented over loudspeakers.

A number of high-level metrics based on the source location data were developed during the investigation of envelopment. Combining these with some of the metrics developed by Conetta [38] in a regression model resulted in an improvement in the prediction of the results from Conetta's envelopment listening tests when compared with a regression model using only Conetta's metrics.
8.9 Continuing and extending the research project

This section consists of a brief summary and discussion of how the research described in this thesis can be continued. Indeed, it also highlights those areas which need to be addressed if the project is to be continued. This section is divided into two parts. The first part discusses the need for onset and offset detection in the model, as this has the potential to improve the performance of the models for all three spatial attributes considered in the thesis. The second part discusses a number of other possible improvements to the individual models developed and investigated in the thesis, together with a very brief discussion of extending the functionality of the artificial listener to include other spatial attributes.

8.9.1 Detection of onsets and offsets

The most important proposed improvement to the artificial listener model is the incorporation of Supper's onset detector [158]. The modifications to Supper's localisation model described in Chapter 5 included the horizontal and vertical normalisation of the look-up tables used to generate the histograms output by the model. It was thought that once consequence of this would be large peaks in the output histograms for those portions of the binaural signals which are easily localisable by the model, and that this would lead to the localisation results for these portions of the signals dominating the final localisation results. As those portions of the binaural signals corresponding to just the direct sound, i.e. before the early reflections, have the most well defined localisation cues, these should have the largest influence on the final localisation results output by the model. Hence, Supper's onset detector was omitted from the investigation into directional localisation reported in Chapter 5, partly for this reason and partly to keep the model as simple as possible.

However, the validation of the localisation model and the subsequent investigation into the effects of reflections showed that the localisation model had major difficulties when reflections were present in the binaural signals. Detecting the onsets in the binaural signals would allow the precedence effect [18] to be modelled by identifying the direct sound in the signals and weighting the output of the localisation model to favour these portions of the signals. This will be effectively suppressing the localisation cues when the early reflections are present in the signals. This should improve the performance of the localisation model when reflections are present in the binaural signals. Also, source width has been associated with the early lateral reflections [7, 9], so using an onset detector to identify the early reflections in the binaural signals may also lead to a more robust model of perceived source width.

Related to the subject of onset detection is offset detection, where the end of each direct sound is detected. Griesinger [66] and Supper [158] both proposed the use of an offset detector for the prediction of listener envelopment, and Supper also suggests that it could prove useful for determining the distance of sources. Incorporating both an onset and offset detector would allow the model to determine whether the direct sound, early reflections or late reflections were present for any given time frame of the binaural signals. This would allow the prediction of envelopment to include the cues due to the relationship between the reflected sound and the direct sound in the signals, which is currently missing from both Conetta's envelopment
models and the models for predicting envelopment developed in Chapter 7.

The main motivation for proposing that the artificial listener model is expanded to include onset and offset detectors is the fact that different groups of time frames of the binaural signals affect the different perceived spatial attributes in different ways. This has been well documented in the literature [7, 25, 66, 162]. The measures for predicting directional localisation and source width which were investigated in this thesis have used a relatively crude method of averaging the output of the models over the duration of each stimuli. Therefore, one area in which the models developed in the thesis could be improved is by developing a more sophisticated method of obtaining a single measure for each spatial attribute from the results output over the duration of each stimulus. Another area of future work related to this is the investigation of stimuli whose spatial characteristics vary over time, e.g. stimuli whose location varies over their duration.

8.9.2 Spatial attribute models

It was noted in the previous section that different portions of the binaural signals affect the perceived spatial attributes in different ways. It was also noted that the time frames of the binaural signals which are well localisable by the modified Supper model result in large peaks in the output histograms. Both these points were used in the development of higher level metrics for predicting envelopment from the output of the modified Supper model, where the final values of the metrics were calculated from only those time frames whose maximum peaks were above a given threshold. One area of further work to the project is to investigate the when a similar method is used for predicting the directional localisation. As the cues for localisation will be least confusing when reflections are not present, applying a threshold to the maximum peaks of the histograms for each time frame may improve the ability of the model to predict localisation angles in the presence of reflections.

The research reported in this thesis could be extended by implementing and investigating some of the improvements to the source width models proposed in Chapter 6. One proposed improvement is to rotate the artificial listener's head to face the perceived source source before predicting the perceived source width. Another approach is to adopt a more Bayesian, probabilistic approach to the calculation of the histograms output by the model. A third possible improvement is to consider explicitly the case when a source is localised at different angles for different critical bands. Another possible avenue of further research is to investigate different methods of creating wide stimuli in order to obtain a wider range of source widths, as this may better reflect how the source width model performs with real program material.

One area in which the work on predicting perceived envelopment reported in Chapter 7 could be expanded is to perform further listening test experiments with multiple listener positions. This would allow the models developed to predict envelopment to be generalised to any listening position. Another, related, area of further work is to modify the existing metrics developed by Conetta [38] so that they all use the same binaural signals as the test subjects. This would bring all of the envelopment model into line with the model framework described in Chapter 3, not just those metrics for the prediction of envelopment which were developed from the modified Supper model.
Finally, the research project can be extended by developing binaural models for the prediction of additional perceived spatial attributes. For example, Rumsey [140] identifies more perceived spatial attributes of reproduced sound than the three investigated in this thesis, including source distance, ensemble width, ensemble depth and scene left-right skew. Note that number of these perceived spatial attributes require significantly more high-level cognitive processing on the part of the listener than the three spatial attributes considered in this thesis, especially directional localisation. Once binaural models have been developed to predict some of these further spatial attributes, these can be incorporated into the model framework in a similar manner to the binaural models considered in this thesis.

8.10 Summary

This chapter has summarised the content and conclusions from each of the preceding chapters. This was followed by a discussion of how the research contained in this thesis has satisfied the aims and requirements described in Chapter 1, together with a summary of the main contributions of the thesis. The last section in the chapter consisted of a discussion of how the research can be continued, together with the areas which need to be addressed if this is the case.

This thesis has described the development of an artificial listener model capable of predicting a number of different perceived spatial attributes at arbitrary locations in the listening area. The three spatial attributes considered were directional localisation, source width and listener envelopment: that these three attributes have been the subject of most research implies that these are the three most significant attributes when perceiving the spatial aspects of sound. The models for predicting each of the three spatial attributes were validated using the results from formal listening tests. Where low correlations were obtained between the listening test results and the model predictions, notably in the case of perceived source width and some groups of stimuli for directional localisation, a number of modifications which may improve the performance of the model have been identified.

Supper’s directional localisation model [158] was extensively modified, leading to an improvement in its ability to predict the localisation results from the listening tests. The modified Supper localisation model was then used to investigate localisation across the listening area for different reproduction systems and also to investigate its limitations when reflections were present in the signals. The investigations into source width and envelopment in Chapters 6 and 7 respectively showed how the modified Supper model could be used as the basis for predicting spatial attributes other than directional localisation. The higher level metrics developed from the output of the modified Supper model for the prediction of envelopment were shown to improve the predictions of listener envelopment when combined in a regression model with existing metrics.
Appendix A

Glossary

**ASW**
Auditory (or apparent) source width. A secondary spatial attribute.

**binaural signals**
The signals at the two ears of a listener.

**correlogram**
A histogram of cross-correlations for different time lags.

**cross-correlation**
A measure of the similarity between two signals, calculated as the sliding scalar product.

**FCS**
Five channel stereo.

**head related transfer function**
The phase and magnitude response of a sound due to the filtering effect of the ear canal, the outer ear, the head and the torso.

**HRIR**
Head related impulse response. The time domain equivalent of a head related transfer function. These can be created by using a dummy head to record the binaural signals caused by an impulse response.

**HRTF**
Head related transfer function.

**IID**
Interaural intensity difference.
ITD
Interaural time difference.

KEMAR
Knowles Electronic Manikin for Acoustic research. A dummy head and torso used to record binaural signals.

LEV
Listener envelopment. A secondary spatial attribute.

ORTF
Office de Radiodiffusion-Television Française. A spaced microphone configuration for use with TCS. See Fig. 4.3

pinna
The external part of the outer ear.

precedence effect
The low-level inhibition of the effect of early reflections on localisation.

primary spatial attribute
The localisation of sound sources.

secondary spatial attribute
Any spatial attribute that is not the localisation of sound sources.

TCS
Two channel stereo.

WFS
Wave field synthesis.
Appendix B

Instructions for the first listening test
Objective Measure of Spatial Quality Listening Test

Purpose of the listening test

The purpose of this listening test is to provide experimental data with which to validate a computational model for objective measures of perceived spatial quality, including the perceived direction of sounds. In these listening tests, the listener is asked to record the direction in which they perceive the sound to be coming from for a variety of stimuli presented over loudspeakers.

What you have to do

1. To start the test, press the red button, which will have the caption “Start the next listening test”.

2. listen to the stimulus, decide which direction you think it is coming from and enter the angle into the user interface. This can be done in one of two ways:
   (i) Click on the plan view of the listening area with the mouse and use the mouse to move the arrow to point in the direction you want.
   (ii) Click on the number box just below the plan view and type in the number from the scale where you think the sound is coming from.

3. If you want to hear a stimulus again then press the green button.

4. Once you are happy with the number you have entered into the interface, either using the mouse of the number box, then click on the red button. This will save your answer and play the next stimulus.

5. Please try to keep your head facing forward throughout the experiment, as changing the direction in which your head faces will affect the results of the experiment.

6. If you are entering your answers using the mouse then please check the number displayed below the plan view, as it is easy to introduce errors when just using your eye to move pointer.

7. If you have any questions or problems then just ask.
Appendix C

Instructions for the second listening test
Instructions for listening test

1. The aim of the listening test is to investigate how listeners localise and perceive the apparent source width of a number of stimuli.

2. The test supervisor will demonstrate the interface.

3. To start the test press the PINK button or ‘s’ on the keyboard.

4. Three responses are required for each stimulus: the direction the sound is coming in, the left edge of the sound and the right edge of the sound.

5. Try to keep your head stationary and facing forward during the test.

6. The mouse can be used to enter the directions by clicking on the picture in the interface, which will cause an arrow to appear in the picture. This arrow can be repositioned using the mouse.

7. Clicking on the number boxes / buttons to the left of the picture will change the current response being entered (i.e. clicking on the top button will allow the arrow for localisation to be entered, the middle button will allow the arrow for the left edge to be entered and the bottom button will allow the arrow for the right edge of the sound to be entered). These can also be changed by using the UP and DOWN arrows on the keyboard.

8. The responses can also be typed on the keyboard using the number keys.

9. SPACE or the GREEN button repeats the stimulus.

10. ‘s’ or the PINK button saves your responses and moves onto the next stimulus.

11. There are 24 stimuli.

12. If you do not enter 3 responses for each stimulus then a message will appear. This can only be cleared by clicking on it. Similarly, if the arrow for the left edge is to right of the direction arrow or the right arrow then a message appears, etc.

13. If you have any problems or queries then please do not hesitate to contact the test supervisor.

14. Thank you for taking part in the test!
Appendix D

Instructions for the third listening test
Instructions for listening test

1. The aim of the listening test is to investigate how listeners localise and perceive the apparent source width of a number of stimuli.

2. The test supervisor will demonstrate the interface.

3. To start the test press the PINK button or 's' on the keyboard.

4. Three responses are required for each stimulus: the direction the sound is coming in, the left edge of the sound and the right edge of the sound.

5. Try to keep your head stationary and facing forward during the test.

6. The mouse can be used to enter the directions by clicking on the picture in the interface, which will cause an arrow to appear in the picture. This arrow can be repositioned using the mouse.

7. Clicking on the number boxes / buttons to the left of the picture will change the current response being entered (i.e. clicking on the top button will allow the arrow for localisation to be entered, the middle button will allow the arrow for the left edge to be entered and the bottom button will allow the arrow for the right edge of the sound to be entered). These can also be changed by using the UP and DOWN arrows on the keyboard.

8. The responses can also be typed on the keyboard using the number keys.

9. 's' or the PINK button saves your responses and moves onto the next stimulus.

10. Each stimulus will automatically repeat until moving onto the next stimulus. Pressing the SPACE bar or the GREEN button will pause the stimulus. Pressing SPACE or the GREEN button again will resume the playing of the stimulus.

11. There are 27 stimuli.

12. If you do not enter 3 responses for each stimulus then a message will appear. This can only be cleared by clicking on it. Similarly, if the arrow for the left edge is to right of the direction arrow or the right arrow then a message appears, etc.

13. If you have any problems or queries then please do not hesitate to contact the test supervisor.

14. Thank you for taking part in the test!
Appendix E

Evaluation of the performance of the listening test subjects
E.1 · The first listening test

Testing for normality

The Shapiro-Wilk test [48] was used to determine whether the localisation results from the first listening test experiment were normally distributed for each stimulus. The sine tone stimuli were omitted due to the difficulty with which these were localised by the test subjects. The null hypothesis for each Shapiro-Wilk test was that the listening test results had a normal distribution. This was tested at the 5% level, and the combinations of stimulus and listening position for which the null hypothesis was rejected are contained in Table E.1. The null hypothesis was rejected for 23 of the 90 possible combinations, approximately 26%. Consequently, non-parametric statistical tests were required to test for inter- and intra-subject consistency.

Testing for inter-subject consistency

For each test subject the correlation was calculated between the subject's localisation results and the results for all the remaining subjects. This was done in order to investigate the performance of each test subject in relation to all the other subjects. It has been shown that not all the localisation results for the first listening test experiment are normally distributed, and consequently a non-parametric measure of correlation was used. The Kendall rank-order correlation T [146] was calculated between each subject's localisation results and the mean results for all the remaining subjects. The results of these calculations are shown in Fig. E.1. From this figure it can be seen that all the test subjects have highly correlated results with the results from the other subjects (the lowest correlation value was 0.85) and that none of the subjects had a noticeably different correlation.

Testing for intra-subject consistency

Each stimulus was repeated once for each combination of test subject and listening position. Consequently there are two localisation angles for each combination of test subject and listening position and these can be used to assess whether each test subject can give consistent results. Again, as it has been shown that not all the localisation results for the first listening test experiment are distributed normally, a non-parametric statistical test was used. Table E.2 shows the results of performing a Wilcoxon signed-rank test [48, 146] on the repeated stimuli. The results of only one test subject, number twelve, were rejected at the 5% level using the test. The effect size, r, of -0.43 for subject twelve shows that the difference between the results for the localisation results from the repeated stimuli is a medium effect [48].
Comparing against stimuli with known source locations

Fig. E.2 shows the root-mean-square errors calculated for each test subject between their localisation angles and the actual location angles for those stimuli played through a single loudspeaker (i.e. those stimuli which had a known source location). From this figure it can be seen that subject twelve gave localisation results with a low root-mean-square error (RMSE), only 3.5°, compared to the other subjects. This shows that subject twelve was more accurate at localising the sources for this subset of stimuli compared to the majority of the other subjects. As it is likely from this that subject twelve's localisation results were also accurate for the other stimuli from the first listening test experiment (excluding the sine tones), subject twelve's results will not be removed.

The average RMSE for the single loudspeaker stimuli across all the test subjects was 7.7° and subject eight had the largest RMSE value (12°). In their 1936 paper, Stevens and Newman [154] reported that the RMSE values were greater when the source was not located either directly in front of or behind the listener. For angles in the range 15° to 90°, Stevens and Newman obtained RMSE values in the range 13.0° to 16.3°, which are larger than the RMSE value of 12° for subject eight's localisation results. Therefore, as the results of subject eight were also consistent both for the repeated stimuli and in relation to the results of the other test subjects, the decision was made not to remove subject eight's results.

Conclusion

Following the assessment of the inter-subject and intra-subject consistency and the assessment of the subjects' localisation results for those stimuli with known source locations, the decision was made not to remove the results for any of the test subjects from the first listening test experiment results.
Table E.1: Results for Shapiro-Wilk test at the 5% level on the localisation results from the first listening test. The null hypothesis of the test is that the results have a normal distribution. The crosses in the table show those combinations of stimulus and listening position where the null hypothesis was rejected, i.e. the results did not have a normal distribution. Those stimuli that had normal results for all three listening positions have been omitted from the table.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
</tr>
<tr>
<td>14</td>
<td>x</td>
</tr>
<tr>
<td>17</td>
<td>x</td>
</tr>
<tr>
<td>18</td>
<td>x</td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>x</td>
</tr>
<tr>
<td>26</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>x</td>
</tr>
<tr>
<td>28</td>
<td>x</td>
</tr>
<tr>
<td>33</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>
Figure E.1: Inter-subject consistency for the localisation results from the first listening test experiment. For each subject the Kendall rank-order correlation coefficient $T$ was calculated between the subject's results and the mean results from the remaining subjects.

Figure E.2: Comparison of the localisation results for each test subject with the stimuli with known locations from the first listening test. Root-mean-square errors between each subject's localisation results and those stimuli played from a single loudspeaker.
Table E.2: Intra-subject consistency for the localisation results from the first listening test experiment. The results of performing the Wilcoxon signed-rank test on the localisation results for each test subject from the first listening test experiment.
E.2 The second listening test

E.2.1 Localisation results

Testing for normality

The Shapiro-Wilk test was used to determine whether the localisation results from the second listening test experiment were normally distributed for each stimulus. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5% level, and the null hypothesis was rejected for the localisation results for 4 of the 24 stimuli, namely stimuli numbers 1, 9, 15 and 23. As not all the results were normally distributed, non-parametric statistical tests were used to test for inter-subject consistency.

Testing for inter-subject consistency

The Kendall rank-order correlation $T$ was calculated between each subject's localisation results and the mean results for all the remaining subjects. The results of these calculations are shown in Fig. E.3. From this figure it can be seen that all the test subjects had results that were highly correlated with the results from the other subjects (the lowest value was 0.95) and that none of the subjects had a noticeably different correlation. Note that some of the calculated correlations are actually slightly higher than one. This is due to the correction for tied observations: Table E.3 shows the Kendall rank-order correlation coefficients calculated both with and without the tied observations correction.

Comparing against stimuli with known source locations

Fig. E.4 shows the root-mean-square errors calculated for each test subject between their localisation angle results and the actual location angles for those stimuli played through a single loudspeaker (i.e. those stimuli which had a known source location). The average RMSE for the single loudspeaker stimuli across all the test subjects was 4.5°. Subject five had the largest RMSE value (7.7°), which is still less than the RMSE values obtained by Stevens and Newman [154] (see Section E.1). Therefore, as the localisation results of subject five were also consistent with the localisation results from the other test subjects, the decision was made not to remove subject five's results.

Conclusion

Following the assessment of the inter-subject consistency and the assessment of the subjects' localisation results for those stimuli with known source locations, the decision was made not to remove the results for any of the test subjects from the second listening test experiment localisation results.
E.2.2 Source width results

Testing for normality

The Shapiro-Wilk test was used to determine whether the source width results from the second listening test experiment were normally distributed for each stimulus. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5% level, and the null hypothesis was rejected for the localisation results for 4 of the 24 stimuli, namely stimuli numbers 1, 12, 14 and 20. As not all the results were normally distributed, non-parametric statistical tests were used to test for inter-subject consistency.

Testing for inter-subject consistency

The Kendall rank-order correlation $T$ was calculated between each subject's source width results and the mean results for all the remaining subjects. The results of these calculations are shown in Fig. E.5. The average $T$ value was 0.64, and the $T$ values are much lower than those calculated using the localisation results from the second listening test experiment. Figs. 4.14 and 4.15 show that the source width results from the second listening test experiment have a relatively narrow range of values and that there is considerable variation in the source width results for each stimulus. This means that it is likely that the rank-orders of the source width results for each test subject will differ from each other, which is reflected in the lower Kendall correlation values. Although subjects five, nine and ten all have low $T$ values (0.52, 0.51 and 0.52 respectively), these subjects' source width results will not be removed, as there is a large variation for all the source width results form the second listening test experiment.

Conclusion

Following the assessment of the inter-subject consistency, the decision was made not to remove the results for any of the test subjects from the second listening test experiment source width results.
Figure E.3: Inter-subject consistency for the localisation results from the second listening test experiment. For each subject the Kendall rank-order correlation coefficient $T$ was calculated between the subject's results and the mean results from the remaining subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Kendall rank-order correlation coefficient, $T$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With tied observations correction</td>
</tr>
<tr>
<td>1</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
</tr>
<tr>
<td>4</td>
<td>1.01</td>
</tr>
<tr>
<td>5</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>0.98</td>
</tr>
<tr>
<td>7</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>1.01</td>
</tr>
<tr>
<td>9</td>
<td>0.95</td>
</tr>
<tr>
<td>10</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Table E.3: The effect of the tied observations correction on the calculation of the Kendall rank-order correlation coefficient, $T$, for the second listening test experiment results. The correlation coefficient was calculated between the localisation results for each subject and the mean localisation results from the remaining subjects.
Figure E.4: Comparison of the localisation results for each test subject with the stimuli with known locations from the second listening test. Root mean square errors between each subject’s localisation results and those stimuli played from a single loudspeaker.

Figure E.5: Inter-subject consistency for the source width results from the second listening test experiment. For each subject the Kendall rank-order correlation coefficient $T$ was calculated between the subject’s results and the mean results from the remaining subjects.
E.3 The third listening test

E.3.1 Localisation results

Testing for normality

The Shapiro-Wilk test was used to determine whether the localisation results from the third listening test experiment were normally distributed for each stimulus. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5\% level, and the combinations of stimulus and listening position for which the null hypothesis was rejected are contained in Table E.4. The null hypothesis was rejected for 9 of the 81 possible combinations, approximately 11\%. Consequently, non-parametric statistical tests were required to test for inter- and intra-subject consistency.

Testing for inter-subject consistency

The Kendall rank-order correlation $T$ was calculated between each subject's localisation results and the mean results for all the remaining subjects. The results of these calculations are shown in Fig. E.6. From this figure it can be seen that all the test subjects had results that were highly correlated with the results from the other subjects (the lowest value was 0.92) and that none of the subjects had a noticeably different correlation.

Comparing against stimuli with known source locations

Fig. E.7 shows the root-mean-square errors calculated for each test subject between their localisation angle results and the actual location angles for those stimuli played through a single loudspeaker (i.e. those stimuli which had a known source location). The average RMSE for the single loudspeaker stimuli across all the test subjects was 4.1°. Subjects nine and twelve had the largest RMSE values (5.6° and 6.3° respectively), which are still less than the RMSE values obtained by Stevens and Newman [154] (see Section E.1). Therefore, as the localisation results of subjects nine and twelve were also consistent with the localisation results from the other test subjects, the decision was made not to remove either subject's results.

Conclusion

Following the assessment of the inter-subject consistency and the assessment of the subjects' localisation results for those stimuli with known source locations, the decision was made not to remove the results for any of the test subjects from the third listening test experiment localisation results.
E.3.2 Source width results

Testing for normality

The Shapiro-Wilk test was used to determine whether the source width results from the third listening test experiment were normally distributed for each stimulus. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5% level, and the combinations of stimulus and listening position for which the null hypothesis was rejected are contained in Table E.5. The null hypothesis was rejected for 17 of the 81 possible combinations, approximately 21%. Consequently, non-parametric statistical tests were required to test for inter- and intra-subject consistency.

Testing for inter-subject consistency

The Kendall rank-order correlation $T$ was calculated between each subject's source width results and the mean results for all the remaining subjects. The results of these calculations are shown in Fig. E.8.

The average $T$ value was 0.57, and the $T$ values are much lower than those calculated using the localisation results from the third listening test experiment. Figs. 4.21 and 4.26 show that the source width results from the third listening test experiment have a relatively narrow range of values and that there is considerable variation in the source width results for each stimulus. This means that it is likely that the rank-orders of the source width results for each test subject will differ from each other, which is reflected in the lower Kendall correlation values. Although subject nine had a low $T$ value (0.46), their source width results were not removed, as there is a large variation for all the source width results from the third listening test experiment.

Conclusion

Following the assessment of the inter-subject consistency, the decision was made not to remove the results for any of the test subjects from the third listening test experiment source width results.
Table E.4: Results for Shapiro-Wilk test at the 5% level on the localisation results from the third listening test. The null hypothesis of the test is that the results have a normal distribution. The crosses in the table show those combinations of stimulus and listening position where the null hypothesis was rejected, i.e. the results did not have a normal distribution. Those stimuli that had normal results for all three listening positions have been omitted from the table.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Listening position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>

Figure E.6: Inter-subject consistency for the localisation results from the third listening test experiment. For each subject the Kendall rank-order correlation coefficient $T$ was calculated between the subject's results and the mean results from the remaining subjects.
Figure E.7: Comparison of the localisation results for each test subject with the stimuli with known locations from the third listening test. Root mean square errors between each subject's localisation results and those stimuli played from a single loudspeaker.
Table E.5: Results for Shapiro-Wilk test at the 5\% level on the source width results from the third listening test. The null hypothesis of the test is that the results have a normal distribution. The crosses in the table show those combinations of stimulus and listening position where the null hypothesis was rejected, i.e. the results did not have a normal distribution. Those stimuli that had normal results for all three listening positions have been omitted from the table.

Figure E.8: Inter-subject consistency for the localisation results from the third listening test experiment. For each subject the Kendall rank-order correlation coefficient $T$ was calculated between the subject’s results and the mean results from the remaining subjects.
E.4 Repeated combinations between the second and third listening tests

In the first listening test experiment the test subjects were required only to assess the directional localisation of each stimulus. In comparison, in each of the second and third listening test experiments the test subjects were required to assess the directional localisation and the directions of the left and right edges of each stimulus. Consequently, each stimulus took longer to assess in each of the second and third listening test experiments, due to the more complicated task given to the test subjects. It was also seen in Section E.1 that intra-subject consistency was high for the localisation results from the first listening test experiment. The same pool of listeners was used for all three listening test experiments. These two reasons (the longer time taken for the test subjects to complete the assigned task and the high intra-subject consistency seen in the results for the first listening test experiment) led to the decision not to repeat any of the stimuli in either the second or third listening test experiments.

However, there were a number of similarities between the second and third listening test experiments. Only the artificially widened stimuli were changed between the two experiments, so the stimuli played through a single loudspeaker and the stimuli generated using pair-wise panning were identical in both experiments. Although only the third listening test experiment used the two off-centre listening positions, both experiments used the centre listening position. Lastly, nine of the test subjects who participated in the second listening test experiment also participated in the third experiment at the centre listening position. Therefore, the results from the second and third experiments at the centre listening position and those stimuli which have not been artificially widened can be used to assess how these nine test subjects performed with repeated stimuli. In other words, the intra-subject consistency of these nine test subjects can be assessed for the centre listening position and those stimuli which have not been artificially widened.

E.4.1 Localisation results

Testing for normality

The Shapiro-Wilk test was used to determine whether the localisation results from the second and third listening test experiments where the listeners, listening position and stimuli were repeated were normally distributed. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5% level, and the null hypothesis was rejected for the localisation results for 6 of the 18 stimuli, namely stimuli numbers 1, 2, 9, 11, 20 and 21. As not all the results were normally distributed, non-parametric statistical tests were used to test for intra-subject consistency.
Testing for intra-subject consistency

Table E.6 shows the results of performing a Wilcoxon signed-rank test [48, 146] on the repeated stimuli. The null hypothesis for each of these Wilcoxon signed-rank tests was that there was no significant difference between the localisation results from the second and third listening test experiments, tested at the 5% level. As can be seen in the table, the null hypothesis was rejected by the results of three of the nine test subjects. The effect sizes, $r$, for subjects six, seven and eight are -0.62, -0.57 and -0.54 respectively, showing that these are all medium to large effects [48]. The eighteen stimuli that were used in both the second and third listening tests include the group of stimuli played through a single loudspeaker, where each stimulus has a known location angle, i.e. the angle of the loudspeaker relative to the listener. Table E.7 shows the root-mean-square error (RMSE) values of the localisation results from subjects six, seven and eight (whose results rejected the null hypothesis in the Wilcoxon signed-rank tests shown in Table E.6) compared with the actual angles of the single loudspeaker stimuli. This shows that these three subjects had more accurate localisation results in the third listening test compared to the second test. This supports the proposal made in Section 4.4.6 that the changes made to the experimental procedure have improved the results of the experiment and so account for the differences between the localisation results shown by the Wilcoxon signed-rank tests in Table E.6.

Conclusion

Following the assessment of the intra-subject consistency, the decision was made not to remove the results for any of the test subjects from the second and third listening test experiment localisation results.

E.4.2 Source width results

Testing for normality

The Shapiro-Wilk test was used to determine whether the localisation results from the second and third listening test experiments where the listeners, listening position and stimuli were repeated were normally distributed. The null hypothesis for each test was that the results were normal distributed. This was tested at the 5% level, and the null hypothesis was rejected for the localisation results for 5 of the 18 stimuli, namely stimuli numbers 6, 9, 14, 21 and 22. As not all the results were normally distributed, non-parametric statistical tests were used to test for intra-subject consistency.

Testing for intra-subject consistency

Table E.8 shows the results of performing a Wilcoxon signed-rank test on the repeated stimuli. The null hypothesis for each of these Wilcoxon signed-rank tests was that there was no significant difference between the source width results from the second and third listening test experiments, tested at the 5% level. As
Table E.6: Intra-subject consistency for the localisation results from the second and third listening test experiments. The results of performing the Wilcoxon signed-rank test on the localisation results from the two experiments having the same combination of test subject, listening position and stimulus.

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Test statistic, $T$</th>
<th>Null hypothesis accepted at 5% level</th>
<th>Effect size, $r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.5</td>
<td>✓</td>
<td>-0.27</td>
</tr>
<tr>
<td>2</td>
<td>24.5</td>
<td>✓</td>
<td>-0.27</td>
</tr>
<tr>
<td>3</td>
<td>47.0</td>
<td>✓</td>
<td>-0.26</td>
</tr>
<tr>
<td>4</td>
<td>13.5</td>
<td>✓</td>
<td>-0.48</td>
</tr>
<tr>
<td>5</td>
<td>43.5</td>
<td>✓</td>
<td>-0.10</td>
</tr>
<tr>
<td>6</td>
<td>8.0</td>
<td>×</td>
<td>-0.62</td>
</tr>
<tr>
<td>7</td>
<td>21.5</td>
<td>×</td>
<td>-0.57</td>
</tr>
<tr>
<td>8</td>
<td>20.0</td>
<td>×</td>
<td>-0.54</td>
</tr>
<tr>
<td>9</td>
<td>41.0</td>
<td>✓</td>
<td>-0.17</td>
</tr>
</tbody>
</table>

Table E.7: The root-mean-square error (RMSE) values between the localisation results from the test subjects and the location angles of those stimuli played through a single loudspeaker. Only the test subjects whose localisation results were judged by the Wilcoxon signed-rank test to be significantly different between the second and third listening tests have been included. Also, only the results from the centre listening position have been included.

<table>
<thead>
<tr>
<th>Subject</th>
<th>RMSE (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Second listening test experiment</td>
</tr>
<tr>
<td>6</td>
<td>5.1</td>
</tr>
<tr>
<td>7</td>
<td>3.7</td>
</tr>
<tr>
<td>8</td>
<td>5.6</td>
</tr>
<tr>
<td>Stimulus</td>
<td>Test statistic, T</td>
</tr>
<tr>
<td>----------</td>
<td>------------------</td>
</tr>
<tr>
<td>1</td>
<td>15.0</td>
</tr>
<tr>
<td>2</td>
<td>17.5</td>
</tr>
<tr>
<td>3</td>
<td>22.0</td>
</tr>
<tr>
<td>4</td>
<td>18.0</td>
</tr>
<tr>
<td>5</td>
<td>42.5</td>
</tr>
<tr>
<td>6</td>
<td>11.0</td>
</tr>
<tr>
<td>7</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>24.0</td>
</tr>
<tr>
<td>9</td>
<td>35.0</td>
</tr>
</tbody>
</table>

Table E.8: Intra-subject consistency for the source width results from the second and third listening test experiments. The results of performing the Wilcoxon signed-rank test on the source width results from the two experiments having the same combination of test subject, listening position and stimulus.

can be seen in the table, the null hypothesis was rejected by the source width results of the same three test subjects who rejected the null hypothesis of the Wilcoxon signed-rank test on the localisation results. In Section 4.4.6 it was discussed how the changes made to the experimental procedure between the second and third listening test experiments could account for the differences between the results of the two experiments. Similarly, in Section E.4.1 it was seen how there was an improvement in the localisation results between the second and third experiments Therefore, as there are possible explanations for the differences between the source width results for subjects six, seven and eight in the two listening test experiments, none of the source width results for these three test subjects will be removed.

Conclusion

Following the assessment of the intra-subject consistency, the decision was made not to remove the results for any of the test subjects from the second and third listening test experiment source width results.

APPENDIX E. EVALUATION OF LISTENING TEST SUBJECTS | 279
Appendix F

Errors in correlogram calculation

This appendix contains details of the mathematical derivations of some of the equations given in the discussion of the calculation of the correlograms in Section 5.1.5. Recall from Eq. 5.2 that the component correlogram at time $t$ is calculated as

$$C_M(t) = \begin{bmatrix}
L_{t+N}(-M) \times R_{t-N}(-M) \\
L_{t+N-1}(-M+1) \times R_{t-N+1}(-M+1) \\
L_{t+N-2}(-M+2) \times R_{t-N+2}(-M+2) \\
\vdots \\
L_{t+M-M} \times R_{t-M-M} \\
L_{t+M-M-1} \times R_{t-M-M-1} \\
L_{t+M-M} \times R_{t-M-M}
\end{bmatrix},$$

where $L_t$ and $R_t$ are the left and right binaural streams respectively, $M$ determines the length of the correlogram. The units of time are chosen to be the time between consecutive samples in the binaural streams. Also recall from Eq. 5.3 that each correlogram used to calculate an ITD is the sum of a given number, $N$, of consecutive component correlograms:

$$J_{M,N}(t) = \sum_{n=0}^{N-1} C_M(t+n)$$

(F.2)
Recall from Eq. 5.4 that the correlogram $H_k$ used to calculate the $k$th ITD value is given by

$$H_k = J_{9,14}(6k + c), \quad (F.3)$$

where $c$ is the initial offset used as an index into the two binaural streams. The Supper model makes use of the following property to reduce the number of calculations required:

$$H_k = \sum_{n=0}^{19} C_9((6k + c) + n)$$

$$= \sum_{n=0}^{19} C_9(6(k - 1) + 6 + c + n)$$

$$= \sum_{n=0}^{19} C_9(6(k - 1) + c + n) + \sum_{n=0}^{19} C_9(6(k - 1) + c + n)$$

$$\Rightarrow H_k = H_{k-1} - J_{9,6}(6k - 6 + c) + J_{9,6}(6k + 8 + c) \quad (F.4)$$

Recall from Eq. 5.6 that the original implementation of the Supper model included an error in the indexing:

$$H'_{k} = H'_{k-1} - J_{9,7}(6k - 6 + c) + J_{9,7}(6k + 7 + c) \quad (F.5)$$

### F.1 Proof by induction

This section proves the following equation,

$$H'_k = H_k - \sum_{n=1}^{k-1} C_9(6n + c) + \sum_{n=1}^{k-1} C_9(6n + 7 + c) \quad (F.6)$$

for all $k \in \{1, 2, 3, \ldots\}$. Before the proof by induction, note that Eq. F.5 gives

$$H'_{k} = H'_{k-1} - \sum_{n=0}^{6} C_9(6k - 6 + c + n) + \sum_{n=0}^{6} C_9(6k + 7 + c + n)$$

$$\Rightarrow H'_{k} = H'_{k-1} - J_{9,6}(6k - 6 + c) + J_{9,6}(6k + 8 + c) - C_9(6k + c) + C_9(6k + 7 + c) \quad (F.7)$$
The first step in the proof is to show that Eq. F.6 is true when \( k = 0 \) and \( k = 1 \). Now, \( H_0' \) is not affected by the indexing error, as it is not calculated from the previous ITD correlogram, giving \( H_0' = H_0 \). Using this and equations (F.7) and (5.5),

\[
H_1' = H_0 - J_{g,6}(6 \times 0 - 6 + c) + J_{g,6}(6 \times 0 + 8 + c) - C_g(6 \times 0 + c) + C_g(6 \times 0 + 7 + c)
\]

\[
= H_1 - C_g(6 \times 0 + c) + C_g(6 \times 0 + 7 + c)
\]

(F.8)

This has shown that Eq. F.6 is true when \( k = 0 \) and \( k = 1 \). The second step in the proof is to show that if Eq. F.6 is true for some \( k \in \{1, 2, 3, \ldots\} \), then it is also true for \( k + 1 \). If it is assumed that there exists some \( k \in \{1, 2, 3, \ldots\} \) such that Eq. F.6 is true, then

\[
H_{k+1}' = H_k' - J_{g,6}(6(k + 1) - 6 + c) + J_{g,6}(6(k + 1) + 8 + c) - C_g(6(k + 1) + c) + C_g(6(k + 1) + 7 + c)
\]

(by Eq. F.7)

\[
= H_k - \sum_{n=1}^{k-1} C_g(6n - 6 + c) + \sum_{n=1}^{k-1} C_g(6n + 7 + c)
- J_{g,6}(6(k + 1) - 6 + c) + J_{g,6}(6(k + 1) + 8 + c)
- C_g(6(k + 1) + c) + C_g(6(k + 1) + 7 + c)
\]

(by Eq. F.6)

\[
= H_{k+1} - \sum_{n=1}^{k} C_g(6n + c) + \sum_{n=1}^{k} C_g(6n + 7 + c)
\]

(by Eq. F.4) (F.9)

This has shown that if Eq. F.6 is true for some \( k \in \{1, 2, 3, \ldots\} \) then it is also true for \( k + 1 \). Therefore, as it has already been shown that Eq. F.6 is true for \( k \in \{0, 1\} \), using induction, Eq. F.6 is true for all \( k \in \{0, 1, 2, \ldots\} \). This concludes the proof.
Appendix G

Localisation plots for different reproduction systems at the sweet spot

This appendix contains figures showing the perceived localisation angles calculated by the model for the pink noise stimulus for different angles and different reproduction systems plotted on plan views of the listening area. The modelled listener was positioned at the sweet spot and facing forward (0°) for all the figures in the appendix.
Figure G.1: The perceived localisation angles calculated by the model for intended angles in the range 0° to 90° for TCS and FCS. The top left plot shows the results for the TCS signals created by modelling an ORTF microphone array. The top right hand plot shows the results for the TCS signals created using the constant power panning law: note here that the loudspeaker signals for intended angles greater than 30° were identical to the loudspeaker signals for the 30° intended angle, and consequently all these signals were localised by the model at the same angle. The bottom left plot shows the results for the FCS signals created by modelling a Williams-Du microphone array. The bottom right plot shows the results for the FCS signals created using the constant power panning law. The key at the bottom of each plot shows the relationship between the plotted locations and the intended angles.
Figure G.2: The top row shows the perceived localisation angles calculated by the model for intended angles in the range 0° to 90° for the mono (top left) and 32-channel WFS (top right). The bottom plot shows the perceived localisation angles calculated by the model for sources positioned 3m away from the sweet spot at angles in the range 0° to 90°, i.e. the original sound-field. The key at the bottom of each plot shows the relationship between the plotted locations and the intended angles.
Appendix H.

Stimuli used in the listening test

This section contains specific details of the stimuli used in the listening test.

H.1 Stimuli for Condition 1

Details of Condition 1 stimuli, including types, descriptions, and any associated metadata.
Appendix H

Stimuli used in the envelopment listening tests

This appendix contains details of the stimuli used by Conetta in his envelopment listening tests.

H.1 Stimuli for Conetta’s direct envelopment experiment

Tables D.1 and D.2 contain the details of the stimuli used in the first and second listening tests based on a modified MUSHRA method [132] respectively. Table D.3 contains the details of the stimuli used in the direct envelopment localisation experiment. Constant power panning was used for all the panned signals in these stimuli.
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Number of pan positions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced equally around the circular array (±22.5°, ±67.5°, ±112.5° and ±157.5°)</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±10°</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±25°</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±45°</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±90° (passing in front of the listener)</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±135° (passing in front of the listener)</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±90° (passing behind the listener)</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: -45°, 45°, 135° and -135°</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: 0°, 90°, 180° and -90°</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: -10°, -6.7°, 6.7° and 10°</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: -90°, -60°, 60° and 90°</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: -10°, 10°, -170° and 170°</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: 0° and 180°</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: -90° and 90°</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>1</td>
<td>8 different voices all positioned at 180°</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>4 different voices equally spaced equally around the circular array (±45° and ±135°)</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>4</td>
<td>4 different voices equally spaced between ±10°</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>2</td>
<td>2 different voices, one at 0° and the other at 180°</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>2</td>
<td>2 different voices, one at -90° and the other at 90°</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1 voice at 0°</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>6</td>
<td>8 different voices mixed to mono, the resulting signal then panned to the following positions: 0°, -60°, 60°, -120°, 120° and 180°</td>
</tr>
</tbody>
</table>

Table II.1: (Table continued on next page)

APPENDIX II. STIMULI USED IN THE ENVELOPMENT LISTENING TESTS | 290
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Number of pan positions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>8</td>
<td>6</td>
<td>4 different voices mixed to mono, then panned to 0°, -45°, 45°, -90° and 90°. 2 different voices mixed to mono, then panned to -135°. 2 different voices mixed to mono, then panned to 135°.</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>8</td>
<td>4 different voices mixed to mono, then panned to 0°, -45°, 45°, -90° and 90°. 4 different voices mixed to mono, then panned to -90°, 90°, -135°, 135° and 180°.</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>2</td>
<td>8 different voices mixed to mono, then panned to 0° and 180°.</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>2</td>
<td>8 different voices mixed to mono, then panned to -90° and 90°.</td>
</tr>
<tr>
<td>26</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front and rear loudspeakers, most enveloping.</td>
</tr>
<tr>
<td>27</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front and rear loudspeakers, least enveloping.</td>
</tr>
<tr>
<td>28</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front loudspeakers, background-type material in rear loudspeakers, most enveloping.</td>
</tr>
<tr>
<td>29</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front loudspeakers, background-type material in rear loudspeakers, least enveloping.</td>
</tr>
<tr>
<td>30</td>
<td>-</td>
<td>-</td>
<td>Real stereo (TCS) program material.</td>
</tr>
<tr>
<td>High anchor</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced around the circular array (0°, ±45°, ±90°, ±135° and 180°)</td>
</tr>
<tr>
<td>Low anchor</td>
<td>8</td>
<td>1</td>
<td>8 different voices mixed to mono, then panned to 0°.</td>
</tr>
</tbody>
</table>

Table II.1: Details of the stimuli used in Conetta’s first direct envelopment experiment. The voices were sourced from the anechoic recordings on the Music from Archimedes audio CD produced by Bang and Olufsen [3] on all stimuli except stimuli numbers 26 to 30. (Table continued from previous page)
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Number of pan positions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced equally around the circular array (±22.5°, ±67.5°, ±112.5° and ±157.5°)</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±10°</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±45°</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±90° (passing in front of the listener)</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±135° (passing in front of the listener)</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±90° (passing behind the listener)</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±135° (passing behind the listener)</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±170° (passing behind the listener)</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced between ±45° (passing behind the listener)</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>4</td>
<td>2 different voices each in the following positions: -45°, 45°, -135° and 135°</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: -135° and 135°</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: -45° and 45°</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: -45° and 135°</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>2</td>
<td>4 different voices each in the following positions: -90° and 90°</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>1</td>
<td>8 different voices all positioned at 180°.</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>4</td>
<td>4 different voices equally spaced equally around the circular array (±45° and ±135°)</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>2</td>
<td>2 different voices, one at -135° and the other at 135°.</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>2</td>
<td>2 different voices, one at -45° and the other at 45°.</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>2</td>
<td>2 different voices, one at -90° and the other at 90°.</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1 voice at 0°</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>8</td>
<td>1 voice panned to 8 positions equally spaced between 90°. (passing in front of the listener)</td>
</tr>
</tbody>
</table>

Table H.2: (Table continued on next page)
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Number of pan positions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>4</td>
<td>8</td>
<td>1 voice panned to 0° and 180°, 1 voice panned to −45° and 135°, 1 voice panned to −90° and 90° and 1 voice panned to −135° and 45°.</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>2</td>
<td>8 different voices mixed to mono, then panned to −90° and 90°.</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>2</td>
<td>8 different voices mixed to mono, then panned to −45° and 45°.</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>2</td>
<td>8 different voices mixed to mono, then panned to −135° and 135°.</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>-</td>
<td>High anchor downmixed to TCS (played through loudspeakers at ±45°).</td>
</tr>
<tr>
<td>27</td>
<td>8</td>
<td>-</td>
<td>High anchor downmixed to 3.0 (played through loudspeakers at 0° and ±45°).</td>
</tr>
<tr>
<td>28</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front and rear loudspeakers.</td>
</tr>
<tr>
<td>29</td>
<td>-</td>
<td>-</td>
<td>Real FCS program material, foreground-type material in front loudspeakers, background-type material in rear loudspeakers.</td>
</tr>
<tr>
<td>30</td>
<td>-</td>
<td>-</td>
<td>Real stereo (TCS) program material.</td>
</tr>
<tr>
<td>High anchor</td>
<td>8</td>
<td>8</td>
<td>8 different voices equally spaced around the circular array (0°, ±45°, ±90°, ±135° and 180°)</td>
</tr>
<tr>
<td>Low anchor</td>
<td>8</td>
<td>1</td>
<td>8 different voices mixed to mono, then panned to 0°.</td>
</tr>
</tbody>
</table>

Table II.2: Details of the stimuli used in Conetta’s second direct envelopment experiment. The voices were sourced from the anechoic recordings on the Music from Archimedes audio CD produced by Bang and Olufsen [3] on all stimuli except stimuli 26 to 30. The voices used were different to those used in the first direct envelopment experiment. (Table continued from previous page)
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Corresponding stimulus in Table D.1</th>
<th>Decomposed</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>×</td>
<td>8 different voices equally spaced equally around the circular array (±22.5°, ±67.5°, ±112.5° and ±157.5°)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>×</td>
<td>8 different voices equally spaced between ±10°</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>×</td>
<td>8 different voices equally spaced between ±90° (passing in front of the listener)</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>×</td>
<td>2 different voices each in the following positions: -45°, 45°, 135° and -135°</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>×</td>
<td>4 different voices each in the following positions: 0° and 180°</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>×</td>
<td>4 different voices each in the following positions: -90° and 90°</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>×</td>
<td>8 different voices mixed to mono, the resulting signal then panned to the following positions: 0°, -60°, 60°, -120°, 120° and 180°</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>×</td>
<td>8 different voices mixed to mono, then panned to -90° and 90°.</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to -22.5°</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to -67.5°</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to -112.5°</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to -157.5°</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to 157.5°</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to 112.5°</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to 67.5°</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>✓</td>
<td>1 voice panned to 22.5°</td>
</tr>
</tbody>
</table>

Table II.3: Details of the stimuli used in Conetta’s direct envelopment localisation experiment. The first eight stimuli are identical to stimuli used in Conetta’s first direct envelopment experiment.
Figure H.1: The equipment layout used in Conetta's indirect envelopment experiment. Eight loudspeakers were arranged in an equally spaced circular array with radius 2.2m. The wavy lines represent the acoustically transparent curtain used to hide the positions of the loudspeakers from the listening test subjects.

H.2 Stimuli for Conetta's indirect envelopment experiment

This section describes the stimuli used in Conetta's indirect envelopment experiment. First a model of a large reverberant rectangular room was created in CATT-Acoustics: the dimensions of the room were 100m long, 20m wide and 30m high and the absorption coefficients were chosen to match those of an audience for the floor and marble for all the other surfaces. To the model was added a circular array of eight equally spaced omni-directional microphones centered in the middle of the room with a radius of 5m and a height of 1m. These 8 microphones correspond to the eight loudspeakers used in Conetta's indirect envelopment listening test, with the microphone array orientated to face down the length of the hall. A sound source was modelled in the room 20m to the front and 6m to right of the centre of the room. A set of eight impulse responses were then calculated from the source to each of the eight modelled microphones. The signals to the loudspeakers for the majority of the stimuli were then created by convolving an original signal with each loudspeaker's corresponding impulse response calculated in the CATT-Acoustics model.

The original signals used were either a single voice or a mix of eight different voices sourced from the anechoic recordings on the Music from Archimedes audio CD produced by Bang and Olufsen [3]. In both cases the original signals were low pass filtered at 6kHz before being convolved with the impulse responses. This was done to remove sibilant artifacts from the convolved stimuli. The majority of the indirect envelopment stimuli were created in this way, the differences between the stimuli being due to the combinations of loudspeakers used. Table D.4 contains the details of the stimuli. The method of construction of the remaining stimuli is described in the Notes column of the table. Each of the loudspeaker signals for stimulus 30 were created by combining the signals convolved with the impulse response corresponding to the loudspeaker together with the signal convolved with the impulse response corresponding to loudspeaker 1 (see Fig. H.1). This resulted in the signals in the four loudspeakers being partially correlated.
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Loudspeakers</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>All loudspeakers, 360° coverage.</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>(1, 2, 8)</td>
<td>Convolved signals treated as independent of loudspeakers and panned to 8 locations equally spaced between ±10°.</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>(1, 2, 8)</td>
<td>Convolved signals treated as independent of loudspeakers and panned to 8 locations equally spaced between ±30°.</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>(1, 2, 8)</td>
<td>Convolved signals treated as independent of loudspeakers and panned to 8 locations equally spaced between ±45°.</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>1, 2, 3, 7, 8</td>
<td>Front semicircle.</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>1, 2, 3, 7, 8</td>
<td>Rear semicircle.</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1, 2, 3, 4, 6, 7, 8 (no 5)</td>
<td>No rear loudspeaker.</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>4, 5, 6</td>
<td>Rear three loudspeakers.</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>1, 5, 6, 7, 8</td>
<td>Left semicircle.</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>2, 4, 6, 8</td>
<td>Loudspeakers at ±45° and ±135°.</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>1, 3, 5, 7</td>
<td>Loudspeakers at 0°, ±90° and 180°.</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>Loudspeaker at 0°.</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td>5</td>
<td>Loudspeaker at 180°.</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>7</td>
<td>Loudspeaker at -90°.</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>6, 8</td>
<td>Loudspeakers at -45° and -135°.</td>
</tr>
<tr>
<td>16</td>
<td>8</td>
<td>1, 2</td>
<td>Loudspeakers at 0° and 45°.</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>4, 6</td>
<td>Loudspeakers at ±135°.</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>1, 5</td>
<td>Loudspeakers at 0° and 180°.</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>3, 7</td>
<td>Loudspeakers at ±90°.</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor, high pass filtered at 500Hz.</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor, low pass filtered at 2kHz.</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor with +6dB gain.</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor with -6dB gain.</td>
</tr>
<tr>
<td>24</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor, all loudspeakers, 360° coverage.</td>
</tr>
<tr>
<td>25</td>
<td>8</td>
<td>1</td>
<td>Low anchor, loudspeaker at 0°.</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>1, 2, 4, 6, 8</td>
<td>Approximation of FCS, loudspeakers at 0°, ±45° and ±135°.</td>
</tr>
<tr>
<td>27</td>
<td>8</td>
<td>(1, 2, 4, 6, 8)</td>
<td>Stimulus 26 processed by the AAC+ MPEG codec at 64kbps.</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>(1, 2, 3, 4, 5, 6, 7, 8)</td>
<td>The same impulse response used to convolve the signals for all 8 loudspeakers (i.e. all loudspeakers are identical).</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>(1, 2, 3, 4, 5, 6, 7, 8)</td>
<td>The same impulse response used to convolve the signals for all 8 loudspeakers (i.e. all loudspeakers are identical).</td>
</tr>
<tr>
<td>30</td>
<td>8</td>
<td>(2, 4, 6, 8)</td>
<td>Partially correlated signals in loudspeakers at 45° and 135°: see text in Section II.2.</td>
</tr>
</tbody>
</table>

Table II.4: (Table continues on next page)
<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Number of voices</th>
<th>Loudspeakers</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>High anchor</td>
<td>8</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
<td>High anchor, all loudspeakers, 360° coverage.</td>
</tr>
<tr>
<td>Low anchor</td>
<td>8</td>
<td>1</td>
<td>Low anchor, loudspeaker at 0°.</td>
</tr>
</tbody>
</table>

Table H.4: Details of the stimuli used in Conetta's indirect envelopment experiment. The voices were sourced from the anechoic recordings on the Music from Archimedes audio CD produced by Bang and Olufsen [3]. The numbers in the Loudspeakers column refer to the labels on the loudspeakers in Fig. II.1. When the numbers in the Loudspeakers column are not in parenthesis then each loudspeaker signal in the stimulus was created by convolving the voices with the loudspeaker's corresponding impulse response, as described in Section II.2. When the numbers in the Loudspeakers column are in parenthesis then the method used to create the stimulus is described in the Notes column. (Table continued from previous page)
Appendix I

The linear regression models for predicting envelopment and multicollinearity

This appendix contains tables showing the Variance Inflation Factors (VIFs) calculated for all the linear regression models predicting the data from Conetta's direct and indirect envelopment listening test experiments. Multicollinearity is judged to be a problem if the VIF values for individual metrics are greater than 10 or the average VIF values are greater than 6. From Tables I and I it can be seen that multicollinearity is not an issue for the linear regression models for predicting direct and indirect envelopment from Chapter 7.
### Table I.1: The Variance Inflation Factors (VIFs) calculated for each of the metrics in the linear regression models predicting the data from Conetta's MUSTEA direct envelopment listening test experiments. The standardised coefficients and $R^2$ and RMSEP values for these regression models are contained in Table 7.3. The first column contains the number of the metric set, which corresponds to the row containing the metric set in Table 7.3.

<table>
<thead>
<tr>
<th>Metric set</th>
<th>Number of metrics</th>
<th>IACC0</th>
<th>IACC0*IACC90</th>
<th>TQEnergy</th>
<th>EntropyL</th>
<th>CardsKUT</th>
<th>c90-30</th>
<th>std.id</th>
<th>std.id</th>
<th>std.times</th>
<th>Average VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1.85</td>
<td>-</td>
<td>1.04</td>
<td>1.14</td>
<td>-</td>
<td>2.07</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.83</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.96</td>
<td>-</td>
<td>1.10</td>
<td>1.08</td>
<td>2.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.44</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4.07</td>
<td>4.05</td>
<td>1.11</td>
<td>1.06</td>
<td>2.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.49</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1.96</td>
<td>-</td>
<td>1.08</td>
<td>1.05</td>
<td>2.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.53</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3.70</td>
<td>3.83</td>
<td>1.07</td>
<td>1.04</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.41</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1.17</td>
<td>-</td>
<td>1.06</td>
<td>1.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.20</td>
<td>1.11</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>1.02</td>
<td>-</td>
<td>1.02</td>
<td>1.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.02</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1.94</td>
<td>-</td>
<td>1.02</td>
<td>1.91</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.63</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>1.01</td>
<td>-</td>
<td>1.01</td>
<td>-</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.01</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1.43</td>
<td>-</td>
<td>1.02</td>
<td>1.41</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.29</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1.01</td>
<td>-</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.01</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
</tr>
<tr>
<td>Metric set</td>
<td>Number of metrics</td>
<td>IACC0</td>
<td>IACC0+IACC90</td>
<td>TotEnergy</td>
<td>EntropyL</td>
<td>CardKLT</td>
<td>c90.30</td>
<td>sd.ltd</td>
<td>sd.ltd</td>
<td>sd.plus</td>
<td>Average VIF</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>-------</td>
<td>--------------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>4.98</td>
<td>5.97</td>
<td>1.36</td>
<td>-</td>
<td>2.51</td>
<td>-</td>
<td>1.13</td>
<td>-</td>
<td>-</td>
<td>3.19</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2.40</td>
<td>-</td>
<td>-</td>
<td>1.12</td>
<td>3.07</td>
<td>2.25</td>
<td>-</td>
<td>1.54</td>
<td>-</td>
<td>2.08</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4.94</td>
<td>6.30</td>
<td>4.60</td>
<td>3.55</td>
<td>2.59</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.40</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2.32</td>
<td>-</td>
<td>-</td>
<td>1.03</td>
<td>2.46</td>
<td>-</td>
<td>-</td>
<td>1.07</td>
<td>-</td>
<td>1.72</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4.60</td>
<td>5.26</td>
<td>1.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>3.08</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2.36</td>
<td>-</td>
<td>1.25</td>
<td>-</td>
<td>2.48</td>
<td>-</td>
<td>-</td>
<td>1.13</td>
<td>-</td>
<td>1.81</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>2.48</td>
<td>-</td>
<td>-</td>
<td>1.01</td>
<td>-</td>
<td>1.01</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1.19</td>
<td>-</td>
<td>1.21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.05</td>
<td>-</td>
<td>1.15</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>1.20</td>
<td>-</td>
<td>1.19</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.08</td>
<td>-</td>
<td>1.15</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>1.06</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
<td>2.48</td>
<td>-</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>1.04</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1.01</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.01</td>
<td>-</td>
<td>1.01</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1.03</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.03</td>
<td>-</td>
<td>1.03</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>1.06</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.06</td>
<td>-</td>
<td>1.06</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1.00</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table I.2: The Variance Inflation Factors (VIFs) calculated for each of the metrics in the linear regression models predicting the data from Conetta's MUSTEA indirect envelopment listening test experiment. The standardised coefficients and $R^2$ and RMSEP values for these regression models are contained in Table 7.4. The first column contains the number of the metric set, which corresponds to the row containing the metric set in Table 7.4.
Bibliography


BIBLIOGRAPHY | 305


BIBLIOGRAPHY | 306


BIBLIOGRAPHY | 307


BIBLIOGRAPHY | 311


BIBLIOGRAPHY | 313