Singing in Space(s): Singing performance in real and virtual acoustic environments—Singers' evaluation, performance analysis and listeners' perception.

Judith Sara Brereton, BA(Hons), MPhil

PhD

University of York
Electronics

August 2014

Abstract

The Virtual Singing Studio (VSS), a loudspeaker-based room acoustic simulation, was developed in order to facilitate investigations into the correlations and interactions between room acoustic characteristics and vocal performance parameters. To this end, the VSS provides a virtual performance space with interactivity in real-time for an active sound source - meaning that singers can hear themselves sing as if in a real performance space.

An objective evaluation of the simulation was carried out through measurement and comparison of room acoustic parameters of the simulation and the real performance space. Furthermore a subjective evaluation involved a number of professional singers who sang in the virtual and real performance spaces and reported their impressions of the experience. Singing performances recorded in the real and virtual spaces were compared via the analysis of tempo, vibrato rate, vibrato extent and measures of intonation accuracy and precision.

A stimuli sorting task evaluated listeners' perception of the similarity between singing performances recorded in the real and simulated spaces. A multi-dimensional scaling analysis was undertaken on the data obtained and dimensions of the common perceptual space were identified using property fitting techniques in order to assess the relationship between performance attributes and the perceived similarities. In general significant proportions of the perceived similarity between recordings could be explained by differences in global tempo, vibrato extent and intonation precision. Although there were few statistically significant effects of room acoustic condition all singers self-reported changes to their singing according to the different room acoustic configurations, and listeners perceived these differences, especially in vibrato extent and global tempo.

The present Virtual Singing Studio (VSS) has been shown to be not fully "realistic" enough to elicit variations in singing performance according to room acoustic conditions. Therefore, further improvements are suggested including the incorporation of visual aspect to the simulation. Nonetheless, the VSS is already able to provide a "plausible" interactive room acoustic simulation for singers to hear themselves in real-time as if in a real performance venue.

Contents

Abstract i

List of Tables viii

List of Figures xi

Dedication xix

Acknowledgments xx

Declaration xxii

1 Introduction 1

1.1 Investigating Musical Performance . . . 2

1.1.1 Traditional Auralisation Methods . . . 2

1.1.2 Interactive Room Acoustics Simulation . . . 2

1.2 Hypothesis . . . 3

1.3 Novelty of research . . . 4

1.4 Main contributions . . . 4

1.5 Structure of Thesis . . . 6

2 Simulating Room Acoustics 8

2.1 Introduction . . . 8

2.2 Acoustics . . . 8

2.3 Room Acoustics . . . 10

2.3.1 Sound propagation in rooms . . . 10

2.3.2 Measuring Room Impulse Responses . . . 13

2.3.3 Objective room acoustic parameters . . . 14

2.4 Room Acoustics and Musical Performance . . . 17

2.4.1 Perceptual evaluation of concert hall acoustics . . . 17

2.4.2 Measuring Stage Acoustic Parameters . . . 20

2.4.3 Perceptual Stage Acoustic Parameters . . . 21
2.4.4 Performance spaces and musical style . . . 24
2.4.5 Adjustable room acoustics . . . 25
2.5 Auralisation . . . 28
2.5.1 Auralisation chain . . . 29
2.5.2 Sound Source . . . 29
2.5.3 Room . . . 31
2.5.4 Convolution . . . 33
2.5.5 Rendering and Reproduction . . . 34
2.5.6 Evaluation of Auralisation . . . 35
2.6 Room Acoustic Simulations for Musical Performance . . . 37
2.6.1 Virtual Acoustic Environments . . . 37
2.6.2 Interactive Room Acoustic Simulations . . . 38
2.6.3 Interactive Room Acoustic Simulations for Musical Performance . 39
2.6.4 Evaluation of real-time room acoustic simulations . . . 42
2.7 Summary . . . 43

3 The Virtual Singing Studio- Implementation and Verification 45
3.1 Introduction . . . 45
3.2 Prototype Virtual Singing Studio . . . 46
3.2.1 Methods and materials . . . 46
3.2.2 The Performance Space . . . 46
3.2.3 Measured and Modelled Spatial Room Impulse Response (SRIR) . 47
3.2.4 Acoustic Characteristics of the Space . . . 47
3.2.5 Editing the Impulse Responses . . . 48
3.2.6 Real-time Acoustic Simulation . . . 48
3.2.7 Experimental Protocol . . . 49
3.2.8 Results . . . 50
3.2.9 Discussion . . . 52
3.2.10 Summary . . . 52
3.3 The Real Performance Space . . . 53
3.3.1 Room acoustic configurations . . . 53
3.4 The Virtual Singing Studio Implementation . . . 55
3.4.1 Measurement of Spatial Room Impulse Responses (SRIR) . . . 56
3.4.2 VSS Implementation . . . 58
3.4.3 Capturing the voice signal . . . 60
3.4.4 Rendering the Soundfield . . . 65

3.4.5 Latency . . . 68
3.4.6 Editing SRIRs . . . 70
3.4.7 Calibration . . . 72
3.5 Verification of The Virtual Performance Space . . . 75
3.5.1 Method . . . 75
3.5.2 Room acoustic parameters of Real Performance Space . . . 76
3.5.3 Room acoustic parameters of Virtual Performance Space . . . 82
3.5.4 Comparison of Real and Virtual Performance Space . . . 82
3.5.5 Summary . . . 91
3.6 Singers' Evaluation of the VSS . . . 92
3.6.1 Method . . . 92
3.6.2 Results of questionnaire . . . 92
3.6.3 Discussion of results . . . 94
3.7 Conclusion . . . 95

4 Singing in space(s) 96
4.1 Introduction . . . 96
4.2 Music Performance Research . . . 97
4.2.1 History of Music Performance Research . . . 97
4.2.2 Musical Score and Performance . . . 100
4.2.3 Musical Structure and Performance . . . 100
4.2.4 Music Performance Studies . . . 101
4.2.5 The Singing Voice . . . 102
4.2.6 Vocal Performance Analysis . . . 103
4.3 Music Performance Analysis . . . 104
4.3.1 Timbral Parameters . . . 105
4.3.2 Timbral Attributes . . . 105
4.3.3 Vocal Timbral Attributes . . . 106
4.3.4 Production-related Parameters . . . 110
4.3.5 Vocal Production-related Parameters . . . 110
4.3.6 Intensity-related Parameters . . . 111
4.3.7 Intensity-related Attributes . . . 112
4.3.8 Temporal Parameters . . . 112
4.3.9 Temporal Attributes . . . 115
4.3.10 Tonal Parameters . . . 117
4.3.11 Tonal Attributes . . . 118
4.3.12 Summary of Music Performance Analysis . . . 127

4.4 Room Acoustics and Musical Performance . . . 127
4.4.1 Room acoustics and musical style . . . 128
4.4.2 Importance of aural feedback for musical performance . . . 129
4.4.3 Musicians' Preferences for Room Acoustic Conditions . . . 130
4.4.4 Influence of room acoustics on musical performance . . . 132
4.4.5 Evaluating room acoustics through analysis of performance . . . 133
4.4.6 Conceptual models of performance . . . 137
4.5 Room Acoustics and Speech . . . 138
4.6 Room Acoustics and Singing Performance . . . 139
4.6.1 Singers are special . . . 140
4.6.2 Aural Feedback For Phonatory Control . . . 140
4.6.3 Singing and Listening . . . 141
4.6.4 Singing Performance in Different Room Acoustic Conditions . . . 142
4.7 Case Study I: Quartet singing in the *Real Performance Space* . . . 144
4.7.1 Quartet Singers' subjective responses . . . 145
4.8 Summary . . . 146

5 Singing Performance Analysis and Evaluation 149
5.1 Introduction . . . 149
5.2 Perceptual Evaluation of Audio and Music . . . 149
5.2.1 Psychoacoustic Evaluation Methods . . . 150
5.2.2 Multi-variate Data Analysis . . . 153
5.2.3 Subjective Evaluation of Musical Performance . . . 157
5.3 Correlating Objective and Perceptual Attributes . . . 158
5.3.1 Regression Analysis . . . 159
5.3.2 Principal Component Analysis . . . 159
5.3.3 Music Performance Attributes and Perceptual Correlates . . . 160
5.4 Pilot Listening Test I - Producing Stimuli . . . 161
5.5 Pilot Listening Test II - Quartet Performances . . . 163
5.5.1 Method . . . 164
5.5.2 Results . . . 166
5.5.3 Discussion . . . 168
5.6 Summary . . . 171

6 Singing in Real and Virtual Acoustic Environments 173
6.1 Introduction . . . 173
6.2 Recording in real and virtual performance spaces . . . 174
6.2.1 Method . . . 174

6.2.2 Singer's Evaluation of Solo Singing Performance 175
6.3 Listeners' Evaluation of Solo Singing Performances 176
6.3.1 Producing Stimuli 179
6.3.2 Procedure 180
6.3.3 Data Analysis 180
6.3.4 Results: Test 232a Bass 181
6.3.5 Results: Test 232b Bass 183
6.3.6 Results: Test 221 Tenor 188
6.3.7 Results: Test 212 Mezzo-Soprano 192
6.3.8 Limitation of Sorting Task 198
6.3.9 Summary of listeners' evaluation 199
6.4 Acoustic Analysis of Solo Singing Performances 201
6.4.1 Method 202
6.4.2 Analysis of Tempo 205
6.4.3 Analysis of Intonation 207
6.4.4 Analysis of Vibrato 211
6.5 Correlation of Performance Attributes and Perceptual Evaluation 217
6.5.1 Method 217
6.5.2 Results 219
6.5.3 Test 232a Bass 219
6.5.4 Test 232b Bass 220
6.5.5 Test 221 Tenor 222
6.5.6 Test 212 Mezzo-Soprano 224
6.6 Testing effect of Acoustic Setting and Simulation 227
6.6.1 ANOVA on performance attributes 228
6.6.2 ANOVA on similarity ratings 229
6.6.3 Discussion 229
6.7 Summary 230
6.7.1 Acoustic Configurations 231
6.7.2 Real vs. Virtual 232
6.7.3 Emotional Content 234
6.7.4 Unexplained differences 234
6.7.5 Limitations of the study - future work 235
6.8 Conclusions 235

7 Conclusion **236**
7.0.1 Summary of Thesis 236

7.0.2 Restatement of Hypothesis . . . 239
7.1 Further work . . . 241
7.1.1 Improving the VSS . . . 241
7.1.2 Music Performance Analysis . . . 243
7.2 Application of research . . . 244
7.2.1 Vocal health . . . 244
7.2.2 Concert hall and other architectural design . . . 244
7.2.3 Improving SRIR models . . . 245
7.2.4 Real-time convolution . . . 245
7.2.5 Psychoacoustics . . . 245
7.2.6 Application to Virtual Reality Research . . . 245
7.3 Final Remarks . . . 246

A Index of Supporting Media DVD 248

B Protocol and questionnaire for initial experiment 250

C Instructions for Participants in Main Listening Test 254

D Room Acoustic Parameters of Real Performance Space 258

E Room Acoustic Parameters of Virtual Performance Space 260

F Comments by singers about the performance spaces 262

G Lyrics of recorded fragments 266

H Comments on fragments by listeners 269

I Goodness of fit of MDS solutions in Chapter 6 273

J Vibrato Analysis 276

K Intonation Metrics 281

L Tempo Analysis 285

Acronyms 289

References 291

List of Tables

3.1 Average difference of sung note from mean measured F0 for each note class (cents) and standard deviation (cents) with system turned on and off . . . 51
3.2 Summary of acoustic configurations in the real performance space 55
3.3 List of room acoustic parameters evaluated in the *real performance space* and the simulation (VSS values from [126]) 76

4.1 Table of fundamental frequency values and the difference between notes of the C major scale in equal temperament and just intonation 120
4.2 Summary of three acoustic configurations used in the *real performance space* 144

5.1 Stress and goodness of fit using Kruskal's Stress measure 167

6.1 Summary of performance changes identified by the singers in the different acoustic configurations of the real and virtual performance spaces 176
6.2 Names of Fragments in test 232a 177
6.3 Names of Fragments in test 232b 178
6.4 Names of Fragments in test 221 178
6.5 Names of Fragments in test 212 179
6.6 Mean (St Dev) values of Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) for each test (cents) 210
6.7 Average Mean, standard deviation, median, maximum and minimum values of vibrato rate (Hz) and vibrato extent (cents) across all fragments in each test 214
6.8 Regression analysis between performance attributes and positions in MDS perceptual map for Test 232a 220
6.9 Regression analysis between performance attributes and positions in MDS perceptual map for Test 232b 221
6.10 Regression analysis between performance attributes and positions in MDS perceptual map for Test 221 222

6.11 Regression analysis between performance attributes and positions in 3-D MDS solution for Test 221: dimensions 1 and 3 (boldface $p<0.1$, * $p<.05$, ** $p<.01$) 223
6.12 Regression analysis between performance attributes and positions in MDS perceptual map for Test 212 224
6.13 Regression analysis between performance attributes and positions in MDS perceptual map for Test 212 224
6.14 Results of linear correlation between performance parameters and acoustic configuration, performance space, and song verse 227
6.15 Results of ANOVA on effect of *acoustic* (LC/MR/SP) and *simulation* (virtual vs real) on performance parameters 228
6.16 Results of ANOVA on effect of *acoustic* (congruent vs non-congruent) and *simulation* (congruent vs non-congruent) on similarity ratings of pairs of fragments 229

D.1 Mean values of Early Decay Time (EDT), Reverberation Time (T30), - Early Stage Support (ST_{early}), Late Stage Support (ST_{late}), Total Stage Support (ST_{total}) and Running Reverberation (RR160) averaged across the four performer positions the Large Choral Setting of the Real Performance Space 258
D.2 Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Music Recital Setting of the Real Performance Space 259
D.3 Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Speech Setting of the Real Performance Space 259

E.1 Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Large Choral Setting of the Virtual Performance Space 260
E.2 Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Music Recital Setting of the Virtual Performance Space 260
E.3 Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Speech Setting of the Virtual Performance Space 261

J.1 Vibrato Rate (Hz) and Vibrato Extent (Cents) for notes of phrase Test 232a 276

J.2 Vibrato Rate (Hz) and Vibrato Extent (Cents) for each note of phrase in Test 232b; Mean values include only notes in the phrase with 4 vibrato cycles or more 277
J.3 Vibrato Rate (Hz) and Vibrato Extent (Cents) for notes of phrase Test 221; Mean values include only notes in the phrase with 4 vibrato cycles or more 278
J.4 Vibrato Rate (Hz) for each note of phrase in Test 212 279
J.5 Vibrato Extent (Cents) for each note of phrase in Test 212; Mean values include only notes in the phrase with 4 vibrato cycles or more 280

K.1 Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 232a 281
K.2 Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 232b 282
K.3 Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 221 283
K.4 Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 212 284

L.1 Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 232a . 285
L.2 Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 232b . 286
L.3 Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 221 . . 287
L.4 Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 212 . . 288

List of Figures

2.1 Illustration of the direct sound in a room 10
2.2 Illustration of the early reflections in a room 11
2.3 Illustration of reverberant sound in a room 12
2.4 Active wall system in rehearsal room for symphony orchestra 27
2.5 Representation of image source model 32
2.6 Schematic Diagram of Auralisation Chain 38

3.1 Interior, as viewed from altar, of St. Patrick's Church, Patrington 46
3.2 Mean values of T30 for synthesized SRIR (model) compared with T30 values of measured SRIR of the same space 47
3.3 Graphical representation of real-time convolution process in proto-type Virtual Singing Studio 48
3.4 Illustration of hexagonal loudspeaker array and performer position 49
3.5 Participant in trials of the prototype VSS wearing headset microphone and electrolaryngograph 49
3.6 Average ratings (with standard deviations) by singers for the acoustic simulations: left hand bar in each pair represents the "measured space", right hand bar in each pair represents "modelled space" 51
3.7 Interior of National Centre for Early Music Prior to refurbishment 53
3.8 Concert at National Centre for Early Music 54
3.9 Position of a number of acoustic panels on the back wall 54
3.10 Floor plan of National Centre for Early Music, York 55
3.11 Singer positions relative to back wall 56
3.12 Relative performance positions of vocal quartet indicated by music stand placement 57
3.13 Illustration of Listener SRIR measurement position 58
3.14 Photograph to illustrate listener position Soundfield microphone position at head-height of seated audience member 58

3.15 Position of Soundfield Microphone placed above the Genelec 8040 loudspeaker used to measure performer position SRIRs in the real performance space 59
3.16 Graphical representation of the processing chain involved in the VSS . . 59
3.17 Frequency responses of AudioTechnica Pro45 Cardiod and DPA 4066 microphones 61
3.18 Spectrograms of sung phrase recorded at head-mounted (upper panel) and overhead (lower panel) microphones. 62
3.19 Long-term Average Spectrum (LTAS) of sung phrase recorded at overhead, head-mounted and 1m distance microphones 63
3.20 Difference between long-term average spectra of the overhead (20cm from mouth) and head-mounted (5cm from mouth) microphone recordings of the phrase "Peter Piper" recorded simultaneously 64
3.21 Graphical representation of the position of loudspeakers in the 3 dimensional loudspeaker array 66
3.22 Photograph of the Virtual Singing Studio 67
3.23 Additional acoustic absorption treatment for ceiling of Virtual Singing Studio 67
3.24 Block diagram of system latency measurement illustrating input microphone (DPA4066), VSS processing (in rectangle with dashed border) running through Reaper DAW, microphone inputs and loudspeaker outputs via RME Fireface800 soundcard. 68
3.25 Photograph of microphone placement used to capture balloon pop in measuring end-to-end latency of the system 69
3.26 Diagrammatic representation of loudspeaker and microphone topology . . 70
3.27 Test signal to illustrate processing of RIRs for use in the VSS 71
3.28 Arrangement of source loudspeaker (Genelec 8040) with DPA4066 microphone in front and Soundfield microphone above as used in the calibration experiment 73
3.29 Comparison of ST_{late} values of the *real performance space* and *virtual performance space* at different decoder output levels. 74
3.30 Comparison of ST_{second} values of the *real performance space* and *virtual performance space* at different decoder output levels. 74
3.31 Mean T30 and EDT values evaluated in the Large Choral setting of the Real Performance Space as as measured in the four performer positions . 77
3.32 Mean of Support and RR160 values of the four performer positions evaluated in the Large Choral (LC) setting of the *Real Performance Space* 77

3.33 Mean T30 and EDT values of the four performer positions evaluated in the Music Recital (MR) setting of the *Real Performance Space* 78
3.34 Mean Support and RR160values of the four performer positions evaluated in the Music Recital setting of the *Real Performance Space* 78
3.35 Mean T30 and EDT values of the four performer positions evaluated in the Speech (SP) setting of the *Real Performance Space* 79
3.36 Mean Support and RR160values of the four performer positions evaluated in the Speech (SP) setting of the *Real Performance Space* 79
3.37 Mean T30 values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP) 80
3.38 Mean Early Support values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP) 81
3.39 Mean Late Support values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP) 81
3.40 Comparison of T30 values measured at Listener Position and Performer Position B across seven octave bands in Music Rectial Setting 82
3.41 Mean EDT and T30 values evaluated in the Large Choral setting of the Virtual Performance Space as simulated for the four performer positions . 83
3.42 Mean Support and RR160 values evaluated in the Large Choral setting of the Virtual Performance Space as simulated for the four performer positions 83
3.43 Mean EDT and T30 values of the four performer positions evaluated in the Music Recital (MR)setting of the Virtual Performance Space 84
3.44 Mean Support and RR160 values of the four performer positions evaluated in the Music Recital (MR) setting of the *Virtual Performance Space* . . 84
3.45 Mean EDT and T30 values of the four performer positions evaluated in the Speech (SP) setting of the *Virtual Performance Space* 85
3.46 Mean Support and RR160 values of the four performer positions evaluated in the Speech (SP) setting of the *Virtual Performance Space* 85
3.47 Differences between virtual and *real performance space* EDT values in performer position, for octave bands 125hz - 8000Hz 86
3.48 Differences between virtual and *real performance space* T30 values in performer position, for octave bands 125hz - 8000Hz. 86
3.49 Differences between virtual and *real performance space* Early Support values in performer position, for octave bands between 125hz - 8000Hz 87

3.50 Differences between virtual and *real performance space* Late Support values in performer position, for octave bands 125hz - 8000Hz. 87
3.51 Differences between virtual and *real performance space* Total Support in the performer position, for octave bands 125hz - 8000Hz. 88
3.52 Differences between virtual and *real performance space* RR160 values in the performer position, for octave bands from 125hz - 8000Hz. 88
3.53 Spectrograms of impulse responses measured in the *real performance space* and *virtual performance space* 90
3.54 Mean scores (and standard error) of singers' responses to questionnaire on the Virtual Singing Studio 92
3.55 Mean scores (and standard error) of singers' responses to questionnaire on the Real Performance Space 93

4.1 The impact of the sound modifiers on the voice source sound spectrum to create formant peaks when producing an "ah" vowel 107
4.2 The impact of the *singers formant cluster* on an idealized spectral envelope of an orchestra and singer from [20]. 108
4.3 The spacing of the harmonics at different fundamental frequencies changing the impact of the formant frequencies on the spectral envelope 109
4.4 Illustration of the note onset of a saxophone note 113
4.5 Graphical representation of hypothetical examples of repeated attempts to sing a single pitch class taken from [233] 122
4.6 Schematic model of a performer and a listener in a concert hall 148

5.1 The filter model of the relationship between objective and subjective aspects of sound 150
5.2 Example of 2-D representation grid in Sonic Mapper 153
5.3 Multi-dimensional Scaling (MDS) solution derived from dissimilarity matrices obtained using Sonic Mapper 155
5.4 A scree plot of the stress measure and number of perceptual dimensions used in the MDS analysis 156
5.5 Pilot Test user interface of ABX test to determine if listeners could distinguish between *trimmed* and *untrimmed* source material 163
5.6 Example of user interface in Sonicmapper of COMPARISON task 165
5.7 Example of user interface in Sonicmapper of SORTING task - audio fragments are represented by numbered boxes 166
5.8 Scree plot of stress measure with increasing dimensions used to model pilot test II dissimilarity data 168

5.9 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of pilot test II point configuration modelled in 2 dimensions 169
5.10 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of pilot test II point configuration modelled in 3 dimensions 170
5.11 Nonmetric MDS solution to pilot test II similarities, modelled in 2 dimensions 171
5.12 Nonmetric MDS solution to pilot test II data, modelled in 3 dimensions . 172

6.1 Photograph to illustrate position of head-mounted microphone and masking tape used to replicate position between recordings 175
6.2 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232a modelled in 2 dimensions 182
6.3 Nonmetric MDS solution for listening test 232a, modelled in 2 dimensions 182
6.4 Dendrogram of hierarchical clustering of fragments from a 2-dimensional MDS solution for test 232a 183
6.5 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232b modelled in 2 dimensions 184
6.6 Nonmetric MDS solution for listening test 232b, modelled in 2 dimensions: dimension 1 and dimension 2 plotted 185
6.7 Dendrogram of hierarchical clustering of fragments of a two-dimensional MDS solution for test 232b 185
6.8 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232b modelled in 3 dimensions 186
6.9 Nonmetric MDS solution modelled in three dimensions for listening Test 232b 186
6.10 Nonmetric MDS solution for listening test 232b, modelled in 3 dimensions: dimension 1 vs. dimension 3 187
6.11 Dendrogram of hierarchical clustering of fragments from 3-dimensional MDS solution for test 232b 187
6.12 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 2 dimensions 188
6.13 Nonmetric MDS solution for listening test 221, modelled in 2 dimensions 189
6.14 Dendrogram of hierarchical clustering of fragments from 2-dimensional MDS solution for test 221 189

6.15 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 3 dimensions 190
6.16 Nonmetric MDS solution for listening test 221, modelled in 3 dimensions . 191
6.17 Nonmetric 3-dimensional MDS solution for listening Test 221, Dimension 1 vs. Dimension 3 191
6.18 Dendrogram of hierarchical clustering of fragments from the 3-dimensional MDS solution for test 221 192
6.19 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 212 modelled in 2 dimensions 193
6.20 Nonmetric MDS solution for listening test 212, modelled in 2 dimensions 194
6.21 Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 3 dimensions 195
6.22 Nonmetric MDS solution for listening test 212, modelled in 3 dimensions 196
6.23 Nonmetric MDS solution for listening Test 212, Dimension 1 and Dimension 3197
6.24 Dendrogram of hierarchical clustering of fragments from the 3-dimensional MDS solution 198
6.25 Nonmetric MDS solution for listening test 212 Verse D fragments, modelled in 2 dimensions 199
6.26 Nonmetric MDS solution for listening test 212 Verse E fragments, modelled in 2 dimensions 200
6.27 Illustration of HMM states defined in the AMPACT alignment algorithm 203
6.28 Initial alignment in AMPACT of MIDI file timings (coloured bars)overlaid on a spectrogram of the: blue bars depict vowels, green bars depict voiceless consonants and red bars depict silences 203
6.29 Improved alignment after visual inspection and adjustments made in Sonic Visualiser. 204
6.30 MIDI representation of Test 221 phrase (De Domo) with timings taken from the score 204
6.31 MIDI representation of Test 221 phrase (De Domo) with timings as performed by the singer 204
6.32 Location of bars in phrase for Test 232a 205
6.33 Plot of local tempo (beats per minute) for bars 1-4 over 5 fragments in Test 232a 206
6.34 Location of beats in phrase for Test 232b 206

6.35 Plot of local tempo (beats per minute) for beats 1-6 over eight fragments in Test 232b 206
6.36 Location of beats in phrase for Test 221 207
6.37 Plot of local tempo (beats per minute) for beats 1-6 over eleven fragments in Test 221 207
6.38 Location of beats in phrase for Test 212 207
6.39 Plot of local tempo (beats per minute) for beats 1-4 over 34 fragments in Test 212 208
6.40 Unfiltered fundamental frequency trace of values estimated by the YIN algorithm implemented in AMPACT, for note number 8 of MRVirtV1 fragment in Test 232a 212
6.41 Filtered fundamental frequency trace to provide vibrato contour of note number 8 of MRVirtV1 fragment in test 232a 212
6.42 Mean vibrato rate and standard deviation for each fragment in test 232a 213
6.43 Mean vibrato extent and standard deviation for each fragment in test 232a 213
6.44 Median vibrato extent for each fragment in test 232a 214
6.45 Mean vibrato rate and standard deviation for each fragment in test 232b 214
6.46 Mean vibrato extent and standard deviation for each fragment in test 232b 215
6.47 Median vibrato extent for each fragment in test 232b 215
6.48 Mean vibrato rate and standard deviation for each fragment in test 221 . 216
6.49 Mean vibrato extent and standard deviation for each fragment in test 221 216
6.50 Median vibrato extent for each fragment in test 221 217
6.51 Mean vibrato rate and standard deviation for each fragment in test 212 . 217
6.52 Mean vibrato extent and standard deviation for each fragment in test 212 218
6.53 Median vibrato extent for each fragment in test 212 218
6.54 Vector property fitting of performance attributes onto MDS perceptual map for Test 232a 219
6.55 Vector property fitting of performance parameters onto MDS perceptual map for Test 232b 220
6.56 Vector property fitting of performance parameters onto MDS perceptual map for Test 221 222
6.57 Vector property fitting of performance parameters onto MDS perceptual map for Test 221 Dim 1 vs Dim 3 223
6.58 Vector property fitting of performance parameters onto MDS perceptual map for Test 212 dimension 1 and dimension 2 225
6.59 Vector property fitting of performance parameters onto MDS perceptual map for Test 212 Dim 1 vs Dim 3 226

7.1 Schematic model of a performer and a listener in a concert 247

G.4 Test 212 fragment 267

I.1 Scree plot of stress measure with increasing dimensions used to model test 232a dissimilarity data 273
I.2 Scree plot of stress measure with increasing dimensions used to model test 232b dissimilarity data 274
I.3 Scree plot of stress measure with increasing dimensions used to model test 221 dissimilarity data 274
I.4 Scree plot of stress measure with increasing dimensions used to model test 212 dissimilarity data 275
I.5 Scree plot of stress measure with increasing dimensions used to model test 212 Verse B dissimilarity data 275

Dedication

This thesis is dedicated to the memory of my parents, Philip (1935-2000) and Gwen (1942-2000) Brereton, who in their different ways instilled in me a love of music and a passion for scientific understanding. It is a great sadness that they are not here to see this completed.

Acknowledgments

Firstly, heartfelt thanks go to my supervisors Dr Damian Murphy and Prof David Howard; to Damian in particular, who first encouraged me to do this, and who was always willing to give advice, support, motivation and gin & tonic when it was needed. Thanks also to my thesis advisor Mr Tony Tew for initial advice and helping keep my attention to detail.

Thanks to those who participated as singers and/or listeners: to the quartet of singers who gave of their time to perform in the real performance space and survived the anechoic chamber; Singers who performed in the VSS - who for professional reasons cannot be named here - but you know who you are. I hope you enjoyed the cakes. To all who participated in extensive listening tests - for your patience, enthusiasm and input - this project would not have had any results without you all!

To Delma Tomlin, Gill Baldwin and the staff at the NCEM for allowing me to commandeer your beautiful venue for two days in order to blast it with sine wave sweeps, and then return to record solo singers in the different acoustics, in between whiles balancing on a chair with a long stick to open and close the panels.

To Pip and Jackie who lent me their dining room for a number of days, provided food and drink, whilst I battled with my results chapter. Many thanks indeed.

For those who have supported, helped and inspired me, including past and present members of the AudioLab, both staff and students. Particular thanks go to: Andrew Chadwick for too many things to list, but including building the virtual singing studio single-handedly and putting up (with) my magic curtain; Matt Speed for friendship, advice, surfing and LaTex templates; Simon Shelley for excellent explanations of all things acoustic and diagrams on paper plates; Iain Laird for fun debates about the best way to implement a virtual performance space and for the best acoustic baffle I've seen (duvet); Helena Daffern for always being there when I needed her, love, friendship and chocolate, gin and tonic and much nonsense at the gym. Amelia, Becky and Laurence, who may be relative newbies but who have already made their mark with cakes, paper-planes, crosswords and motivational post-it notes - I could not have completed without them; Aglaia Foteinou for her dedication to impulse response measurements, even when it meant long hours at the venue, calling out for pizza and extra hands to hoik the loudspeakers

around; Gavin Kearney for initial inspiration on the design of the VSS, and for having the cutest car in the world; Steve Oxnard for ridiculousness, soothing guitar music when I needed it most and for the best life-motto ever; Andy Hunt for advice on thesis writing, time management and time juggling - my google calendar isn't quite at your level but I'm working on it....; Alastair Moore for long chats about perception, localization, and how to have a life and do research at the same time; Mr Grammar-man for appreciating my trying to get all the apostrophes in the right places, and for knowing the difference between —and –, and for caring about it.

Last, and definitely by no means least, my lovely husband John and gorgeous children, Arthur and Esme, who have put up with me spending hours at the computer, droning on about singing and acoustics and generally being grumpy whilst trying to get this finished. I could not have completed this without your unswerving support (emotional, financial and practical) and love. Thank you.

Declaration

I hereby declare that this thesis is entirely my own work and all contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced. I also declare that some parts of this program of research have been presented previously at conferences. These conference papers are listed here:

- **Investigating Singing Performance in Different Acoustic Environments using the Virtual Singing Studio,** J.S. Brereton, D.T. Murphy & D.M. Howard, *Acoustic Challenges in Quires and places where they sing*, Institute of Acoustics, 26 June 2013, London, UK, 2013
- **Singing in Space(s): Using the Virtual Singing Studio to investigate singing performance,** J.S. Brereton, D.T. Murphy & D.M. Howard, **Sounds in Space: Research Symposium**, University of Derby, 1st July 2013, Derby, UK, 2013
- **The Virtual Singing Studio: A loudspeaker-based room acoustics simulation for real-time musical performance,** J.S. Brereton, D.T. Murphy & D.M. Howard, *Joint Baltic-Nordic Acoustics Meeting (BNAM)* Odense, Denmark 18-20 June 2012
- **A loudspeaker-based room acoustics simulation for real-time musical performance** J.S. Brereton, D.T. Murphy & D.M. Howard, *25th AES UK Conference: Spatial Audio in Todays 3D World in association with the 4th International Symposium on Ambisonics and Spherical Acoustics*, 25-27 March 2012, York, UK, 2012
- **The impact of vibrato usage on the perception of pitch in early music compared to grand opera** H. Daffern, J.S. Brereton & D.M. Howard, *Acoustics 2012*, Nantes, France, 2012
- **Evaluating the Auralization of Performance Spaces and Its Effect on Singing Performance,** J.S. Brereton, D.T. Murphy & D.M. Howard, *130th Audio Engineering Society Convention*, London, May 2011

Chapter 1

Introduction

It is widely understood that humans adjust their vocal behaviour according to the acoustic environment in which they find themselves. Imagine, for example, the hushed whispers of visitors to a large cathedral, or the way a teacher might alter their vocal output in order to project their voice to the back of a classroom.

It has been shown that musicians adjust their performance according to the acoustic parameters of the performance space. [1, 2]. Musicians respond to the aural feedback provided by the acoustical conditions of the environment and alter their performance in accordance with their perception of how their own sound is being affected by the acoustics of the space [3]. Indeed these changes in vocal performance have been recognised and noted for some time:

> *one sings in one way in churches and public chapels and another way in private rooms. In [church] one sings in a full voice - and in private rooms one sings with a lower and gentler voice, without any shouting.* Zarlino, 1558 [4]

During vocal performance, singers and actors make great use of the aural feedback provided by the surrounding space, and will alter their performance, both consciously and subconsciously, according to how it is being affected by the acoustics of the venue. Such changes can stem either from automatic, reflexive adjustments made by the musician to adapt to his/her surroundings, or from conscious learned changes in technique. The latter may be influenced by what the musician imagines the audience will perceive during the performance.

However, these alterations, such as changes in timing, vibrato and intonation, as well as in vocal function and the resulting spectral balance of the vocal sound have not been rigorously observed, and they are not yet well documented nor systematically investigated. Furthermore, the majority of analyses of the singing voice rely on recordings of singers

made in "laboratory conditions" i.e. in an acoustically non-reflective (anechoic) room, which ultimately removes this important aspect of performance.

1.1 Investigating Musical Performance

In order to undertake any such investigation of singing performance and how it changes in different room acoustics the researcher must be able to alter the auditory environment effectively, so as to undertake laboratory experiments on singing in more natural conditions.

An interactive room acoustics simulation should work in near enough to real-time to allow the singer to perform "in the lab" but hear themselves as if in a real performance venue. A small number of previous authors have undertaken similar work, but as yet the "plausibility" and "realism" of these virtual performance spaces has not been fully verified.

1.1.1 Traditional Auralisation Methods

Traditionally, an auralisation of a soundfield offers a static aural simulation or reproduction of a pre-recorded source in a space. The sound-source itself is usually recorded in an anechoic environment, in order to obtain a "clean" recording which does not contain any sound reflections or reverberation stemming from the natural acoustics of the recording space.

However, anechoic recordings, particularly of singers, are highly unrealistic in terms of musical performance characteristics, which are altered substantially due to adjustments made to compensate for this unusual acoustic environment. Nevertheless auralisation techniques and virtual acoustics applications generally rely on such recordings of source material.

1.1.2 Interactive Room Acoustics Simulation

One of the key aspects therefore that is missing from traditional auralisation or room acoustics simulations is that the source (instrumental, speaker or singer) reacts to the aural feedback received from the space itself.

Although not critical when recreating the experience from a passive perspective of an audience member, this missing element becomes much more important when the listener is also the performer: an active sound source responding to what they hear as the sound they make is projected into the acoustic environment in which they have been virtually placed.

In order to investigate the effect of diverse room acoustic characteristics on the musical performance of singers, one first needs to implement a realistic interactive room acoustics

simulation, optimised for singing performance. It must then be determined whether the simulation provides "sufficient realism" [2] to engender singing performance adjustments relating to changing room acoustic parameters, in the same manner as that of a real performance space.

The main aim of this research project is to show that a vocally interactive acoustic simulation of a performance space can be implemented, which singers accept as being plausible and in which singing performances are sufficiently similar to those in the real performance space. Once this has been shown to work effectively, the "Virtual Singing Studio" will play its own role in alleviating the problems identified with anechoic source material creation as described above.

1.2 Hypothesis

A plausible interactive room acoustic simulation will elicit changes in singing performance which replicate those occurring in different real acoustic environments.

This hypothesis will be tested by:

1. Rendering a virtual simulation of a performance space which allows a singer/speaker to hear their vocal performance in real-time as if in the real performance space - Section 3.4

2. Comparing objective room acoustic measurements of the real space with the virtual simulation - Section 3.4

3. Recording vocal performances in the virtual and real performance spaces - Section 6.2

4. Collecting subjective responses from singers about their own performances in the real and virtual space - Section 3.6

5. Analysing and comparing vocal performance parameters of singing in the real and virtual space - Section 6.4

6. Asking listeners to judge the similarities between vocal performances recorded in the real and virtual space - Section 6.3

1.3 Novelty of research

This research project has two novel aspects. First, the perceptual evaluation and comparison of singing performances in real and virtual performance spaces has not yet been undertaken by others who have used virtual auditory environments in research on musical performance. This project seeks specifically to determine the similarity of singing performances in real and virtual performance spaces.

Although Multi-dimensional Scaling (MDS) has been used recently in the classification of environmental sounds, and other sensory evaluation techniques such as Principal Component Analysis (PCA) in the evaluation of real and virtual auditory environment in concert halls, this research is the first to use MDS to gain an understanding of the perceived similarities and differences between singing performances. In this thesis, MDS analysis is used for dimensionality reduction of the multivariate perceptual data, and complemented by "property fitting" techniques which are used to identify the perceptual dimensions of the similarity ratings. It is the application of these techniques to the question of musical performance similarity which is also novel.

1.4 Main contributions

Virtual Singing Studio

There has been a growing interest recently in providing real-time room acoustic simulations. Favrot [5, 6] has implemented a loudspeaker based room acoustics simulation to facilitate research into auditory perception. Others have included source-sound interactivity, so that the performer can hear his/herself as if in a real performance space. For example, Uneo et al. [7, 8] have simulated a 6-channel real-time acoustic environment for instrumentalists [9], whilst Schärer Kalkandjiev and Wienzierl [10, 11] have simulated stage acoustics for a solo cellist. Woszczyk et al. [12, 13] have provided virtual stage acoustics to enhance the real stage acoustics found in concert hall venues to support instrumentalists in performance, or to recreate the room acoustic conditions of historical venues.

The research outlined here is the only study to implement a sound-source interactive room acoustics simulation over a three-dimensional array of loudspeakers specifically optimised for the singing voice.

Evaluation of the Room Acoustics Simulation

Although others have simulated room acoustics for the performing musician in real-time the present study seeks to assess the plausibility and performance of the simulation not only by comparing objective measures of room acoustic parameters, but also through analysing

and comparing singing performances made in the simulation and the real performance venue on which the simulation is based. In this way a fuller assessment can be made of the effectiveness of the simulation, not only in terms of verification, meeting the design specifications of simulating the real performance space, but also in terms of validation, meeting the needs of the user (singer).

Singing in Spaces

Investigations by Ueno, Kato et al. [14, 15] into musicians' adjustment of musical performance according to room acoustic conditions include a number of different instrumentalists and one singer, but they found that the participants altered their musical performance in different ways, and so could not generalise across the musicians in the study.

The present study concentrates solely on singing voice looking in particular at those performance attributes which singers and listeners suggest are altered according to acoustic environment, including measures of vibrato, tempo and intonation accuracy and precision. Due to the fact that a number of different room acoustics configurations could be simulated and similarly presented in the real-performance venue, alterations between different room acoustic conditions could also be investigated as well as the assessment between the real and simulated venue.

Quantitative and Qualitative Data

The present study attempts to evaluate the plausibility of the implemented room acoustic simulation by combining qualitative data obtained through interviews and questionnaires, with quantitative data on the accuracy of the room acoustic simulation assessed through objective room acoustic parameter analyses.

Similarly, the perceptual evaluation of the singing performances recorded in the real and virtual spaces combines quantitative assessments of perceived similarity with listeners' subjective comments on the singing. In addition it combines perceptual mapping with the analysis of objective singing performance attributes.

Dimensionality Reduction

Gygi [16] used MDS techniques to assess perceived similarity/dissimilarity between environmental sounds, and Bonebright [17] used MDS to investigate methods of sonification for auditory display of weather data. Lokki et al. [18, 19] have used other dimensionality-reduction techniques on subjective sensory data to assess listeners' preferences for concert hall acoustics.

This study is unique in using well-established multi-dimensional scaling techniques to bear on the perception of similarity between musical phrases, specifically sung phrases.

1.5 Structure of Thesis

The thesis is organised as follows:

Chapter 2 - Simulating Room Acoustics introduces room acoustics, the behaviour of sound in rooms and Section 2.3 describes how room acoustics are captured and analysed via the measurement of Room Impulse Responses (RIR) . It outlines the aspects of room acoustic conditions which have been investigated in studies of concert hall acoustics. It then goes on to look at stage acoustic parameters which are important for the performing musician. Section 2.4 (Section 2.6) goes on to describe techniques for providing virtual room acoustics and how they have recently been used to simulate the room acoustic conditions of musical performance spaces.

Chapter 3 - Virtual Singing Studio: Implementation and Verification describes the steps taken in the design and implementation of the Virtual Singing Studio, which is designed to provide an interactive room acoustics simulation for the musician. It first outlines a pilot project carried out to build and test a prototype VSS (Section 3.2). After a description of the performance venue which is to be simulated (Section 3.3) the final improved design for the VSS is described in Section 3.4. A number of objective measurements are made in the real and simulated space, and compared in order to verify the implementation of the VSS (Section 3.5). Finally a number of professional singers are asked to perform in the simulated and real performance spaces and are asked to describe their subjective impressions of the VSS (Section 3.6).

Chapter 4 - Singing in Space(s) looks at the ways that musical performance in general, and singing in particular, is understood to change according to the acoustic environment in which it takes place. The chapter begins with a brief history of the quantitative analysis of music performance (Section 4.2). Section 4.3 sets out the variety of objective parameters which can be extracted from music performance for analysis and comparison. Results of recent research into the changes in singing performance according to room acoustic conditions are explored in Section 4.6. Section 4.7 describes a case study of a vocal quartet which was undertaken to begin to investigate the room acoustic conditions of the real performance venue.

Chapter 5 - Singing Performance Analysis and Evaluation describes some of the many available methods of undertaking perceptual evaluations of audio material, and examines the possibility of extending these methods to the subjective evaluation of musical performances. Techniques which allow the researcher to correlate objective analysis with the results of perceptual evaluation are outlined in Section 5.3. Sections 5.4 and 5.5 present the results of two pilot listening tests which are undertaken to inform the design of the main listening test described in Chapter 6.

Chapter 6 - Singing in Real and Virtual Acoustic Environments presents the results of the main listening test carried out in this thesis. First of all it summarises the singers' own assessment of how their singing changed according to the room acoustic conditions in the virtual and real spaces (Section 6.2). The results of the perceptual listening test and subsequent dimensionality reduction analyses are presented in Section 6.3. Section 6.4 describes the objective acoustic analysis of singing performances recorded in the real and virtual spaces. These objective attributes are used to inform the interpretation of the results of the perceptual evaluation of the recorded singing performances as described in Section 6.5. In addition analyses of variance (ANOVA) are performed to test for the effects of room acoustic conditions and simulation (real or virtual) on the measured performance attributes and listeners' similarity judgements. (Section 6.6)

Chapter 7 - Conclusion brings the results of the objective and subjective evaluations of the VSS together and examines the evidence to support the main hypothesis. Some future work which could be undertaken to improve the VSS is suggested in Section 7.1 together with ideas for further work in the analysis of singing performance. Section 7.2 summarises some of the possible areas of application for the research in this thesis and the Chapter concludes with some final remarks on the implementation and evaluation of the VSS.

Chapter 2

Simulating Room Acoustics

2.1 Introduction

This chapter opens with a brief introduction to sound as a wave (Section 2.3), and the behaviour of sound inside rooms (Section 2.3.1).

Section 2.3.2 outlines some of the methods available for capturing and recording the acoustic properties of rooms and Section 2.3.3 goes on to define some of the objective room acoustic parameters which can be measured and evaluated. Section 2.4 discusses the influence of room acoustic characteristics on the experience of music-making for listeners and performers. Measurements of room acoustic properties which are of specific importance to the performing musician are described in Section 2.4.2 and performers' perceptual evaluation of concert hall stages are outlined in 2.4.3.

Methods of auralisation, that is making audible the data captured in room acoustic measurements so that one can "hear the room" are laid out in Section 2.5. Section 2.6 discusses the implementation of virtual room acoustic simulations which operate in near to real-time in order to allow a performing musician to play and hear him/herself as if in a concert hall or other venue. The subjective and objective evaluation of such systems are considered in Section 2.6.4.

2.2 Acoustics

Sound is a variation in pressure which is sensed by the human auditory system and perceived as sound. The human auditory system is sensitive to variations in pressure which occur within the frequency range of 20 Hz to 20,000 Hz. Sound propagates through different media - gas, liquid or solid - as a wave.

Sound waves in the free field, which do not interact with objects or surfaces, propagate outwards in three dimensions, and no acoustic energy is lost as the original sound-producing

disturbance spreads. However, sound intensity decreases as the acoustic energy is in effect spread over a sphere. Sound intensity, which is measured in Watts/m^2, is a measure of the amount of energy which passes through a unit area in a one second, and is inversely proportional to the distance from the source. This relationship is captured by the Inverse Square Law

$$I \propto \frac{W_{source}}{4\pi r^2} \tag{2.1}$$

where I is sound intensity (in W/m^2), W_{source} is the power of the source (in Watts), and r is the distance from the sound source (in m).

After calculating various distances using Equation 2.1 it can be seen that a doubling of distance from the source to receiver results in a 6dB drop of sound intensity level (except in the near-field, which is approximated by a radius equal to the size of the source itself).

Sound Intensity Level (SIL) is expressed as the ratio of the sound in question to a standard reference level, and expressed in decibels (dB).

$$SIL = 10log_{10}\frac{W_1}{W_2} \tag{2.2}$$

where *SIL* is sound intensity level, W$_1$ is the actual sound intensity and W$_2$ is the reference level 1 picowatt/m^2 (10^12 W/m^2).

Whilst sound intensity measures the amplitude of a sound wave at a particular point it is not the most perceptually relevant measurement of the amplitude of a sound. The human auditory system is sensitive to pressure, and so measuring the pressure level is a more relevant way of describing the amplitude of sound waves.

Sound intensity is proportional to the square of the Sound Pressure Level (SPL), which is also measured relative to a reference source, in this case 20 micro Pascals (μ), and expressed in decibels.

$$SPL = 20log_{10}\frac{P_1}{P_2} \tag{2.3}$$

where *SPL* is sound pressure level, P_1 is the pressure of the source in question, and P_2 is the reference source of 20 micro Pascals (μ).

The difference in Sound Pressure Level (SPL) between two sound sources can also be expressed in decibels using equation 2.3, where P_1 and P_2 are the sound pressure levels of the two sound sources respectively. A doubling in SPL is equivalent to an increase of 6 dB, and a halving in SPL is equivalent to a decrease of 6 dB.

A thorough introduction to physics of sound and its perception can be found in [20] and [21].

2.3 Room Acoustics

Many sounds, including musical sounds, are heard inside rooms and hence the behaviour of sound waves as they interact with objects and boundaries must be considered, as it is this behaviour which characterises sound in rooms.

2.3.1 Sound propagation in rooms

Within a room, sound waves are transmitted through the air from sound source to listener, but the surrounding walls, ceiling, floor and other reflective surfaces within the room all play a part in transforming the sound that eventually reaches the listener's ears.

Direct Sound and Early Reflections

Direct sound travels in a straight line from the sound source to the listener and is the first element to arrive at the listener's ears, as illustrated in Figure 2.1.

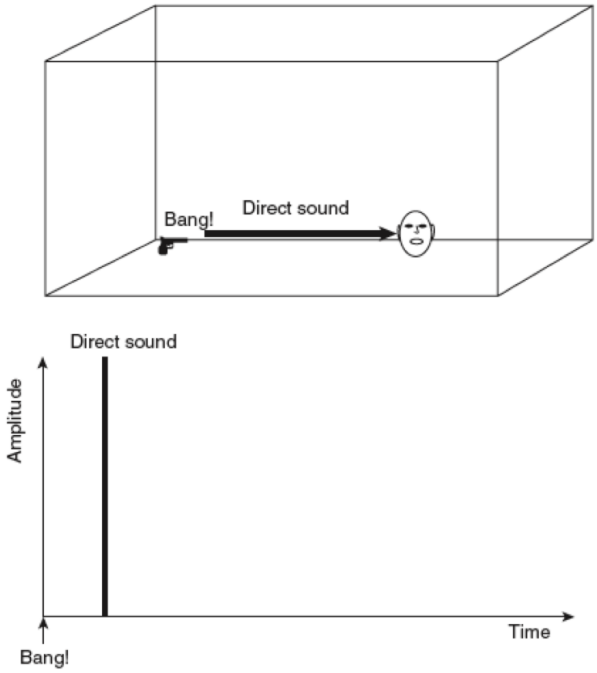


Figure 2.1: *Illustration of the direct sound in a room, from [20, p.262] used with permission*

A number of early reflections follow soon after, which have reflected off two, three or more surfaces before arriving at the listener position. (Figure 2.2). In most concert hall environments the first early reflection arrives at the listener very shortly after the direct sound. This first order reflection is lower in amplitude than the direct sound, as it has travelled from the source and reflected off one surface, before arriving at the listener. The next successive early reflections decrease in sound intensity at a rate of 6 dB for

each doubling of distance travelled, however their timing and frequency content will differ according to the material, size and positioning of the walls and ceiling in the space and their absorptive or reflective qualities [22, 23].

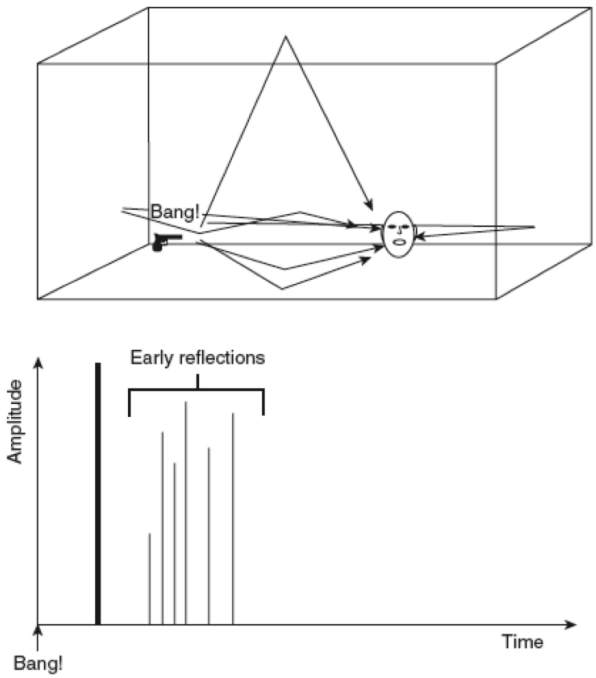


Figure 2.2: *Illustration of the early reflections in a room, from [20, p.262] used with permission*

Within a large room early reflections typically arrive at the listener within about 50 - 80 milliseconds of the direct sound. However, early reflections which arrive within this time period are not perceived as separate events, but rather reinforce and/or colour the direct sound [24].

In general, in medium to large concert halls the first reflections to reach the audience position are from the nearest side walls (lateral reflections), or from the ceiling. However, the exact timing and level of early reflections for the listener will of course depend on the audience member's position within the audience areas. Similarly, on the stage of the concert hall, the timing and level of early and late reflections are important for the performing musician in order that they can hear their own sound, as well of that of other musicians in the ensemble, and furthermore gain an impression of the size and shape of the auditorium.

Reverberation

After a certain time reflections arrive at the listener which are much lower in amplitude and temporally closely spaced; these late reflections are perceived as reverberant sound (reverberation) as illustrated in Figure 2.3.

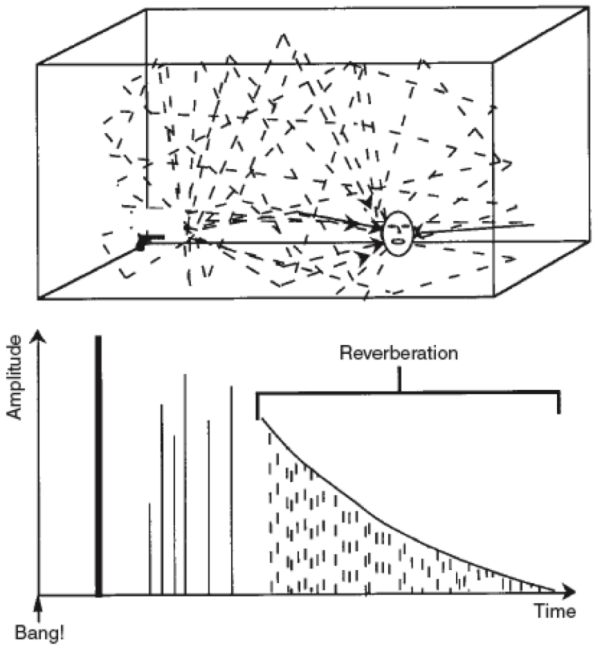


Figure 2.3: *Illustration of reverberant sound in a room, from [20, p.262] used with permission*

The reverberant sound level in a space is reached when the rate of energy supplied by a continuous sound source is equal to the rate at which the sound is being absorbed by the room [24].

Source localisation

In general terms, a listener will locate the position of a sound source relative to their own position according to the direction at which the direct sound arrives at the ears. In addition, the timing and the balance of direct sound and early reflections will help a listener to judge their distance from the source. The Inter-aural Time Difference (ITD), the difference in arrival time of the direct sound at each ear plays the biggest role in source localisation at lower frequencies. In contrast the Inter-aural Level Delay (ILD), the difference in level between the two ears, is more useful at higher frequencies, due to the shadowing effect of the head.

When close to the sound source the direct sound dominates whilst early reflections are delayed and much weaker relative to the direct sound. When the listener is positioned further away from the sound source, early reflections arrive much sooner after the direct sound, and are only slightly lower in amplitude, since they have not travelled so much further than the direct sound itself [24].

Additionally, impurities in air do serve to absorb some sound energy, most importantly at higher frequencies, so a sound will not only decrease in amplitude level as the listener

moves away from the source, but will also become "duller" (less energy in the higher frequency regions of the spectrum)

2.3.2 Measuring Room Impulse Responses

A Room Impulse Response (RIR) can be thought of as the "acoustic signature" of a room i.e. the response of a room to an excitation signal for a given source and receiver combination. RIRs can be either measured in situ or simulated through computer modelling.

Measuring a RIR involves playing an audio excitation signal in the room and recording the response. The excitation signal used must have sufficient energy across the spectrum to ensure a good signal-to-noise ratio at all frequencies. Two main methods exist : through direct measurements with impulsive excitation signals or through indirect measurements via wideband signals, such as noise, or narrowband signals, such as sine sweeps or time-stretched pulses.

Impulsive excitation

Traditionally loud impulsive excitations such as gun-shots, canon or electrical sparks were used to record room impulse responses. No post-processing is needed in this method, as the recorded gun-shot in the space represents the room impulse response directly in the time domain. These methods have fallen out of favour, mainly because the frequency response of such excitation signals are not perfectly flat across the frequency spectrum and it is difficult to guarantee a sufficient energy level of the signal.

Noise excitation

Noise signals can be used for the purpose of RIR measurement utilising random or pseudo-random broadband noise output from a loudspeaker. The decay curve of the room response can be measured once the noise source has been switched off. Time-based room acoustic parameters can be obtained this way, but energy-based parameters such as Clarity (C_{80}) cannot be evaluated since this method does not produce an impulse response to allow for the calculation of energy content across the spectrum.

Sine sweep excitation

The exponential swept sine (ESS) technique allows the measurement of the impulse response of a system, and at the same time, any distortion present in the system can be removed (see Section 2.3.2). It has been developed by Farina [25] and has been shown to be robust for many applications.

A logarithmic sine sweep is synthesized with constant amplitude, increasing exponentially in frequency per time unit. The resulting sweep moves slowly through low frequencies but travels faster through the higher frequencies, meaning that the spectrum is attenuated by 3dB per octave.

The sine sweep is output to the room, recorded and then deconvolved with an time-inversed filter of the original input sine sweep. The time inverse filter must account for the amplitude envelope of the input sweep, and hence an amplitude envelope must be applied in order to eventually obtain an impulse response with a flat spectrum. This results in a linear impulse response with an initial delay equal to the length of the input signal [26, 25].

Any harmonic distortion present in the loudspeaker and recording equipment is now seen in the time domain, and appears as a series of lower amplitude impulse responses prior to the main RIR which can be easily excluded by editing out the initial time delay. The matlab code used to generate the inverse filter and deconvolution of recorded sweeps used in this thesis can be found on the data CDs - see Appendix A.

The exponential swept sine (ESS) technique has been used for many in-situ room impulse response measurements in concert halls and other performance spaces (e.g. [27, 28, 25, 10]

The many different techniques for the measurement of Room Impulse Responses are not addressed in depth here, instead the reader is referred to [26] in which Fausti and Farina have authored a thorough comparison of RIR measurement methods. Methods for synthesizing RIRs from computer models are outlined briefly in Section 2.5.3. A number of objective room acoustic parameters can be calculated once the impulse response of a room has been obtained, either through measurement or computer modelling.

The term “parameter” is used throughout this chapter (and thesis) in accordance with ISO-3382 [29], although others prefer terms such as ”attribute” or ”measure”. However, there is no intention to imply that room acoustic ”parameters” are able to be independently manipulated.

2.3.3 Objective room acoustic parameters

The most well-known and widely quoted room acoustic parameter is Reverberation Time (RT60), but of course this one measure alone cannot fully describe the complex room acoustic conditions of performance spaces such as concert halls or churches. RT60 and a small selection of some of the other more common room acoustic parameters are explained briefly in this section, with particular attention paid to those which are of relevance to the musician during performance. The equations for the objective parameters summarised in this section are taken from [29] unless otherwise stated.

Reverberation Time (RT60)

Reverberation Time (RT60) is traditionally defined as the time it takes for the sound level in a room to decrease by 60 dB. It was Wallace Sabine (1868−1919) [30] who first formulated the relationship between reverberation time, room volume and the total area of sound absorbing surfaces within the room and their absorption properties.

Through exhaustive measurement and mathematical modelling, he found that reverberation time for a theoretically ideal room, where all room surfaces are similarly absorbent, is proportional to the ratio of volume to surface area and can be estimated as:

$$RT = 0.161V/A \tag{2.4}$$

where V is the room volume and A is a measure of the total absorption of all the surfaces of the room.

A is calculated by multiplying each surface area by the correct absorption coefficients for the surface material in octave frequency bands. An absorption coefficient of 1 is equivalent to an "open window", and so a low absorption coefficient corresponds to a more reflective material.

$$A = S_1a_1 + S_2a_2 + .. + S_na_n = \sum S_ia_i \tag{2.5}$$

where A is the absorption of the room, S_n is the area of the surface (m^2) and a_n is the absorption coefficient of the surface.

Sound is also absorbed by the air inside the room especially at high frequencies and needs to be taken into account in large auditoria. To take account of air absorption equation 2.4 can be rewritten as

$$RT = 0.161V/(A + mV), \tag{2.6}$$

where m is the air absorption coefficient, which is dependent on air temperature, humidity and frequency.

Reverberation Time can be obtained directly from the RIR by measuring the rate of decay after estimating the slope of the backward integration of the squared impulse response [29] between -5dB and -30dB, which gives T30. Similary T20 is measured from -5dB to -25dB. If the decay is linear then all measures of reverberation time, RT60 (T60), T20 and T30 will be equal [25, 31].

Reverberation time varies according to frequency, meaning that RT60 values should be calculated for frequency bands in order to characterise the room acoustics of a performance space properly.

Early Decay Time (EDT)

EDT is evaluated in the same way as reverberation time measurements outlined above, but is measured from 0 dB to -10dB and then extrapolated to a drop in level of 60dB. In this way it is another equivalent measure of Reverberation Time (RT60) and when assessed together with T30 it can form an impression of the shape of decay within the room. For example a shorter EDT than T30 would show that the initial rate of decay was quicker; a rapid initial decay can be perceived as an overall shorter reverberation time [32].

The initial rate of decay of reverberant sound appears to be more perceptually important than the total reverberation time. A rapid initial decay is interpreted by the human ear as meaning that the reverberation time is short [33]. Section 2.4.1 gives more on the perceptual correlates of object room acoustic parameters.

Strength

Strength (G) is the level of the total sound in the room, relative to the free-field direct sound energy at a distance of 10m [29].

$$G = 10log\left\{\frac{\int_0^\infty p^2(t)dt}{\int_0^\infty p_{10}^2(t)dt}\right\}, dB \tag{2.7}$$

where $p^2(t)$ is the sound pressure of the impulse response at the measurement point, and $p_{10}^2(t)$ is the sound pressure measured at a distance of 10m in the free field.

Clarity

Clarity measures, C_{50} and C_{80} are calculated to quantify the ratio (in dB) between energy arriving in the first 50 (or 80) milliseconds of the sound, and the energy arriving later.

$$C_{50} = 10log\left\{\frac{E_{0-50ms}}{E_{50-\infty}}\right\}, dB \tag{2.8}$$

$$C_{80} = 10log\left\{\frac{E_{0-80ms}}{E_{80-\infty}}\right\}, dB \tag{2.9}$$

where E_{0-50ms} is the early energy in the first 50 ms of the impulse response and $E_{50-\infty}$ is the late energy after 50 ms. For C_{80} the upper integration limit in the calculation of early energy is 80 ms rather than 50 ms.

Brilliance (BR)

Defined as the ratio of energy decay in high frequencies to mid-frequencies, that is the ratio of EDT_{2000} to EDT_{mid}, where EDT_{mid} is the average of EDT values at 500 and 1000 Hz.

Initial Time Delay Gap (ITDG)

Initially suggested by Marshall [34] as the delay between the direct sound and the first reflection in the stalls and correlated with a sense of intimacy, (see Section 2.4.1). ITDG is also a main perceptual cue for estimating a listener's distance from a sound source.

Other objective parameters

Less frequently used parameters are not outlined here, save to mention that authors often introduce new objective parameters when investigating room acoustics, which are in the main reformulations of those outlined above. For example, in evaluation of acoustic conditions of concert halls Bradley, [35] introduced "early-arriving relative sound level" G_{80} and "late-arriving relative sound level" G_L to describe listeners' subjective impression of spaciousness in the concert hall.

2.4 Room Acoustics and Musical Performance

A number of authors (e.g. [36, 24, 37, 38, 26, 32, 39, 40]) have studied the acoustics of concert halls in order to identify which room acoustic characteristics are judged by listeners to be important to the experience of listening to music in a live concert hall setting. Knowledge of room acoustic parameters and their subjective counterparts are used to improve the acoustic conditions of existing concert halls and to inform the design of performance spaces in the future.

2.4.1 Perceptual evaluation of concert hall acoustics

In early studies of concert hall acoustics perceptual evaluations had to be made in situ in the concert halls in question. Although evaluating concert halls this way provides the most realistic and reliable listening experience, there are a number of drawbacks to this method, including of course the time and cost involved in visiting the real venues, but also the difficulty of comparison between halls [41]. Laboratory based methods of presenting concert hall acoustics using auralisation techniques have been developed and the results of these more recent studies are outlined in Section 2.6.4.

A number of recent studies have sought to correlate subjective impressions of concert hall acoustics with objective room acoustic parameters such as [37, 38, 32]. It should be noted that some of the perceptual parameters outlined here have names similar to the objective parameters summarised above. However, not all relationships between qualitative and quantitative measures are clear cut. For example, a listener's impression of *Reverberance* can be correlated not only to reverberation time but also to the tonal quality (bass ratio and treble ratio) of the sound in the hall.[42].

In order to distinguish concepts and indices with similar names, from this point in the text subjective room acoustic parameters will be italicised.

For example, Cerda et al. [33] identified four perceptual categories which they found correlated to the objective attributes of sixteen concert halls which are outlined below:

- *Spatiality*: measures of late lateral sound level and Inter Aural Cross-correlation (IACC)
- *Clarity-Balance*: C_{80} (Clarity), Speech Clarity (C_{50}) , Brilliance (Br), Bass Ratio (BR)
- *Envelopment*: Late Strength (G_{late})
- *Reverberation*: EDT_{mid} (EDT) in mid-frequency range

Reverberance

Longer reverberation times RT60 can produce a sensation of "fullness of tone" and are also related to a perception of warmth [43].

Early Decay Time (EDT) is very important to the perception of reverberation since only the early part of reverberant decay is audible in continuous music or speech, whereas late reverberant sound is only audible when there is a gap in speech or a period of silence in music.[37]. *Reverberance* has also been found to correlate strongly with EDT and T30 [32].

Loudness

The perceived loudness of the sound in a concert hall or other room is described by the Sound Strength parameter [31](see Section 2.3.3).

Clarity

In subjective responses from audience members in Venetian churches Howard et al. found that clarity was highly correlated with C_{80} [32]. Higher values for *Clarity* correlate with a perceived sense of "definition", low values can add to the perceived "fullness of tone", whereas very low values can cause the impression of "muddiness".

Warmth

The impression of *Warmth* can be related to long reverberation times, especially long bass reverberation time [24]. It has also been shown to relate to strong low frequency levels and bass ratio of reverberance [42].

Intimacy

Intimacy describes the subjective impression of the size of the concert hall, and how close the listener feels to the sound source. A feeling of intimacy in the concert hall can be affected by the listener's proximity to the performers. Beranek argued [43] that ITDG, (see Section 2.3.3 above), the time between the direct sound and the first reflection, can be used to quantify a feeling of "acoustical intimacy" .

In his 1962 survey of 100 concert halls [43] Beranek found that the most well-liked concert halls had Initial Time Delay Gap (ITDG) values of less than 20 ms and were judged as more intimate. However, since the calculation of ITDG is dependent on the location of the position in the auditorium relative to the stage, this link between intimacy and short ITDG has been disputed. For instance, for seats further from the stage, the delay between the direct sound and the first early reflection is shorter, and therefore should be judged as more intimate, but subjective surveys of concert halls suggest that the opposite effect is experienced.

Marshall [34] found also that the direction and strength of early reflections arriving from the sides played an important role in the sense of concert hall intimacy.

Apparent source width

Early reflections which arrive from the side add to the listener's sense of the width of the source for example an individual instrument or the orchestra on stage. It can be described as the "auditory width of the sound field created by a source as perceived at a particular listener position" [44]. Marshall suggests that Apparent Source Width is influenced by the level and time of arrival of early reflections [1].

Some room acoustic parameters seem to correlate well with the objective parameters described above whereas for others the relationships are not simple, and inter-relationships

occur between a number of objective and subjective characteristics.

Optiumum concert hall acoustics

Concert hall studies (for example [43, 2, 45, 34, 46, 24]) have suggested that optimum room acoustic conditions for music performance have to strike a good balance between clarity, sound intensity and liveness, all three of which depend on reverberation time and the reverberant level of the sound. Sometimes this can be achieved through effective design of the stage and auditorium, for example, by balancing a short EDT with a longer RT60 to provide *clarity* and "liveness" to the music.

Sometimes, the optimum room acoustic conditions for a performance venue must also depend on how the space is used, whether for music, presentations, orchestral or choral music amongst other activities. In order to try and accommodate multiple uses more effectively some performance spaces are designed with adjustable acoustic configurations (see Section 2.4.5), usually through the use of absorbing panels and/or drapes, such as the National Centre for Early Music (York) which is detailed further in Section 3.3.

2.4.2 Measuring Stage Acoustic Parameters

Although research into room acoustic conditions in concert halls has a long tradition, it is only really since the late 1970s that the room acoustic preferences of performing musicians have been investigated (for example [34, 2, 47, 24, 48, 45, 49]).

In 1989 Gade [2] produced a seminal study on room acoustic parameters from the performer's perspective. Gade [2] investigated the subjective room acoustic aspects which contributed to orchestral musicians' preferences for concert hall acoustics and tried to relate them to objective room acoustic parameters which can be measured from the Room Impulse Response. He also attempted to understand how concert hall design influenced objective parameters, and in turn the musicians' preferences.

Acoustic parameters of concert halls stages (podiums) can be calculated in a similar way to parameters describing the room acoustics of the concert hall from the audience area. However, the position of the source used in the RIR measurement needs to reflect the performing position(s) on stage, and thus for his measurements Gade placed the microphone receiver at a distance of 1m from the source in order to replicate the topology of player and instrument [49].

With regards to podium acoustics it is generally more difficult to separate objective parameters from their subjective correlations, especially since many of the studies in this area arose from a desire to quantify musicians' preferences for particular concert halls and concert hall stages. For this reason the next section is not split along subjective/objective

division, but rather each objective parameter is explained together with a note on subjective preferences of musicians where these have been investigated.

2.4.3 Perceptual Stage Acoustic Parameters

The perceptual stage acoustic parameters outlined in this section are those which authors have found most important for the musician in performance, and mostly relate to acoustic characteristics found on the stages or platforms of concert halls.

The area of stage acoustics is relatively new, and therefore the number of parameters defined and evaluated is small but some authors have recently investigated the relationships between objective acoustic parameters and subjective impressions for musicians on stage (podium), for example [50, 45, 10].

In a very recent study of the correlation between perceptual stage acoustic parameters, perception of concert hall acoustics and musical performance Kalkanjiev et al. [10] identified four main stage acoustic parameters, namely RT60, ST_{late} (Late Support - see Section 2.4.3), Early Strength (G_e) and Br, which they interpreted as *perceived duration of reverberation*, *reverberant energy*, *early acoustical support* and *timbre of reverberation* respectively.

Musicians' preferences for concert hall and stage acoustics are varied and the same performer may indeed express preferences for different acoustic characteristics of the performing venue. The desire for two different types of feedback from the performance venue is summed up in a quote from harpsichordist Tom Beghin who took part in a study of virtual stage acoustics by Woszczyk et al.

> *Musicians prefer smaller, narrower spaces where much of the emitted sound returns to them relatively early, but they also like rooms of larger cubic capacity where ambient sound does not become excessively loud or reverberant* [13].

Stage acoustic preferences of musicians can also be influenced by the repertoire performed, or by the performing forces employed.

In general, stage acoustic conditions are concerned with levels of support for the musician's own sound, ease of hearing others on stage and the effect of other musicians and musical instruments (as well as platforms, scenery, chairs etc.) on the stage [45]. These attributes are related to the ratio of measured levels of early and late energy in the sound on the stage, and hence Stage Support and Ease of Ensemble are considered here.

Early Ensemble Level *Hearing Others*

Gade[39] initially introduced Early Ensemble Level (EEL) measured across the stage, to reflect the balance of early to later arriving sound, in order to quantify the impression of

being able to hear fellow performers *Hearing Others*.

$$EEL = 10log\left\{\frac{E_{0-80ms}}{E_{0-10ms}}\right\}, dB \tag{2.10}$$

where E_{0-80ms} is the energy present in the impulse response between $0-80$ ms, and E_{0-10ms} is the energy present in the first 10 ms of the impulse response.

However, his later work found that EEL was directly related to ST_{Late} which can be used for assessing the ease of mutual hearing between musicians on stage for ensemble playing.

It is worth noting that stage acoustical conditions are very diverse; even within a single hall stage support values can vary in as much as 10dB (e.g. [51, 49]).

Stage Support *Hearing Oneself*

Gade [2] was the first to suggest the objective parameters of Stage Support (ST1 and ST2) to quantify the musician's impression of *Support* or *Hearing Oneself*.

Measures of support describe the energy ratio between the direct sound (of the performer) and reflected sound (from the room or stage area). It is measured with a source on stage at 1m above the floor and a microphone at a distance of 1m, to represent the relative position of musician and instrument.

$$ST1 = 10log\left\{\frac{E_{20-100ms}}{E_{0-10ms}}\right\}, dB \tag{2.11}$$

$$ST2 = 10log\left\{\frac{E_{20-200ms}}{E_{0-10ms}}\right\}, dB \tag{2.12}$$

$$\tag{2.13}$$

where $E_{20-100ms}$ is the total energy present in the impulse response between $20-100$ ms, $E_{20-200ms}$ is the total energy in the impulse response between $20-100$ and E_{0-10ms} is the total energy present in the first 10 ms of the impulse response. The reference time window of $0-10$ ms is used to include the direct sound but to exclude the floor reflection from the instrument.

Gade [39] found that musicians' impression of *Support* was significantly correlated to ST1 and ST2, and a ST2 value of around -12dB was preferred by orchestral players. He found that ST1(equivalent to ST_{early} was significantly correlated for the performing musician with the subjective impression of being able to "hear oneself".

Several measures of stage support are now used in the evaluation of concert hall stages (for example [43, 51, 49, 49]) including ST_{early}, ST_{late} and ST_{total}, where ST_{early}

is equivalent to ST1 as defined by Gade [2]. ST_{late} is the balance of energy between the direct sound and the late arriving energy, and ST_{total} is the balance of energy between the direct sound and the sound in the first second of the sound, calculated as follows:

$$ST_{late} = 10log\left\{\frac{E_{100-1000ms}}{E_{0-10ms}}\right\}, dB \tag{2.14}$$

$$ST_{total} = 10log\left\{\frac{E_{20-1000ms}}{E_{0-10ms}}\right\}, dB \tag{2.15}$$

Dammerud [45] compared measures of Stage Support on stages and found limited subjective relevance for musicians' sense of their own sound being *supported*. Although he did find that ST_{total} was useful for assessing the support from the room for hearing the sound from the musician's own instrument [45].

Strength

Dammerud suggested that measuring Strength (G) in both the audience area and on stage of concert halls as more robust than Stage Support (ST), not only relating to performers' subjective impression of Support but also in terms of objective measurements. He argues that although ST measurement should still be made to characterise stage acoustic conditions, future studies should also measure C_{80}, G, G_e and G_{late} both on stage as well as within the audience area [52, 45].

Voice support *Voice comfort*

Brunskog et al [53] studied spoken voice use and speaker comfort in a number of different rooms, including an anechoic chamber, a medium size lecture room and large auditoria and found that lateral and vertical early reflections were necessary for the speakers to feel comfortable with a perceived good level of support for the voice.

Brunskog, Gade, Payá-Bellester & Reig-Calbo [53] proposed a new measure of "Room Gain" (G_{RG}) which is described as "the degree of amplification offered by the room to the voice of a speaker at his/her ears, considering only the airborne paths" and defined as:

$$G_{RG} = L_E - L_{E,ach} \tag{2.16}$$

where L_E and $L_{E,ach}$ are the overall impulse energy level of an impulse response taken between the mouth and ear of a dummy head torso measured in the room and in the anechoic chamber respectively. Room gain has been shown to correspond to the subjective sense of the room adding *Support* to the voice.

In 2011, in a comment on the Brunskog et al. paper [53], Pelegrín-García [54] proposed a new measure of "Voice Support (ST_V)", which is an equivalent metric to Room Gain (G_{RG}) but using a more sensitive measurement method by introducing restrictions on the placement of the Head and Torso Simulator (HATS) used in the measurement. Making sure that the HATS is at least 1m away from any reflecting surface, ST_V can be calculated from a single impulse response measurement by windowing the impulse response signal to evaluate levels of direct and reflected sound separately. ST_V is then defined as the difference between the reflected sound and the direct sound from the HATS' mouth to ears impulse response.

$$ST_V = L_{E,r} - L_{E,d}, dB \tag{2.17}$$

where $L_{E,r}$ is the energy level of the reflected sound and $L_{E,d}$ is the energy level of the direct sound [54, p.1162]. Pelegrín-García noted that the measurement of Voice Support is related to Room Gain through the formula:

$$G_{RG} = 10log(10^{ST_V/10} + 1), dB \tag{2.18}$$

Reveberation *Perception of Reverberation*

Running Reverberation has been posited by Griesinger [55] as a measure of reverberance as perceived by a musician whilst playing.

Following listening tests where musicians were asked to match varying reverberation times and levels, Griesinger noted the levels of reverberation which musicians perceived to be matching had similar ST levels if the integration times were adjusted to 160ms. This lead him to propose that an objective parameter to account for subjective *Musician Self Support* might be Running Reverberation (RR160)

$$RR_{160} = 10log\left\{\frac{E_{0-160ms}}{E_{160-320ms}}\right\} \tag{2.19}$$

where $E_{0-160ms}$ and $E_{160-320ms}$ is energy in the impulse response in the two time intervals 0-160 ms and 160-320 ms respectively.

2.4.4 Performance spaces and musical style

Performance characteristics

It has already been suggested that musical performance changes according to the room acoustics of the performance space, and this will be discussed in further detail in Chapter 4.

It is also widely agreed that different styles of music are suited to different room acoustics. For example, highly contrapuntal music with many independently moving musical lines will be blurred and muddied in a performance space with long reverberation times. The same long reverberation time, however, would suit slow moving homophonic choral music or plainchant.

In their study of eleven Venetian churches Howard and Moretti found the "largest churches poor for performance of complex choral music involving advanced polyphony and/or multiple choirs" whereas in the larger churches higher frequencies were strongly dampened, leading to low values for clarity and brilliance, which meant they were perhaps best suited to plainchant consisting of one single melodic line [32].

Compositional styles

Distinct performance styles stemming from differences in room acoustics will in turn influence the compositional style of music written for a specific place. Throughout history musicologists, musicians and composers have noted that different spaces suit different styles and vice versa. A composer writing with a particular performance space in mind will alter their compositional style accordingly. For example, Henry Purcell's compositions for the resonant open spaces of Westminster Abbey differ in style to his music composed for the smaller Chapel Royal [56].

It is interesting to compare how performances of different compositional styles (ideally from the same composer) change to suit the room acoustics of their original performance space. In Section 4.7 three pieces by the same composer but in different styles have been recorded to allow such a comparison.

2.4.5 Adjustable room acoustics

Adjustable acoustic systems are often installed in buildings which are not primarily designed as concert halls, or in multi-purpose venues, where a number of different room acoustic settings are needed to accommodate performance of a wide range of musical genres, as well as spoken presentations.

Passive adjustable acoustic systems which rely on the physical opening/closing of boxes and drawing out of drapes are termed "passive acoustic systems", whereas electronic systems, employing microphones and loudspeakers are termed "active acoustic systems".

Active acoustic systems

An active acoustic system, sometimes known as a Reverberation Enhancement System (RES), can be used in auditoria or halls where the room acoustic properties are not

desirable, in order to improve the architectural acoustic characteristics, or to optimize parameters such as RT60 for specific purposes.

An active acoustic system detects sound in the auditorium via a number of microphones, processes the microphone signals electronically and then outputs the processed sound back into the auditorium [57].

Two main types of active acoustic systems exist, and there are a number of commercial systems are currently available:

- **in-line systems** employ a small number of directional microphones close to the performance platform and generate early reflections and late reverberation which is fed to the main auditorium (seating areas). Acoustic feedback is avoided by maintaining a high ratio of direct to reverberant sound [58].
  - LARES system (Lares -Lexicon Inc [59])
  - System for Improved Acoustic Performance (SIAP [60])
- **non-in-line (regenerative) systems** use microphones placed around the auditorium to increase reverberation time and signal processing methods are used to avoid instability [57, 61].
  - Wenger V-Room virtual rehearsal room
  - Meyer Sound Constellation System
  - Yamaha Active Field Control ([62])

For non-in-line (regenerative) systems a number of methods have been developed in order to increase the Gain Before Instability (GBI) of the system, to avoid or lessen the risk of acoustic feedback loops. For instance, by introducing time-varying gain to the microphone signals or using time-variant filters in the reverberation algorithms deployed [62].

Griesinger has an informative article on his own website [40] about his recent experiences with active acoustic systems in concert halls, and Poletti [57] gives a good overview of systems currently in use.

Electro-acoustic enhancement of rehearsal rooms

Lokki and Hippaka implemented an RES "active wall" system in a rehearsal room, consisting of an anechoic wall fitted with a number of loudspeakers, fed by microphones in the "stage" area of the room. Such a system can be used to enhance the reverberation in a (medium to large) rehearsal room [63]. Early reflections are present from the room

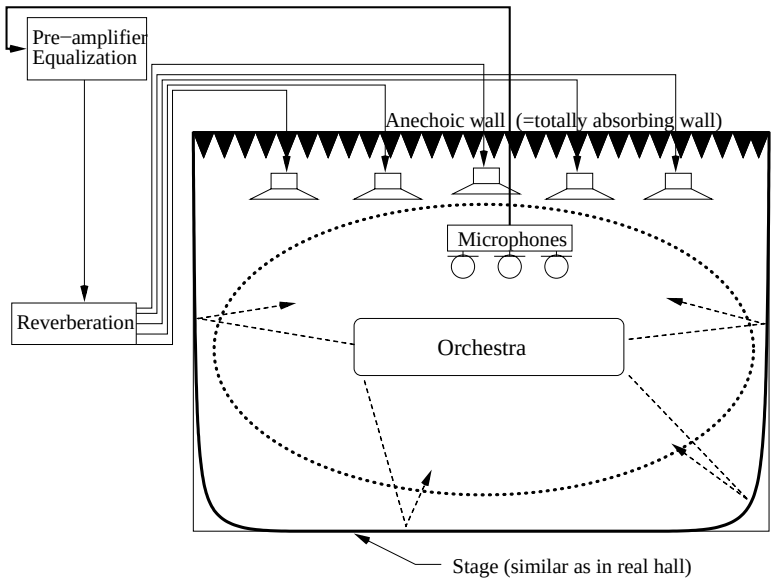


Figure 2.4: *Active wall system in rehearsal room for symphony orchestra from [63]*

itself, but later reflections and diffuse reverberation, which would otherwise stem from the main body of the auditorium, are modelled and output to the loudspeakers in the "active wall".

Lokki and Happaka's system incorporates a time variant reverberation algorithm, in which each of the four channels includes a delay line, a low-pass filter and a comb-all pass filter. The feedback coefficient of the comb-all pass filter is modulated over time by another continuous signal, such as a low frequency sinusoid. This has the effect of avoiding the positive acoustic feedback which would otherwise arise between microphone and loudspeakers at certain frequencies, leading to the characteristic "ringing" of feedback loops in the system and eventual instability. The time-varying delays introduced in the reverberation algorithm allow the gain before instability (GBI) to be higher. There are no perceived pitch changes in the reverberant sound in this system since the modulation shifts the frequency peaks in the spectrum in different directions.

Pätynen has developed a similar system suitable for use in small practise rooms with good results. However, the system is not designed to simulate early reflections, and this was recognised in the subjective responses from players using the "acoustically enhanced" practice rooms, who advised that more support for ensemble playing was desired [64]. Indeed one player suggested that "there can never be too much support" [64, p68].

Such systems, where the stage acoustics (early reflections) stem from the physical room, and the later reverberation is enhanced electronically, can be thought of as a step towards a full interactive Virtual Acoustic Environment (VAE) (discussed in Section 2.6).

The importance of room acoustic characteristics for performers and audience alike, together with methods used to enhance the room acoustic conditions of performance spaces have been outlined above. The next sections go on to describe the processes required in providing "virtual room acoustics" for a performing musician, which might facilitate the proposed investigation of musical performance attributes as they change with varied room acoustic conditions. Section 2.5 begins by detailing the technique of "auralisation", which is part of the fuller process of implementing a Virtual Acoustic Environment or room acoustic simulation.

2.5 Auralisation

As outlined in Section 2.3.3, an RIR can be evaluated through signal processing techniques to enable measurement of various time and energy based room acoustic parameters. However, in order to allow subjective evaluation of a room's characteristics, "to hear the room", the numerical data contained in a room impulse response must be "auralised".

Vorländer defines auralisation as: " the technique for creating audible sound files from numerical (simulated, measured, synthesized) data" [22, p103]. In this sense it is analogous to "visualisation" which renders numerical data in a visual format.

Kleiner et al. define auralisation more specifically relating to room acoustics as:

> ...the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled (sic) space [65].

Kleiner et al [65] describe four main types of auralisation :

- "Fully computed auralisation " - computer model used to predict binaural RIR, which is then convolved with sound source material
- "Computed multiple-loudspeaker auralisation " - computer model used to synthesize RIR, convolved with sound source on multiple-channels and presented over multiple-loudspeakers
- "Acoustic scale-model technique" - a physical scale model is produced, and audio source material played into the model after being scaled in terms of frequency, the resulting sound field is recorded and reproduced over headphones/loudspeakers
- "Indirect acoustic scale-model auralisation " - a physical scale model is made, binaural RIR of the model measured and convolved with sound source material

The first two in this list are now the most commonly used, although scale-model techniques were once popular, see for example [66, 67, 46, 24].

An RIR could also be obtained by in situ measurement (see Section 2.3.2) rather than modelling, in which case the process is more correctly understood as "room acoustic simulation" rather than complete auralisation where all elements of the auralisation chain are modelled.

2.5.1 Auralisation chain

Auralisation is a multi-step process in which each element of the process, or auralisation chain, must be carefully implemented in order to avoid undue colouration of the eventual auralised sound field.

2.5.2 Sound Source

Two sound sources in the auralisation chain may be identified, namely the sound source used in the original Spatial Room Impulse Response (SRIR) measurement and the sound used as the source material (often music or speech) for the auralisation. In studies of concert-halls, the measurement sound-source has traditionally been an omnidirectional loudspeaker on the stage, and the RIR captured at different points in the audience area.

Sound source for SRIR Measurement

Methods of SRIR measurement were outlined in section 2.3.2. Early auralisation techniques used omni-directional point sources as the measurement source and monophonic anechoic recordings as source material. The poor perceptual quality of auralisations made in this way means these early attempts have largely been superseded by techniques which now aim to capture and retain the directional properties of the sound source, i.e. the frequency dependent radiation characteristics of the musical instrument, speaker or singer within the auralisation chain.

To enable successful and natural sounding auralisation the directivity of the eventual source sound material must also be taken into account and replicated in the auralisation chain.

The directivity information could be included at the SRIR measurement stage, by outputting the measurement source signal with a loudspeaker which replicates the source directivity of the instrument or voice which will eventually be included in the auralised sound field.

Kearney has devised a method whereby a number of measured room impulse responses (RIRs) are combined in suitable ratios in order to approximate the directivity characteristics

of the source sound which improved the subjective evaluation of virtual acoustic recordings.

Kearney [44] showed that including some source directivity information in the signal chain, even if in a simplified form, for example, averaged across perceptually relevant frequency bands, increased listeners rating of the naturalness in auralisations.

Source Material - Directivity

At its most basic level the sound source in an auralisation can be represented by a monophonic source. However, for complex sound sources, such as music ensembles or orchestras, a point source will not suffice, since much of the spatial information of the orchestra and also the variable and frequency dependent radiation patterns of orchestral instruments are lost in this method [68].

Approaches in this area include directional filtering of an omni-directional source [69], or the use of multi-channel recordings [70]. Lokki and Pätynen have produced anechoic orchestral recordings where each instrument is recorded individually (with timing and dynamics aided by the use of a video of the orchestral conductor) and instrumental radiation characteristics are maintained [71].

Other authors, (e.g. [72, 73, 69]) have developed multi-channel recording techniques in order to preserve the source radiation characteristics of a single musical instrument, singer or group of instrumentalists.

Wang and Vigeant [74] found that using an omni-directional source for auralisations can lead to erroneous reproductions in terms of measures of C_{80} and RT60. Although low frequencies from the source loudspeaker are almost omnidirectional, higher frequencies output will be more directional in nature, so that the room in effect is less excited at higher frequencies and the receiver microphone captures less reverberant energy at higher frequencies. The direct sound then has more relative energy at higher frequencies than in the lower parts of the spectrum, leading to higher clarity values being calculated in the upper octave bands. Their subjective testing revealed that, when convolved with anechoic recordings of instrument or voice, highly directional sources (corresponding to a sixteenth-tant of a sphere) were distinguished from those using an omni-directional source.

Recent implementations of virtual auditory environments for performance (speech and music) have measured SRIRs using directional sources. Rindel [75] used a large number of microphones arranged around the instrumentalist and used the resulting directivity pattern in the sound source in a modelled auralisation. Subjective testing showed that a directional sound source was preferred by listeners.

Head and Torso Simulator (HATS) have been used to capture the Binaural Room Impulse Response (BRIR) for measurement of performance spaces for singers/speakers such as classrooms and concert halls, and for the implementation of virtual performance

spaces for singing or speech. For example, the measure of "room gain" used to characterise the support offered to the voice user in a room proposed by Brunskog [53] (see Section 2.4.3) is based on the measurement of BRIR using a HATS.

Cabrera et al [76] use Oral-binaural room impulse response (OBRIR) to describe the room acoustic response from the mouth to the ears of a dummy or real head measured with a HATS in particular to measure "room gain" and "stage support" as this represents more accurately the directivity of the spoken or sung voice. Although there is little published data in this area Cabrera et al. [77] show that the singing voice is highly directional especially in the 2 kHz and 4 kHz octave bands.

Anechoic source material

All studies mentioned above use anechoically recorded source material in the auralisations. However, there is an inherent problem in the use of anechoic source material for auralisation, in that a musical performance in an anechoic chamber will differ in many ways from a performance given in a concert hall. Musicians adapt their performance to suit the acoustic characteristics of the surrounding performance. Moreover, the anechoic chamber itself is an unnatural acoustic environment which many musicians find negatively impacts their performance due to the lack of auditory feedback from the room.

2.5.3 Room

Techniques for measuring Spatial Room Impulse Responses (SRIR) have already been outlined in Section 2.3.2. SRIRs can also be synthesized by computer models, for which a number of different methods exist.

SRIR Synthesis

A number of methods exist to computationally derive a SRIR from a 3-dimensional computer model, the most common of which are scale models, wave-based methods and ray-based methods.

Ray-tracing methods

Ray-tracing techniques consider sound waves as a ray, and track these rays as they travel and reflect within a room. The main disadvantage of ray-tracing techniques is that sound does not travel like a ray, and therefore a number of densely spaced rays need to be used in order to try to model the sound field faithfully.

Image source methods

A listener in a room perceives each reflection of a source within the room as if it were radiating from a point beyond the reflecting wall. Image source methods model reflections off the boundaries of a room as if they mirror virtual (image) sources beyond each boundary.

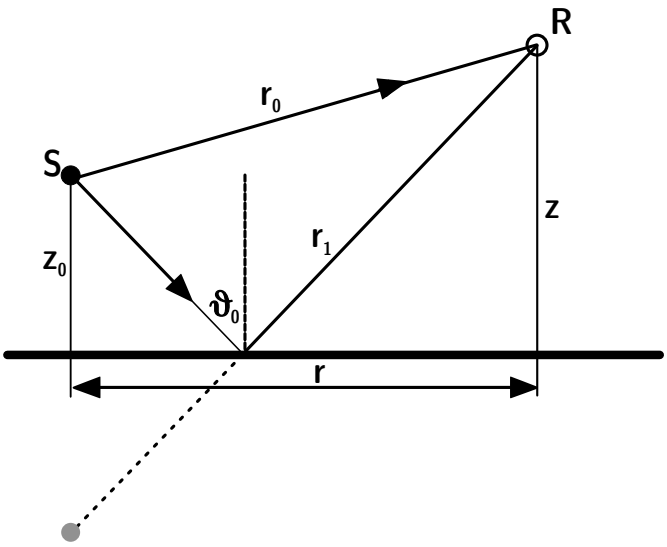


Figure 2.5: *Representation of image source model from [22, p.200]*

Figure 2.5 illustrates a simple case of this model; **S** is the sound source within the room, **R** is the receiver and the grey circle represents the virtual source which is modelled to mimic the reflection path from source to receiver as it would reflect off the virtual surface **r**. The main disadvantage of image-source models is that a large number of source images need to be computed to accurately model the sound field and scattering and diffraction characteristics of acoustic reflections are hard to achieve.

Wave-based methods

The wave equation is the complex mathematical equation which describes the propogation of sound waves in time and space. Wave-based methods attempt to solve the wave equation numerically, by generating a mesh of points to cover the space inside a room (Finite Element Method (FEM)) or the surfaces of the room (Boundary Element Method (BEM)). Using such methods a complex transfer function of the room can be obtained in the frequency domain which can then be transformed into a SRIR by inverse Fourier transformation. Finite-difference Time Domain (FDTD) methods make a time domain approximation of the wave equation by discretizing time and space and calculating the pressure or particle velocity for each point. Currently all such methods need to generate

large numbers of elements in order to accurately simulate the sound field, which is heavy on computation time.

Hybrid models

Vorländer [22] argues that some of these methods alone are not yet capable of simulating sound fields with sufficient accuracy, but that hybrid models which simulate specular and diffuse reflections achieve much more plausible results.

Current research seeks to provide hybrid models for RIR synthesis in order to keep the advantages of both methods, whilst ameliorating the drawbacks. One such model is the hybrid digital waveguide mesh [78, 79, 80] which uses FDTD for low frequencies and geometric modelling methods for high frequencies.

Comparison of modelled and measured SRIRs

At present auralisations using computer-modelled RIRs have never been judged to be as "authentic" as those based on real acoustic measurement data (measured RIRs) [44, 81]. However, auralisations based on measured SRIRs are challenging since a large number of listener positions need to be measured within the space.

On the other hand, a number of listener positions within an acoustic model can be generated, and techniques are being developed to facilitate the interpolation between such points to enable a virtual walk-through of a space, for example [82, 44].

Binaural Room Impulse Responses

A Binaural Room Impulse Response can be defined as "the signature of the room response for a particular sound source and human receiver" [65]. A BRIR is needed if the final auralisation is to be rendered binaurally over headphones or loudspeakers with cross-talk cancellation. Previous room acoustic simulations for singers have shown that many singers prefer not to wear headphones whilst singing (See Section 2.6.3 and Section 4.6.2 for more information on this point). BRIRs are not necessary in loudspeaker-based multi-channel auralisation reproduction methods and are therefore not explored further in this thesis.

2.5.4 Convolution

Convolution of the RIR with a sound source can be calculated either in the time domain using finite impulse response filters (FIR) or in the frequency domain using Fourier transformations (FFT) [22, p138].

2.5.5 Rendering and Reproduction

Spatial sound rendering is a two stage process which encompasses the encoding of audio signals to contain spatial information, and the subsequent decoding of the encoded signals to allow playback over loudspeakers or headphones.

There exist numerous spatial sound rendering techniques for example,

Wavefield Synthesis (WFS) Attempts to recreate the spatial sound field by synthesising wave fronts via a large number of loudspeakers [83]

Ambisonics Represents the spatial sound field with four co-incident microphone signals which are then decoded to simulate the pressure and velocity components of the sound field at the central listening position [84, 85]

Vector Based Amplitude Panning (VBAP) recreates the positions of virtual sound sources over multiple loudspeakers by amplitude panning between pairs or triplets of loudspeakers [86]

Spatial Impulse Response Rendering (SIRR) Spatial Impulse Response Rendering (SIRR) analyses the room impulse response to ascertain the direction and arrival time of reflections which are synthesised using Vector Based Amplitude Panning (VBAP) techniques; diffuse sound is reproduced across all loudspeakers [87, 88]

All the methods listed above are used in auralisations, but as yet there has been no systematic objective or subjective comparison of different methods. However, some authors have compared spatial audio techniques for specific purposes. For example, Kearney compared different techniques for rendering audio to multiple listener scenarios and found that VBAP was good for localization of stationary sources, but that Ambisonics was best for moving sound sources [44, 89].

Reproduction methods

Reproduction of 3-D sound fields can be implemented over multiple loudspeakers or headphones. Headphone based auralisation is not examined in great depth here due to the inherent problems in headphone use by singers (see Section 4.6.2 for more on this point).

For Kleiner et al., [65] the advantage of multi-channel convolution for multiple loudspeaker presentation is that the natural directionality of the sound field can be preserved. However, the disadvantage of such an approach has been that the auralisation needs to be presented in an anechoic chamber in order to avoid colouration from the room acoustics of the listening room. Nevertheless, research on "active techniques" to allow multiple loudspeaker systems to be used in rooms with some reverberation is ongoing [64, 90].

Ambisonics is in effect both a rendering and reproduction method. The spatial sound field is captured by four co-incident microphone signals (A-format) and then translated into B-format which comprises 4 signals; the omni-directional sound field (W channel) and X (front –back), Y(left –right) and Z (up –down) directions. The four channel B-format representation can be decoded to any number of arrangement of loudspeakers.

Ambisonics is simple to implement; indeed a number of software packages and digital audio workstation (DAW) plug-ins are now available for Ambisonic reproduction over multi-channel loudspeaker arrays (e.g. [91, 92, 93]). It is easy to adapt for presentation over any number of loudspeakers and the decoding process means that the whole sound field can be manipulated easily and rotated in space by applying simple trigonometric functions.

The disadvantages of Ambisonics include the size of the suitable listening area (sweet spot) which is extremely limited, since the method seeks to reproduce a sound field that has been captured by co-incident microphones. In addition the presence of a listener within the recreated sound field itself will lead to colouration of the sound field due to obstruction and reflection effects. These disadvantages are somewhat ameliorated through the use of higher order Ambisonics, that is adding groups of more directional components to the original B-format signals [85]. Despite its limitations, Ambisonic reproduction has been used successfully by a number of authors in auralisations for music and speech.

Guastavino et al [94] looked at subjective ratings for transaural (cross-talk cancelled binaural reproduction over loudspeakers), Ambisonics and stereo and found that Ambisonics was rated as more enveloping and immersive than the other reproduction methods.

In a loudspeaker-based room acoustic simulation for auditory research, Favrot [6, 95] found that a lower Ambisonics order was sufficient for auralisation of rooms where source localization is not of immediate concern.

2.5.6 Evaluation of Auralisation

Since there exist a number of variables in the auralisation chain, results of auralisation can differ greatly according to the procedures used and the choices made at all points in the process and indeed all of the components in the auralisation chain have the potential to spectrally colour the resulting sound field. A number of studies have evaluated the perceptual relevance of choices made at each point in the auralisation chain whilst others have used auralisation methods to make evaluations of room acoustic conditions.

Traditionally, research in the area of auralisation of concert hall acoustics has concentrated on simulating and evaluating room acoustic conditions from the listening position in the audience [96, 43, 46, 97, 36, 98, 97]. In addition most studies are undertaken under ideal listening conditions for a single listener.

Objective evaluation

One aim of evaluating auralisation methods is to compare the simulation against the real listening space (e.g., concert hall) to verify that the auralisation has produced sufficiently realistic results. The overall quality of an auralisation can be assessed objectively by comparing the room acoustic parameters of the modelled space with those of the real space.

For example, Lokki [99] compared reflection density, reverberation time (T30), EDT and C_{50} between modelled impulse responses and measurements in the real-space.

Favrot and Buchholz [5] tested a room auralisation over multiple-loudspeaker by comparing six different room acoustic parameters (RT60, EDT, C_{80}, G, Speech Transmission Index (STI) and IACC) and found only small differences occurred. However, in the same set-up speech intelligibility scores were improved by using fourth order Ambisonics instead of first order Ambisonics.

Just Noticeable Difference (JND)s for C_{80} values were studied by Cox et al., [38] by varying C_{80} values whilst keeping RT60 the same through delay and effect units played over 8 channels. they found that different musical motifs produced different JNDs; a Handel motif had a C_{80} JND of 0.44dB whereas a musical phrase by Mendelssohn had a C_{80} JND of 0.92dB.

Subjective evaluation

Auralisations can be evaluated objectively through comparison of acoustic parameters, but the subjective impression of the resulting acoustics should also be evaluated.

The use of auralisation techniques allows the researcher to control room acoustic parameters or simulate the room acoustics of a concert hall under laboratory conditions and greatly facilitates the subjective evaluation and comparison of auralisations (e.g. [41, 100, 101, 18, 19]).

However, listening test methodology for assessment of auralisations is not standardised, and yet the design of subjective experiments and subsequent analysis of results are not trivial. Since auditory memory is short, ideally a listening test will give the listener adequate opportunity to compare short audio examples a number of times, or to switch between auralisations in real-time.

Lokki and Savioja suggest that the evaluation methodology from audio codec quality testing can be used with good results [100, 102, 19]. A number of different methods exist which fall broadly into three groups; 1) absolute evaluations of audio signals in terms of subjective parameters e.g.*pleasantness* or *warmth* 2) different kinds of paired comparison test or 3) ratings of similarity. Methods for the subjective evaluation of audio material

are discussed further in Chapter 5.

Room acoustic evaluations can be influenced by the source material used, e.g. different musical instruments or styles of music. For example, Chiang and Huang [103] compared varied acoustic environments through binaural auralisation over headphones to assess listener preferences when convolved with two different musical sources, solo cello and solo xylophone. Overall listeners preferred RT60 values of 1.4s, but preferred longer RT60 times for cello music than for the xylophone.

Farina and colleagues [104] have compared the room acoustic conditions of five auditoria using a subjective questionnaire and related the objective room acoustic parameters to the subjective responses of listeners. Results have not been conclusive, for example they found that ratings of room size were not clearly related to the actual size of the auditorium simulated, but they did suggest that the stimulus SPL (strength factor) can substantially influence the auditory perception of distance.

Farina et al., [28] asked listeners to use nine pairs of adjectives to assess room acoustics whilst being able switch in real-time between auralisations of different concert halls using the same sound source material. They found that subjective results were not well correlated with objective parameters when the auralisation was reproduced over headphones, but results were improved when the listening test was reproduced over loudspeakers in a listening room.

2.6 Room Acoustic Simulations for Musical Performance

2.6.1 Virtual Acoustic Environments

The previous section outlined the steps involved in auralisation, that is the process of rendering a modelled or measured room acoustic sound field audible. Many of the techniques needed for room acoustic simulation are similar to those used in the production of virtual acoustic environments, of which auralisation is a subset.

The implementation of a VAE –also referred to as a virtual acoustic display –as defined by Savioja et al., [68] is a three step process involving : 1) Definition, 2) Modelling and 3) Reproduction (as illustrated in Figure 2.6).

The terms auralisation and virtual acoustics are often used interchangeably in the literature, but Savioja et al. [68] use the term “Virtual acoustics” specifically to cover the modelling of three main aspects of acoustic communication: 1) source, 2) transmission medium (room) and 3) receiver (listener).

Whereas they suggest that

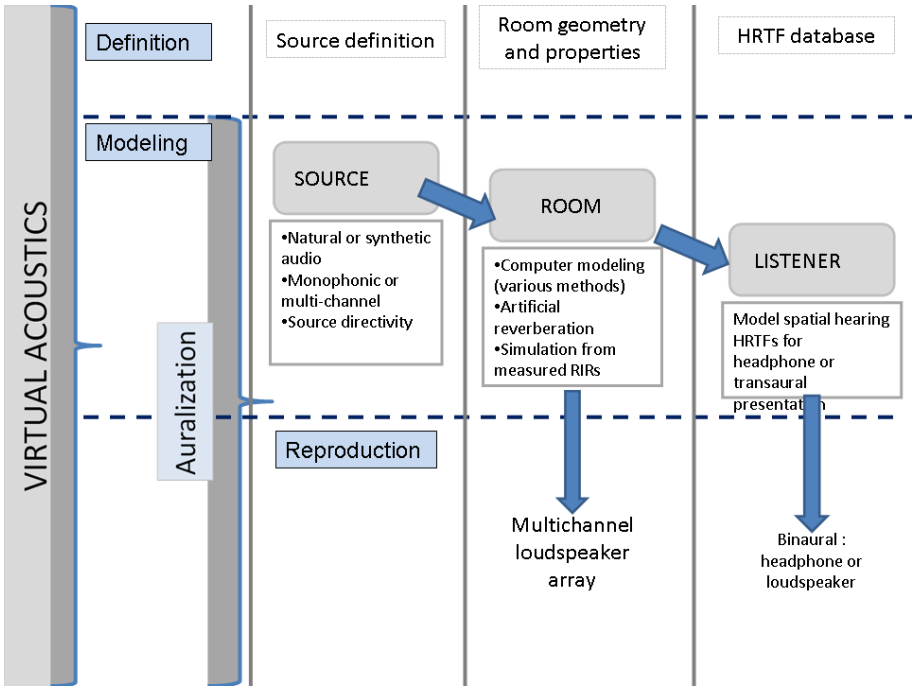


Figure 2.6: *Schematic Diagram of Auralisation Chain, adpated from [68]*

> the term "auralisation" is understood as a subset of the virtual acoustic concept referring to the modeling and reproduction of sound fields. [68, p675]

Thus, auralisation encompasses steps 2 and 3 of the implementation of a VAE. In order to implement a real-time interactive room acoustic simulation for the performing musician, auralisation techniques will be utilised, but certain modifications will be required at a number of points in the auralisation chain.

2.6.2 Interactive Room Acoustic Simulations

Most of the studies outlined above use "non real-time" auralisation i.e. the audio material is processed in advance of being presented to the listener for evaluation. However, there is an increasing interest in the use of "real-time" auralisation (within interactive VAEs) to investigate subjective responses to virtual acoustic displays, both from the perspective of the listener and performer.

Recent work by Lokki et al. has added interaction into a VAE which allows a listener to change listening positions in a simulated concert hall [68]. Another type of interactivity is to allow the listener to move about physically within a room auralisation, for example [105, 106].

2.6.3 Interactive Room Acoustic Simulations for Musical Performance

In his seminal work in 1989 [2, 39] Gade investigated musicians' impressions of a number of room acoustic parameters within a simulated sound field presented in an anechoic room. His main finding for solo musicians was that the impression of *Support* was related to the presence of early reflections. However, there were limitations to the simulated sound field where the direct sound, a small number of early reflections and the reverberation was simulated and fedback to the musician over four loudspeakers. Technical issues including the risk of acoustic feedback (also present in active acoustic systems, as considered in Section 2.4.5), sound signal colouration due to the use of a closely placed unidirectional microphone and problems with calibrating the levels involved led Gade to suggest that:

> The primary requirement for carrying out relevant experiments is that room acoustic sound fields –of proper realism and with the possibility of changing variables of potential importance –can be presented to musicians while playing [2].

A real-time room acoustic simulation for performance allows researchers to gather subjective responses from musicians to different room acoustic parameters e.g. [107, 3, 13, 49, 108].

Whereas a full VAE comprises three components which must be modelled –source, medium and receiver –a virtual acoustic for real-time performance always include a real source and receiver, and hence only the room acoustic characteristics are modelled (from measured or synthetic RIRs). For this reason such systems for real-time performance should be thought of as a subset of virtual acoustics, and might be more properly named a Real-time Room Acoustic Simulation.

A Real-time Room Acoustic Simulation (RRAS) for performance must include "sound interactivity", i.e., the listener is also able to make a sound and hear back the response of a simulated room in real-time. It must be noted that such a system not operate strictly in real-time as there is always some delay inherent in the processing and sound rendering. The term "real-time" is used here to contrast to "off-line" processing, and encompasses systems which operate in "pseudo-real-time" that is, as near to real-time as can be achieved and with only short latency times which are either not noticed, or accepted, by those who use the system.

A small number of research teams are using measured RIRs specifically to recreate a performer's position on stage/in the hall for presentation of an interactive acoustic environment, or room acoustic simulation, for the performer.

The implementation of a real-time room acoustic simulation with source interactivity involves some considerations to be made at all points of the auralisation chain and certain challenges arise, as are outlined in the following sections.

Sound source

The source signal –i.e., the musician's sound –must be accurately captured in a real-time room acoustics simulation (RRAS), whilst avoiding undue colouration of the signal due to microphone placement.

Ideally the source radiation pattern of the measurement source should reflect those of the eventual source in the real-time auralisation. For example, Martens and Woszczyk [109] modelled the source directivity during the RIR measurement in historic concert halls by using a group of omnidirectional loudspeakers in order to simulate the complex directional radiation of the pianoforte. During their subsequent performance experiment in the VAE a spaced microphone array was used to capture the sound of the pianoforte which was then convolved with the measured SRIR.

Source and receiver positions

In contrast to non-real-time auralisation methods, in real-time room acoustics simulations for performance the performer's sound acts as the sound-source and must be convolved with the RIR of the simulated space in (as near as possible to) real-time. In order for this to be properly achieved the Room Impulse Response must be modelled or measured from the position of the performer in the space, rather than from a listening position in the audience as is usual, i.e. the source and receiver positions should be co-located.

RIR editing

The RIR used in a real-time room auralisation (whether synthetic or measured) has to be edited to remove the direct sound. Whilst editing the RIR the Initial Time Delay Gap (ITDG) must be maintained i.e., the time between the direct sound and the first reflection. If the listening room used for the presentation of the RRAS has a floor, then the first floor reflection should also be removed [110, 8].

Ueno et al. [7, 9] presented an RRAS for musicians in a six-sided anechoic room, but noted the difficulty of simulating the first floor reflection since the floor in a real performance venue could be 1.5 m from the musician (or closer if seated), much closer to the musician than the loudspeakers located beneath the mesh floor in the anechoic chamber.

Real-time Convolution and Latency

"Off-line" convolution of the source sound with the SRIR can be performed in the time or frequency domain. Frequency domain processing is quicker but both methods can be computationally expensive, therefore, another method is needed for real-time applications. In real-time convolution the input signal is segmented and processed frame by frame, with the results being output in sequence [111].

In contrast to "off-line" auralisation outlined in Section 2.5 the sound source needs to be convolved in real-time with the RIR of the simulated room either via hardware or software applications (such as a VST plug-in in a Digital Audio Workstation), whilst ensuring that latency is low so that no delay in the early reflections and reverberant sound is perceived.

Miller et al. [112] infer from studies of the minimum audible movement angle for real sources, that the minimum perceptible end-to-end latency for a virtual audio system is about 70 ms for head movements (assuming a source velocity of 180°/s). They presume that the same threshold would apply for all types of source-listener motion, including when the source is fixed and the listener moves.

Wenzel [113] proposes a method of measuring end-to-end latency (which she terms TSL, Total System Latency) and similarly expects thresholds of 92 ms, 69 ms and 59 ms for slow, moderate and fast moving sources respectively.

Chafe et al. [114] showed that musicians asked to clap a rhythm with a partner were able to do so successfully when the partner's sound was delayed with time delays of up to 77 ms. However, delays above 14ms led to a deceleration of tempo, whereas short delays up to 11.5 ms helped stabilise the tempo.

Rendering and Reproduction

Most singers report that wearing headphones alters the balance of bone conducted and airborne aural feedback, which can be detrimental to their singing performance. Similarly Libeaux et al., [115] carried out investigations to determine whether a virtual environment could be used to assess the effect of room acoustic conditions and choir formation on singers' voices. They reproduced a virtual choir environment over loudspeakers and headphones, to assess vocal parameters such as intonation and vocal loudness. Singers preferred the virtual environment to be presented over an array of loudspeakers rather then a binaural rendering over headphones. Singers reported that they found timing, rhythm and intonation easier to control in the loudspeaker reproduction.

Avoiding instability

In a multi-channel room acoustic simulation the sound-source is located within an array of loudspeakers, which greatly increases the possibility of acoustic feedback loops arising, and the system becoming unstable. A close microphone can be used to capture the sound-source in order to ensure the level of direct sound is high relative to the reverberant sound and hence loudspeaker output levels can be higher without instability. However, close miking loses the directivity of the sound-source and can also cause colouration of the output signal.

2.6.4 Evaluation of real-time room acoustic simulations

This section outlines some of the recent studies which seek to evaluate the effectiveness of providing real-time room acoustic simulations for performing musicians.

Objective evaluation

As was seen in Section 2.5.6, an impulse response of a multichannel room auralisation can be measured and compared with the impulse response of the real room. In her room acoustic simulation for musicians, Ueno et al. found "quite a good accordance of the early reflection and reverberation process ... between the real field and the simulated one" [48].

Nevertheless, as is the case for auralisation in general, the objective comparison of real and simulated performance spaces needs also to be supplemented by subjective assessments from performing musicians.

Subjective evaluation

Lokki et al. [105] adjusted a simulated room impulse response in a VAE and asked participants to assess the size and shape of the virtual room with some good results, with most of the participants reporting that they could imagine that they were in a real space, although some small artefacts appeared such as echoes or "something unnatural" in the end of the reverberant sound.

Ueno et al. [3] simulated a sound field for the musician using a 6-channel system incorporating specially measured RIR on the stages of a number of concert halls of differing sizes. The instrument sound was picked up by a unidirectional microphone and convolved with the directional RIR in real-time. In an earlier study by the same team, the participating musicians gave subjective responses to the simulated sound field [48]; most players found the simulated sound field gave a natural impression of playing on a stage in a concert hall, although there was some reported tonal coloration in the higher frequencies, due to limitations in the simulation system.

Ueno et al [9, 3, 107, 110] used measured RIRs in order to examine musicians' responses to differing room acoustic conditions, by asking the musicians to evaluate their own performances in a number of simulated concert halls. In addition they have taken objective measures of performance such as tempo, vibrato rate and extent (see Section 4.3.11) and correlated these to subjective impressions of the room acoustics.

Woszcyk et al. [12] presented virtual stage acoustics in real-time to two violinists, whilst being able to adjust early, mid and late parts of the sound field. Through questionnaire and interviews with the musicians they found that three aspects were important for performance; 1) the balance of direct sound to reverberation 2) the loudness level of support and 3) the angle from which the early reflections arrive.

In another experiment where virtual stage acoustics were presented to a harpsichordist in real-time, [13] Woszczyk and Martens investigated the player's response to the room auralisation. The harpsichordist described different aspects of the room acoustics as forming a "triangle of listening", where the direct sound enabled him to perceive the sound of instrument, the early reflections and support were important for him to hear his own playing, and the reverberation in the concert hall allowed him to imagine what the performance might sound like to the audience.

2.7 Summary

After a brief introduction to acoustics in general and room acoustics in particular (Section 2.3), the objective and subjective evaluation of room acoustic conditions in concert halls has been described. (Section 2.3.3). Section 2.4 outlined some of the room acoustic parameters which are understood to be of importance for performing musicians, especially measures which describe the musician's impression of "hearing onself" such as ST, G, Room Gain and Voice Support (ST_v).

The chapter went on to describe the process of auralisation, that is, making numerical data audible, which is at the heart of any room acoustic simulation (Section 2.5). Recent and ongoing work in this area was described and the considerations necessary at all parts of the auralisation chain were considered, in order that the most natural sounding results are achieved.

The presentation of Real-time Room Acoustic Simulation for musical performance is a growing area of interest at present with research teams in Japan, Canada and Finland active in this area. The implementation of a VAE for real-time interactivity differs in a number of ways from that outlined in Section 2.6 and certain challenges arise. These challenges include the measurement of SRIR for use in the simulation, preserving the directivity characteristics of the source sound and convolving the input source sound with

the SRIR in near to real-time with no noticeable delay.

Whilst Ueno and Kato [15] have compared recordings of musicians in a number of simulated concert hall environments, there has not yet been any thorough comparison of musical performances produced in simulated and real performance spaces.

The following chapter describes the implementation and objective evaluation of a room acoustic simulation (Virtual Singing Studio (VSS)) which is designed to facilitate such an investigation. The VSS should ideally provide the singer with the required impression of *Support*, without undue colouration effects from the signal processing involved, whilst avoiding any risk of instability and acoustic feedback. These requirements will be addressed further in Chapter 3.

Chapter 3

The Virtual Singing Studio- Implementation and Verification

3.1 Introduction

This chapter describes the design, implementation and evaluation of the Virtual Singing Studio (VSS) - a (pseudo) real-time room acoustic simulation which will allow singing performances to be recorded in a simulation of a real performance venue, subjective evaluations by singers to be collated, and singing performances in the *real* and *virtual performance spaces* to be compared. In order to be able to change acoustic variables of the simulation, a real performance venue with manually adjustable acoustics was chosen as the basis for the simulation. This allowed the measurement of a number of different room acoustic configurations in one venue.

Section 3.2 describes an initial pilot experiment with the prototype VSS, which was undertaken to test methodology for the main set-up.

Section 3.4 outlines how room impulse response measurements were taken to provide the basis for the VSS and Section 3.3 gives more details of the real performance venue chosen for the VSS.

Room acoustic parameters are evaluated in the *real performance space* and the *virtual performance space* (as provided by the VSS) and presented in 3.5 in order to check that the simulation was correctly implemented.

Section 3.6 then goes on to describe subjective evaluation of the room acoustic simulation by a number of professional singers.

Figure 3.1: *Interior, as viewed from altar, of St. Patrick's Church, Patrington*

3.2 Prototype Virtual Singing Studio

3.2.1 Methods and materials

The production of the prototype Virtual Singing Studio is a multi-step process which involves capturing or synthesizing a SRIR, editing the SRIR to remove the direct sound and the floor reflection, real-time convolution of the microphone input (singer) on 3 channels (Ambisonic B-Format W, X and Y channels) and decoding the resulting output for presentation over a number of loudspeakers.

3.2.2 The Performance Space

The performance space which was simulated in this pilot study is the Parish Church of St. Patrick, Patrington, East Yorkshire. A large parish church, it is cruciform in shape with a length of 46m, width of 27m and internal volume of 8078 m^3. This church was chosen because its acoustic properties are known and documented and a number of computer models of the church have already been made and perceptually tested [81, 116].

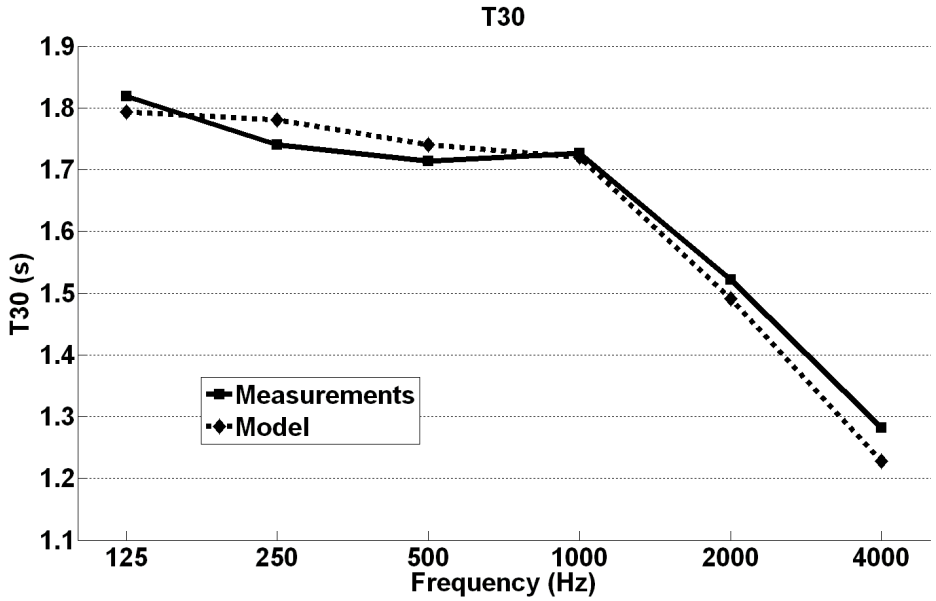


Figure 3.2: *Mean values of T30 for synthesized SRIR (model) compared with T30 values of measured SRIR of the same space*

3.2.3 Measured and Modelled SRIR

Measurements taken in the performance space are further documented in [116]. To summarize: an omni-directional source (Genelec S30D) emitted a 15 second long log sine sweep based on the Exponential Swept Sine (ESS) Method, as described in [25]. The recorded sine wave sweeps were deconvolved using "AirSupply", software written by Dr Simon Shelley (Audio Lab, University of York) available online at www.openairlib.net and described in [117].

The source was positioned in the middle of the crossing i.e. the area under the central tower, and the receiver was positioned at a distance of around 8.7m in the nave, outside the critical distance of the space which is approximately 3.76 m.

A model of the church had previously been produced by a fellow researcher (in ODEON 7.0 Auditorium), with surface absorption and scattering co-efficients, which had been selected and optimised as part of another study whose aim was to replicate the measured SRIR of the space as closely as possible. [116]

3.2.4 Acoustic Characteristics of the Space

In terms of reverberation time T30 the measured SRIR and the synthetic SRIR are closely matched, as can be seen in Figure 3.2. T30 is 1.73s at 1kHz and the mean EDT is 1.4s as quoted in [116].

3.2.5 Editing the Impulse Responses

In the virtual acoustic environment used in this performance experiment, the singer in effect provides the sound source (see Figure 2.6 in Section 2.5.1), so it is necessary to edit the SRIR to remove the direct sound component. It is also necessary to remove the floor reflection from the SRIR, as this is provided by the floor in the studio room in which the test takes place. The relative timing, i.e. the gap between the direct sound and the first lateral reflection, is retained in the edited SRIR. See section 3.4.6 for more on editing impulse responses in this way.

3.2.6 Real-time Acoustic Simulation

In the performance experiment the microphone signal was convolved in real-time over three channels (Ambisonic B-format, but the Z channel was not used as no overhead loudspeakers were available) with the edited SRIRs in Reaper [118] using the "ReaVerb" plug-in. The resulting signals were then decoded for a hexagonal Ambisonic loudspeaker array using the VST plug-in "B-dec High Resolution First Order Ambisonic B-format Decoder" [93].

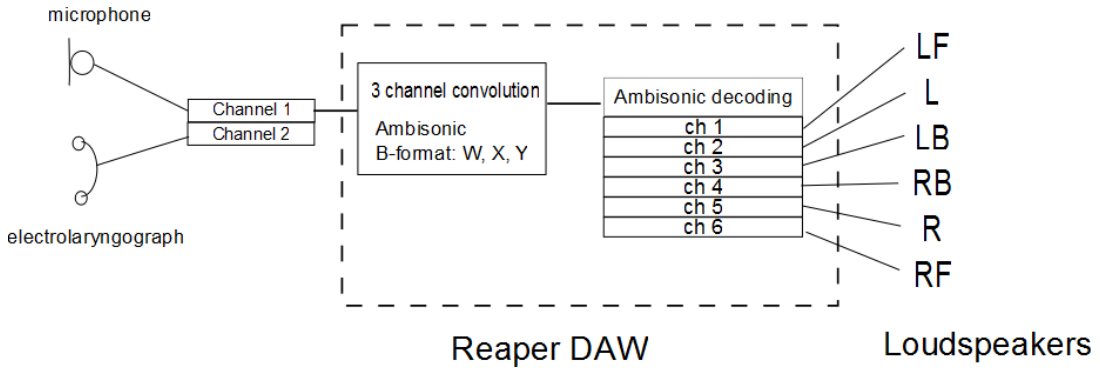


Figure 3.3: *Graphical representation of real-time convolution process in proto-type Virtual Singing Studio*

Six Genelec 8040a loudspeakers were used as illustrated in Figure 3.4, mounted at ear-height and positioned 1.75m from the central position for the performer.

A head-mounted AKG CK77 omnidirectional condenser capsule microphone was positioned approximately 5cm from the mouth, to capture the output signal from the singer as illustrated in Figure 3.5.

Singers were also asked to wear an electrolaryngograph, which captures a small electrical signal measuring the contact between the vocal folds during phonation. The electrolaryngograph signal was simultaneously recorded via the Reaper DAW to allow for potential closed quotient analysis (see Section 4.3.5).

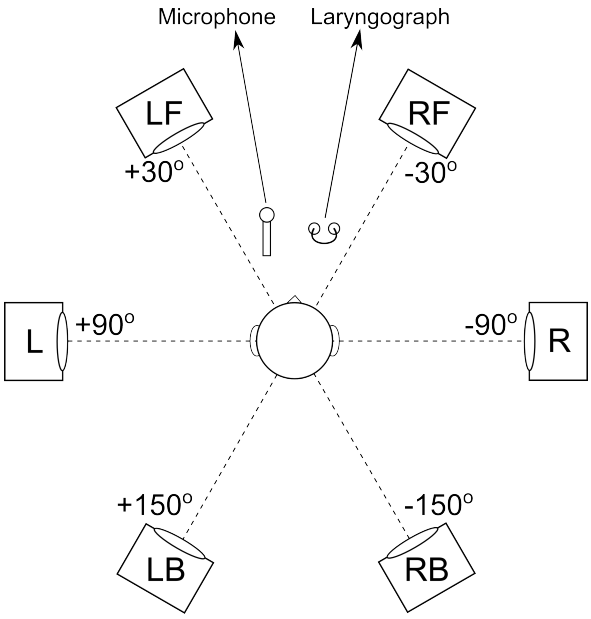


Figure 3.4: *Illustration of hexagonal loudspeaker array and performer position*

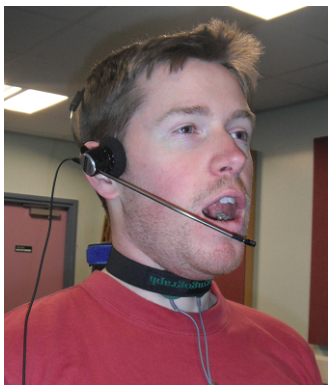


Figure 3.5: *Participant in trials of the prototype VSS wearing headset microphone and electro-laryngograph*

Two versions of the performance space simulation were presented to the singer (in a randomly chosen order); one using the synthetic SRIR ("modelled space") and one using the measured SRIR ("measured space").

3.2.7 Experimental Protocol

The singer was asked to warm up the voice as usual before a performance and to perform a prepared piece of their choice. After performing in one of two simulations the singer was asked to complete a short questionnaire based on a standard questionnaire produced and used by Arup Acoustics as published in [37, 32] (Appendix B includes the experiment protocol and a copy of the questionnaire).

Participants were asked to rate, on a scale between 1 to 10, characteristics of the room acoustics as follows:

- *Volume*: loud - quiet
- *Clarity*: muddy - clear
- *Reverberance*: dry - live
- *Envelopment*: frontal - enveloping
- *Intimacy*: remote - intimate
- *Warmth*: harsh/thin - warm
- *Brilliance*: dull - bright
- *Timbre*: unpleasant - beautiful
- *Overall impression*: poor - excellent

Eight participants took part in the study; six singers who sang and two participants who spoke in the space. One of the participants knew the space well and was able to give extra information about the naturalness of the simulation in comparison to his acoustic memory of the church surroundings.

3.2.8 Results

Subjective evaluation of the simulations

Figure 3.6 illustrates the average ratings from the participants of the two simulations i.e. the measured space (green) and modelled space (blue).

The modelled space was judged as *quieter*, with less *clarity*, *reveberance*, *Warmth* and *brilliance* than the measured space. Seven out of eight participants preferred the measured space when asked supplementary questions, and all felt that both versions presented a frontal soundfield, which is reflected in equal low ratings for both spaces of *envelopment*. The measured space scored more highly than the modelled space in terms of *overall impression*.

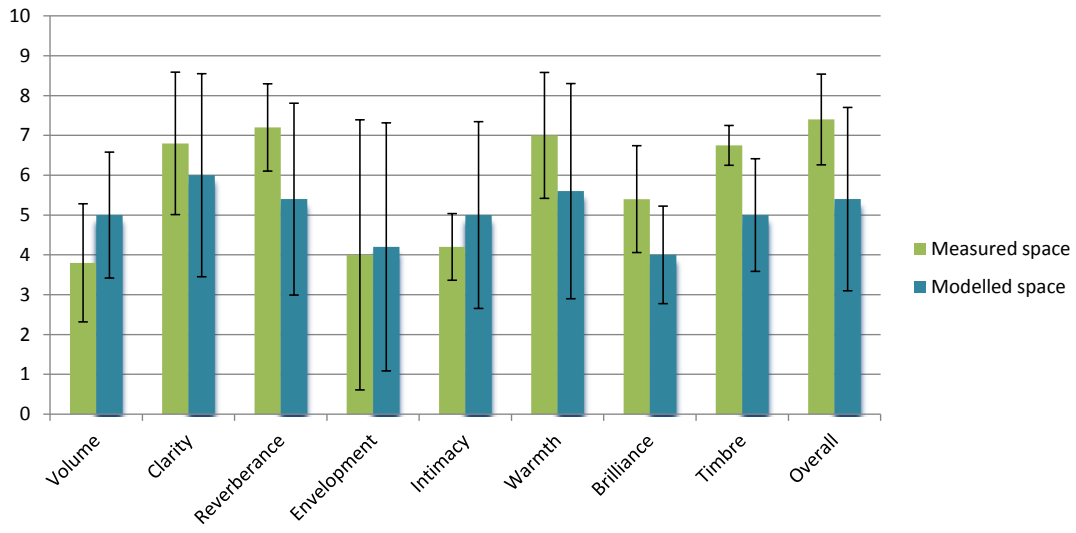


Figure 3.6: *Average ratings (with standard deviations) by singers for the acoustic simulations: left hand bar in each pair represents the "measured space", right hand bar in each pair represents "modelled space"*

Analysis of vocal performances

Some initial analysis of the vocal performance was carried out, although results are difficult to generalise since a very small number of singers took part, all with different levels of training and singing in different styles.

One singer in particular reported that singing in the listening room (an acoustically treated studio, with short reverberation time used for audio recording and production) without the "virtual singing studio" system resulted in difficulties maintaining tuning, which he thought stemmed from the lack of aural feedback provided by the acoustically dead room. However, an analysis of the average fundamental frequencies within note classes of sung pitches G3 (196Hz) and C4 (261Hz) for this singer shows that intonation patterns with the system turned on or off (treated listening room) are not consistently distinct.

	System ON		**System OFF**	
Note name	***Mean F0 deviation***	***St.Dev.***	***Mean F0 deviation***	***St. Dev***
G3	15.9	2.2	24.0	3.3
C4	21.2	3.5	12.7	2.8

Table 3.1: *Average difference of sung note from mean measured F0 for each note class (cents) and standard deviation (cents) with system turned on and off*

3.2.9 Discussion

Subjective evaluation

This investigation involved a small number of participants and yet produced some interesting results for the further improvement and development of a Virtual Singing Studio.

The subjective questionnaire included some characteristics which participants felt were not very relevant for the experience of singing in the real-time simulation. For instance participants found it hard to assess *Clarity, Intimacy, Timbre* and *Brilliance* of the simulation. This area needs further investigation as certainly some of the traditional room acoustic parameters are not relevant when the listener and sound source are the same person. For example, *Timbre* and *Brilliance* relate to the spectral quality of the sound in the performance space and participants seemed to find this hard to assess.

Although T30 values are well matched between the modelled and the measured space it is notable that all participants rated the measured space as more reverberant. It is interesting to note that a study by Bonsi [37] looked at the audience response in eleven Venetian churches, and concluded that the subjective evaluation of *Reverberance* correlated not only to T30, but also EDT. The EDT value for the modelled space is 1.7s and 1.4s for the measured SRIR. All participants also rated the measured space as "warmer" than the modelled space. Bradley has shown that "warmth" can be related to high levels of energy at low frequencies and the bass ratio of reverberance [42].

SRIR measurement

The measured SRIR used in this study was obtained with the source and receiver positions at a distance of around 8.7m and the synthetic impulse response was based on source and receiver at the same positions. This may explain the partcipants' rating of the sound field as frontal and not enveloping. In contrast, in order to correctly simulate the sound source and sound receiver positions of a single performer (mouth and ears) the source and receiver positions should be very close, or indeed co-located as far as this is possible within the measurement procedure (see Section 3.4.1).

3.2.10 Summary

The main conclusion from this pilot study was that although participants rated the two different simulations as being "good enough", they were able to identify a difference between the simulation based on a synthetic SRIR and that based on a measured SRIR. Moreover the simulation based on the modelled SRIR was rated less well in terms of *Reveberance*, *Warmth*, *Clarity* and *Overall Impression*. It was therefore decided to use

only measured spatial room impulse responses for the main version of the Virtual Singing Studio as described in the following section.

3.3 The Real Performance Space

The National Centre for Early Music is housed in the medieval church of St. Margaret's in York. In 1999/2000 the redundant church was refurbished and converted into a performing arts venue, especially for concerts of Early Music (music written prior to 1750). Figure 3.7 shows the interior of the church space prior to renovation, and Figure 3.8 illustrates how the space looks when used for a concert.

Figure 3.7: *Interior of National Centre for Early Music Prior to refurbishment, from the archives of the NCEM, used with permission*

The performance space has an internal volume of 3600m^3 and is equipped with a passive variable acoustics system which gives performers the opportunity, via manipulation of wall panels, boxes and ceiling drapes, to adjust the overall absorption in the space and subsequent reverberation time. Some of these panels are shown in Figure 3.9; the lighting racks in the roof void can also be seen in this picture, although the black ceiling drapes which are also housed here are difficult to make out against the dark woodwork.

3.3.1 Room acoustic configurations

The adjustable acoustics system allows at least five different configurations of the performance space. For this research project three different configurations were chosen

Figure 3.8: *Concert at National Centre for Early Music to illustrate position of staging against back wall (right of picture)*

Figure 3.9: *Position of a number of acoustic panels on the back wall*

for measurement and investigation as outlined in Table 4.2. Mean T30 values at 1kHz, measured across 26 receiver positions for the three acoustic configurations considered are quoted in Table 3.2 and are evaluated by Foteinou [81] who undertook a full survey of the space to inform room acoustic models.

Full details of room acoustic parameters evaluated in the *real performance space* for

Activity	Acoustic Boxes	Ceiling Drapes	Acoustic Qualities of Space	Mean T30 at 1kHz (s)
Large choral (LC)	All boxes closed	All drawn back	Highest reverberation, warmth, spaciousness	2.22
Music recitals(MR)	All boxes closed	All drawn out	Even balance between clarity and reverberation discrete sounds stand apart clearly but ample reverberation	1.83
Lectures and speech (SP)	All fully open	All drawn out	Sound absorbent space, giving maximum clarity for speech	1.32

Table 3.2: *Summary of acoustic configurations in the real performance space*

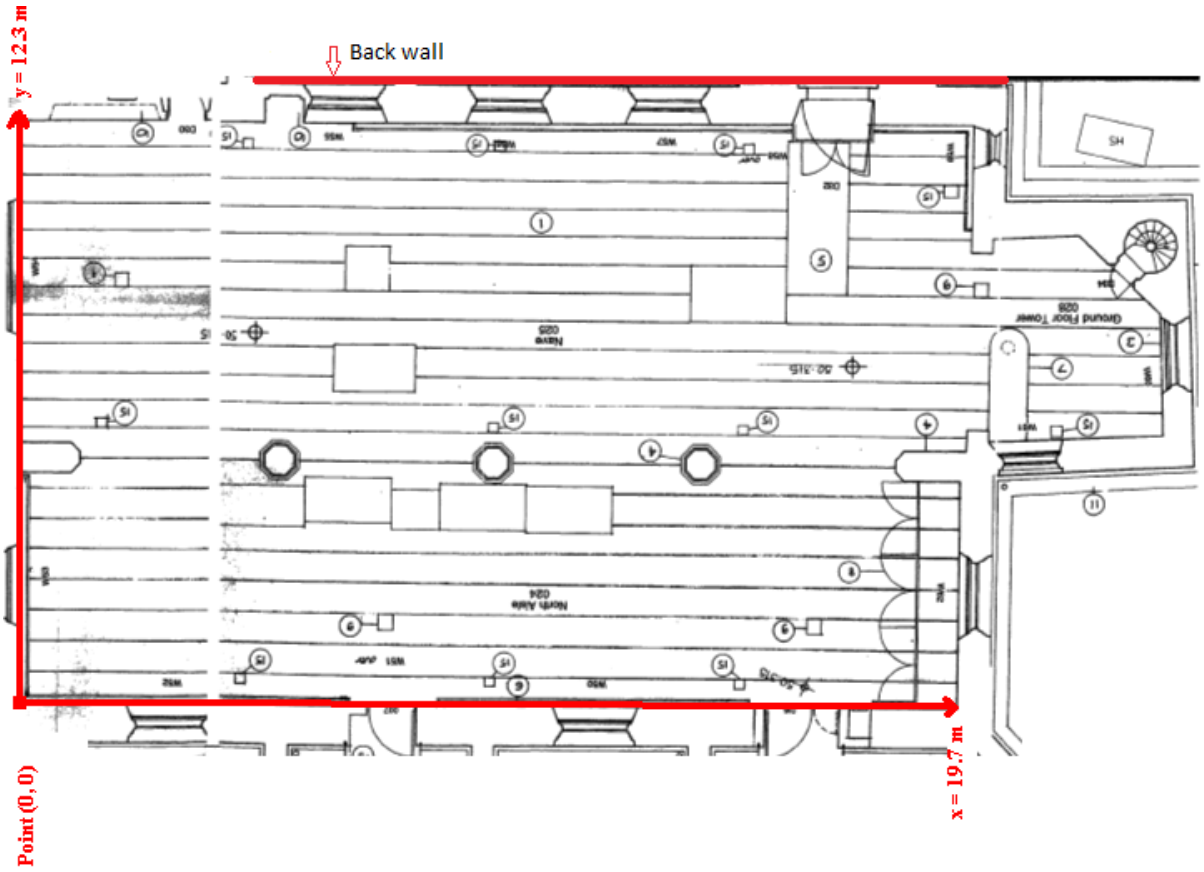


Figure 3.10: *Floor plan of National Centre for Early Music, York*

performer positions A - D and at the listener position can be found in Appendix D and are summarised in Section 3.5.2.

3.4 The Virtual Singing Studio Implementation

The investigation of a prototype version of the Virtual Singing Studio (Section 3.2) showed that all singers who took part in the pilot study preferred the measured SRIR simulation,

judging it to sound 'warmer', 'more natural' and 'more reverberant' and so it was decided to use measured SRIRs for the main implementation of the Virtual Singing Studio as described in this section.

3.4.1 Measurement of Spatial Room Impulse Responses (SRIR)

SRIR measurement sound source

The B-format room impulse responses were measured in the *real performance space* at a sampling rate of 96 kHz using the Exponential Swept Sine (ESS) Method as developed by Farina and described in [25] and outlined in Section 2.3.2.

Although traditionally an omni-directional source is used for the measurement of spatial room impulse responses, more recently authors have used directional sources for virtual auditory environments as is discussed in Section 2.5.2. Directivity of the source is also important during the auralisation chain as is discussed in Section 3.4.3 and others have used a Head and Torso Simulator for the measurement of the SRIR for this reason. As no HATS was available for this thesis it was decided to use a Genelec 8040 loudspeaker as an approximation of a suitable directional source for auralisation purposes.

SRIR for listener

SRIRs were measured from four separate positions to reflect the placement of individual members of a quartet of singers (see Section 4.7). Four performance positions were chosen, reflecting the positions freely chosen by a quartet of singers.

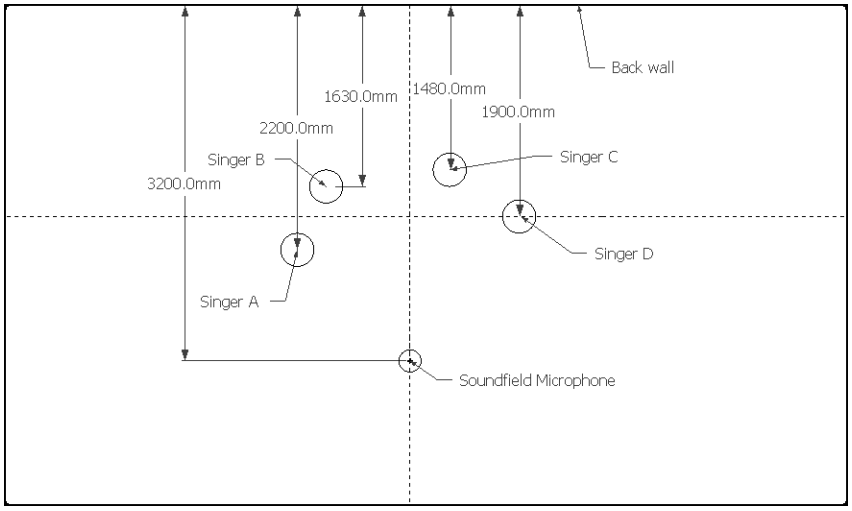


Figure 3.11: *Singer positions relative to back wall*

Singer positions are marked: A - Soprano; B - Alto; C - Tenor; D - Bass and their placement relative to the far wall (back wall) of the venue are illustrated in Figure 3.11

Figure 3.12: *Relative performance positions of vocal quartet indicated by music stand placement*

and a photograph showing the placement of music stands for the quartet of singers using these positions is shown in Figure 3.12.

For each performer position the loudspeaker height reflected the actual height of the singer involved, and four sets of measurements were made with the loudspeaker positioned at four azimuth angles (0°, 90°, 180°, 270°).

A Soundfield SPS422B microphone was positioned to reflect a typical audience position (2nd row of audience) as seen Figures 3.13 and 3.14. Due to time limitations only one set of the four source performer positions SRIRs were measured in this way in the second acoustic configuration (Music Recital (MR)).

Measuring SRIR for performer

SRIRs were also recorded to specifically emulate the performer's experience i.e. with source and receiver co-located. A Genelec 8040 loudspeaker emitted a 15 second long log sine sweep, but in this case the Soundfield SPS422B Microphone was positioned directly above the loudspeaker as pictured in Figure 3.15.

As a HATS was not available, this loudspeaker and microphone arrangement was chosen to mimic the mouth and ears of the performer as closely as possible.

Performer positions

At each performer position (A, B, C, D) the loudspeaker height was adjusted, and four sets of measurements were made with the loudspeaker positioned at four azimuth angles

Figure 3.13: *Illustration of Listener SRIR measurement position*

Figure 3.14: *Photograph to illustrate listener position Soundfield microphone position at head-height of seated audience member*

(0 °, 90 °, 180 °, 270 °)

3.4.2 VSS Implementation

The signal chain involved in the implementation of the VSS is illustrated in Figure 3.16 and further details of all elements are found in the following sections. The singer's voice is captured by the head-worn DPA4066 microphone and is converted into digital format via the RME Fireface800 external soundcard (Section 3.4.3). The voice signal is then

Figure 3.15: *Position of Soundfield Microphone placed above the Genelec 8040 loudspeaker used to measure performer position SRIRs in the real performance space*

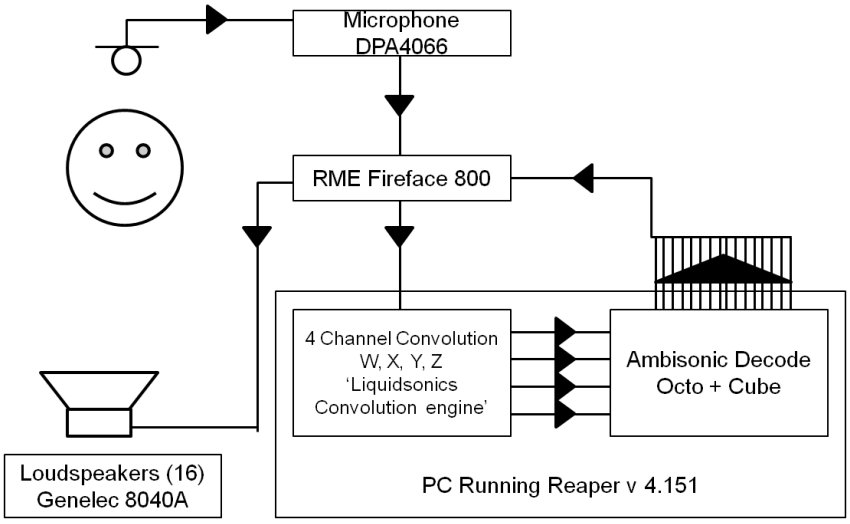


Figure 3.16: *Graphical representation of the processing chain involved in the VSS*

convolved over four channels in the Reaper Digital Audio Workstation (Section 3.4.4) with the measured SRIR (Section 3.4.1). The reverberant sound is then decoded for Ambisonic presentation over a 3 dimensional array of loudspeakers (Section 3.4.4).

3.4.3 Capturing the voice signal

Using a head-worn DPA4066 microphone means that the ratio of direct-to-reverberant sound captured at the microphone is high, which ensures that the problem of acoustic feedback and instability can be avoided (the output level of the loudspeakers can be higher before instability becomes a danger). It is also unseen so that the singer is not effected by the "PA effect" of seeing and knowing they are using a microphone, which could in turn cause the singer to alter their performance to incorporate a perceived need for microphone technique.

As an alternative to a head-mounted microphone, a cardioid microphone attached to the ceiling over the head of the singer was trialled for use in the VSS by a small number of singers. Although the overhead microphone affords a high degree of realism, due mostly to being unseen by the performer and therefore not contributing visually to the PA effect, some singers reported informally that they perceived the reverberated sound field being played back at a higher pitch. It was thought that this perceptual effect might arise from spectral balance of the reverberant field produced in the VSS stemming from colouration in the auralisation chain due to microphone type and placement.

Method

In order to compare a number of different microphone placements and types, a recording of a female singer was made in an anechoic chamber simultaneously on 3 different microphones:

Head-mounted microphone DPA 4066 (as used in the VSS) positioned 5 cm from singers mouth

Baseball cap microphone DPA 4066 attached to the peak of a baseball cap worn by the singer (a possible alternative placement) at 10 cm from singers mouth

Overhead microphone AudioTechnica PRO45 Cardioid Microphone positioned at 20cm directly above the singers head

Far microphone AKG C414 XLS at a distance of 1m from the singers mouth

The frequency response plots of the overhead (AudioTechnica) and head-mounted (DPA 4066) microphones are shown in Figure 3.17.

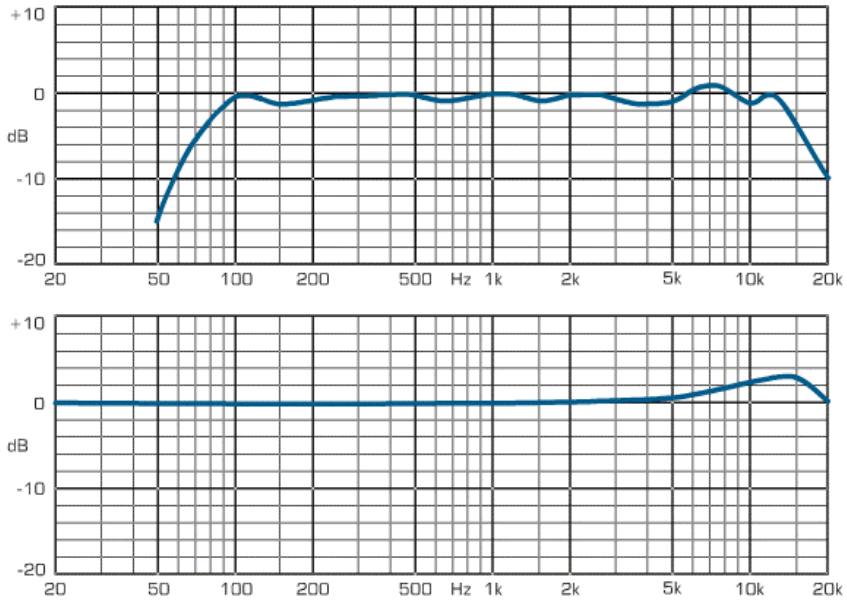


Figure 3.17: *Frequency responses of AudioTechnica Pro45 Cardiod (upper panel) and DPA 4066(lower panel) microphones from [119]*

The singer was asked to perform an excerpt from a prepared piece, and three English language spoken word tongue-twisters on a sustained pitch (e.g. "Peter Piper picked a peck of pickled pepper"). Recordings were made using a TASCAM DR-680 recorder at a sampling frequency of 96kHz. Wideband spectrograms are plotted of the recorded singing for each microphone, using the voicebox MATLAB toolbox [120] for a frequency range up to 10,000 Hz using a Hamming window with bandwidth of 200 Hz. Long-term average spectra are calculated by computing the average power spectra across short time frames, to provide a 'typical' spectral envelope of the sound source recorded [121] allowing longer term spectral characteristics of the recordings to be more easily compared. Long-term average spectra (LTAS) were computed using a MATLAB function based on Monsons [121] analysed over 2048 data points resulting in a frequency resolution of 46.87Hz (96000Hz/2048).

Results

Comparisons are made here between recordings captured at the **Head-mounted Microphone** and the **Overhead Microphone**. It is striking in Figure 3.18 that although there seem to be gaps in the spectrum of the head-mounted microphone recording around 5 kHz and 8 kHz, energy is still present in these frequency regions during the impulsive release of the plosive /t/ seen as vertical lines in the spectrogram above, for example, at around 1.5 secs and just before 6 secs.

In addition it can also be seen in Figure 3.18 that the spectrogram of the phrase

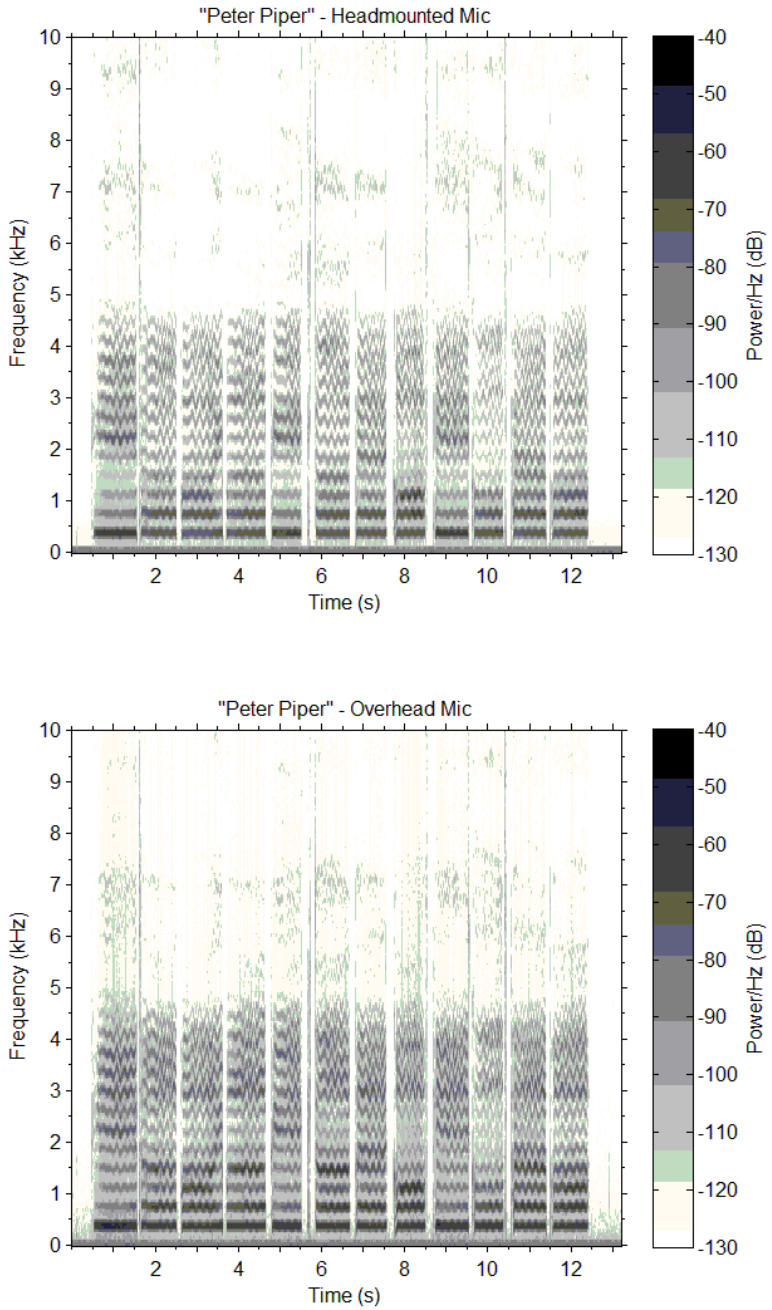


Figure 3.18: *Spectrograms of sung phrase recorded at head-mounted (upper panel) and overhead (lower panel) microphones.*

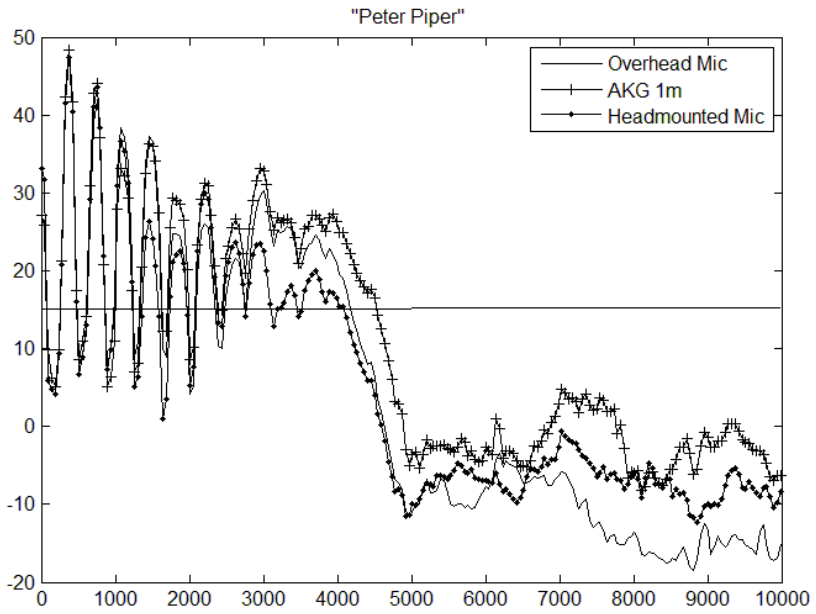


Figure 3.19: *LTAS of sung phrase recorded at overhead, head-mounted and 1m distance microphones*

recorded using the overhead microphone shows generally more energy in the spectrum above 5 kHz than the head-mounted microphone recording. These differences at higher frequencies can be more readily distinguished when long-term average spectra are plotted for the same recorded phase using different microphones (Figure 3.19) whereas at lower frequencies the microphone responses are very similar.

If the difference in response between the overhead and head-mounted microphone is plotted, as in Figure 3.20, the higher energy in frequencies above 5000Hz is more easily seen.

The overhead microphone placement also appears to boost frequencies below 100 Hz due to the proximity effect caused by the cardioid microphone. On the other hand there are prominent dips in the spectrum around 1500Hz and 3000Hz which are most probably due to the shadowing effect of the singers head.

Summary

Although the overhead cardioid allows a realistic feel to the simulation because it is out of sight, some singers have perceived the simulated reverberant field to be sharper in pitch than expected. This could be due to colouration introduced by this microphone in the region above 5 kHz, which shifts the spectral locus - the balance of energy between higher frequencies and the region around the fundamental frequency - of the reverberated sound. A shift in spectral locus has been shown to affect listeners' perception of the pitch of a tone [122] even if the fundamental frequency remains unchanged [122].

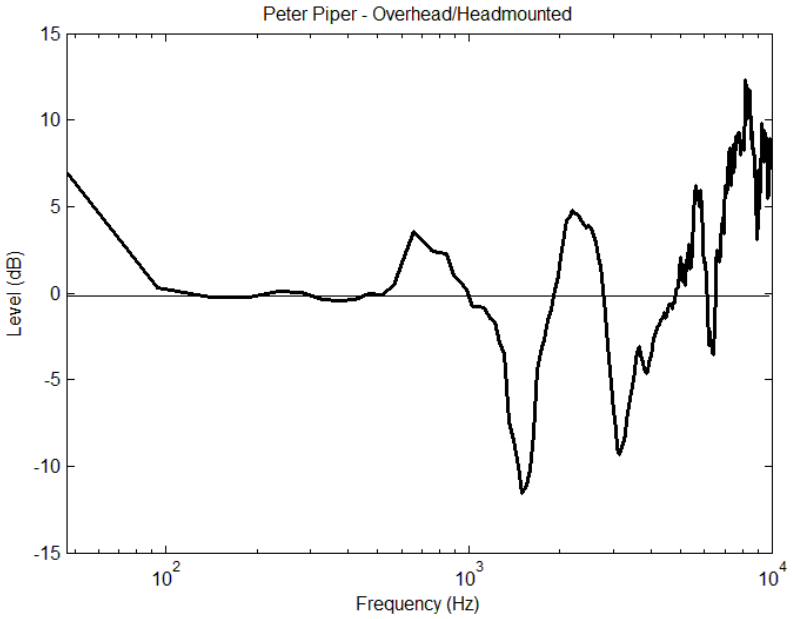


Figure 3.20: *Difference between long-term average spectra of the overhead (20cm from mouth) and head-mounted (5cm from mouth) microphone recordings of the phrase "Peter Piper" recorded simultaneously*

The head-mounted DPA4066 microphone was generally well liked by singers who trialled the VSS, and although one concern is that this microphone gives undue emphasis to mouth noises and plosive sounds (/p/,/t/,/k/ etc.), as can be seen in spectrograms in Figure 3.18 ; moving the microphone further back and away from the air-stream ameliorates this effect.

Directivity of sound source

There are however some disadvantages associated with using a head-mounted microphone, one of which is the loss of directivity information in the signal. The voice is, in effect, "spatially sampled" and treated as a point source within the auralisation, losing the directivity pattern of the voice signal. This point source then convolved with a SRIR which has originally been captured with a directional source (non omni-directional loudspeaker). So, although the direct sound includes the normal directivity pattern with which the singer is familiar, the simulated room reflections and reverberated sound are fashioned with a directivity pattern (that of the source loudspeaker used in the initial SRIR measurements) which does not fully correspond to that of the singing voice.

However, it is hoped that since the directivity of the Genelec 8040a loudspeaker used has similar directivity characteristics [123] to the singing voice - subcardiod at low frequencies becoming increasingly directional in nature in the higher octave bands - that this would not impair the plausibility of the simulation unduly. Future improvement

to the VSS might include some method of capturing and replicating the directivity of the singing voice in the simulation by developing real-time implementation of methods proposed by Vigeant [70] and Kearney [44] (see Section 2.5.2).

3.4.4 Rendering the Soundfield

A number of spatial sound rendering techniques exist as outlined in Section 2.5.5. Ideally, to ensure portability and future-proofing the encoding process in the simulation should be "blind" to the eventual playback system used, so that possible different loudspeaker layouts can be used to reproduce the 3 dimensional sound field with the correct decoding.

Since first order Ambisonics was used successfully in a loudspeaker-based room acoustics simulation by Favrot [6] and localization of the source is not an issue for the simulation in this project (since the singer is both the source and listener) first order Ambisonic decoding was chosen to render the soundfield.

Real-time convolution

The vocal signal captured by the head-mounted microphone is convolved over four channels (Ambisonic B-format) with an edited version of one of the measured Perfomer SRIRs (see section 3.4.6) in Reaper [124] using the low-latency convolution reverb audio processor Reverberate plugin by Liquidsonics [118].

The forward facing SRIR (0°) was replicated in the VSS only and singers using the system were asked to remain facing in one direction, which all singers did naturally

Ambisonic decode

The first order Ambisonic decoder used is provided as a VST plugin by Bruce Wiggins (Wigware) [85] and freely available at [92].

Reproduction

It was shown by Libeaux et al. [115] (see Section 2.6.3) that singers preferred a virtual environment to be presented over an array of loudspeakers rather than a binaural rendering over headphones and that with loudspeaker presentation singers reported that they found timing, rhythm and intonation easier to control.

Wearing headphones alters the signal chain between the mouth and the ear of the singer; having an enclosure over the pinnae alters the balance of low-frequency to high-frequency energy transmitted to the singer's ear. Singers rely on four types of acoustic feedback : bone conducted sound, air-borne direct sound, reflected sound and kinesthetic feedback (see Section 4.6.1 for more on this point). For many singers the disruption to the normal

balance of air-borne and bone-conducted sound which arises from wearing headphones makes for an uncomfortable singing experience. Since the goal of this study is to recreate the performance environment as effectively as possible, loudspeaker presentation was chosen: Sixteen Genelec 8040 loudspeakers, eight in the horizontal ring at ear-height (loudspeakers numbered 1 - 8 in Figure 3.21, together with a cube arrangement of ceiling and floor loudspeakers (ceiling A-D and floor E-H as illustrated in Figure 3.21).

The system is housed in an acoustically-damped listening room measuring 4.7 m x 10.8 m. x 2.6 m with a short reverberation time (T30 0.196 s and EDT of 0.162 s at 1kHz). The performer stands at the central point of the loudspeaker array, at a distance of 1.95m from the loudspeakers. Figure 3.22 shows the positions of floor, ceiling and horizontal ring of loudspeakers. Also suspended from the ceiling centrally is a metal housing for a data projector - when the VSS is in use this metal housing is covered by (removable) absorption treatment as pictured in Figure 3.23 and seen also in Figure 3.22 above the head of the singer's poistion marked by KEMAR (head and torso mannequin).

A heavy acoustic black curtain (kilowool) surrounds the loudspeakers to provide additional mid and high-frequency acoustic absorption. After initial trials with singers, who reported that being able to see loudspeakers might affect the way they would perform, an additional curtain of thin white cloth (muslin) was manufactured and hung in a circle between the singer and the loudspeakers obscuring them from view.

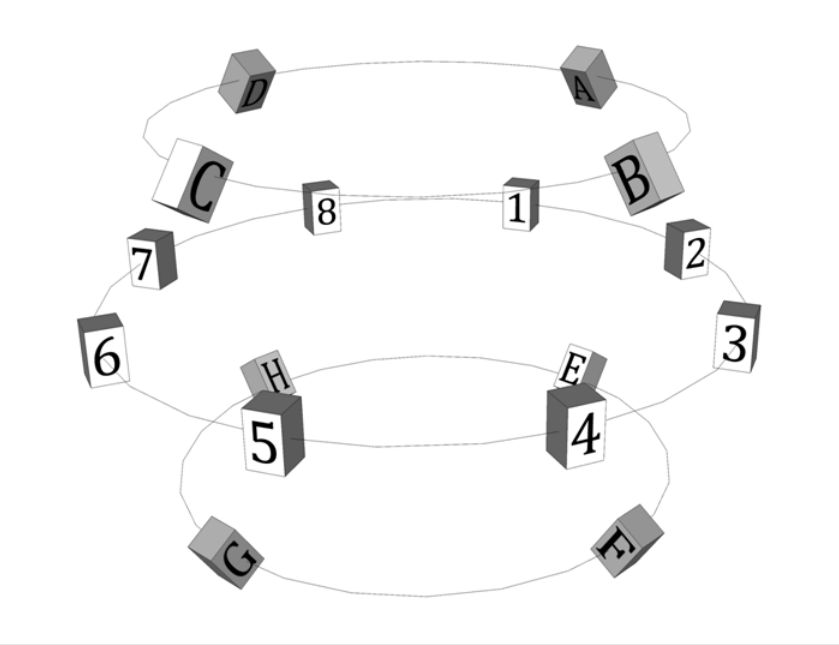


Figure 3.21: *Graphical representation of the position of loudspeakers in the 3 dimensional loudspeaker array*

Figure 3.22: *Photograph of the Virtual Singing Studio as set up in the acoustically treated listening room (KEMAR denotes position singer would take)*

Figure 3.23: *Additional acoustic absorption treatment for ceiling of Virtual Singing Studio*

Orientation of singer

In the *virtual performance space* the vocal signal is convolved with a pre-chosen SRIR; for example, the singer can decide to face into the performance space by using the SRIR measured at a particular performer position oriented forward (labelled 0°). However, if the singer then rotates on the sweet spot in the centre of the loudspeaker array by 180°, the sound field remains static but the singer (and their ears) are facing away from the auralised soundfield meaning that the soundfield is simulated as if the singer were singing

out from the back of the head. However, in practice, having to stand in the sweet spot facing in one direction was not a problem for the singers in the VSS, as in fact, this reflects usual practice for performance.

Acoustic instability

Acoustic instability was avoided in the VSS due to the use of a close head-mounted microphone, as considered in Section 3.4.3, giving a large direct to reverberant sound ratio, meaning that gain before instability was sufficiently high (see Section 2.6.3).

3.4.5 Latency

As discussed in section 2.6.3 any detectable latency in the VSS would impair the plausibility of the simulation and recent research in this area has suggested a perceptual threshold of 59ms for fast moving sound sources in a VAE.

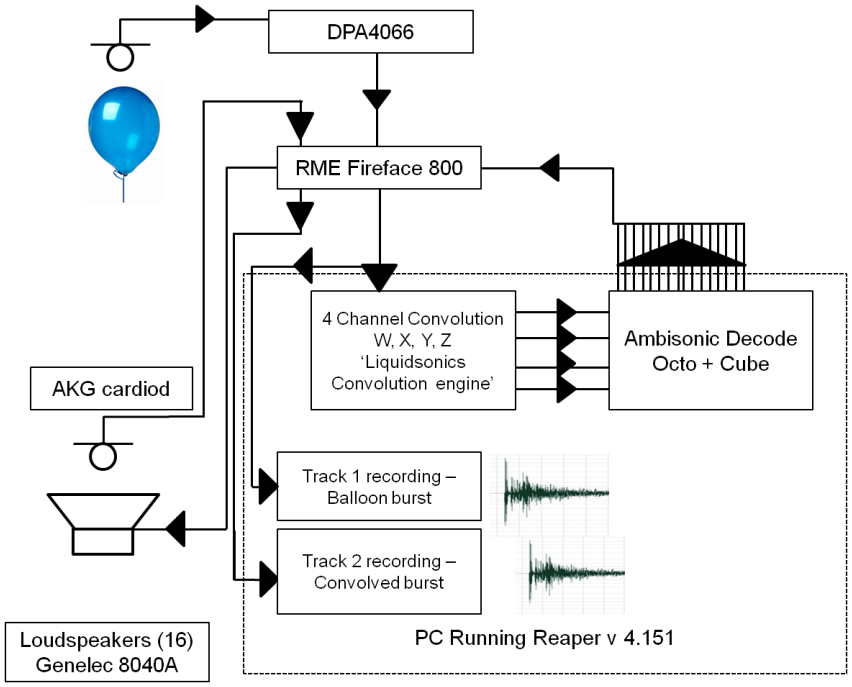


Figure 3.24: *Block diagram of system latency measurement illustrating input microphone (DPA4066), VSS processing (in rectangle with dashed border) running through Reaper DAW, microphone inputs and loudspeaker outputs via RME Fireface800 soundcard.*

In order to evaluate the end-to-end latency of the VSS system the following method was used (Figure 3.24 provides a diagrammatic representation of this method):

An impulse-like signal (balloon pop) was captured at the central point of the loudspeaker array via the DPA4066 microphone used in the VSS and input into the VSS

convolution system (which is described in 3.4.4). A recording of the inital balloon pop was made on a single track (track 1 recording) within the Reaper DAW.

Using the same software and hardware as in the VSS the balloon pop was convolved over 4 channels (Ambisonic channels W,X,Y,Z) with a 4 second long Dirac impulse approximation (comprising one sample of amplitude 1 at time point 1), decoded for Ambisonic presentation and output to the 3 dimensional loudspeaker array. The Dirac impulse used for convolution here allows the processing time of the convolution plug-in (Liquidsonics Reverberate [118]) to be assessed within the overall latency measurement.

Figure 3.25: *Photograph of microphone placement used to capture balloon pop in measuring end-to-end latency of the system*

The convolved balloon pop was additionally captured at one of the head-height loudspeakers by an AKG cardiod microphone placed at a distance of 1cm in front of the loudspeaker (See Figure 3.25). The DPA4066 and cardiod microphone parallel signals were recorded on separate tracks in Reaper and inspected. The time delay between the initial balloon burst (recorded on track 1), and the arrival of the convolved balloon burst (recorded on track 2) was measured as 14ms. The additional time taken for this burst to arrive at the listener central position was calculated to be 5.67ms, giving an overall end-to-end latency of the VSS of 19.67ms. The latency time evaluated in this way is equivalent to the latency which occurs for a singer in the VSS as it includes all the steps

in the processing where latency might arise i.e. convolution, ambisonic decode, output to loudspeaker and input via microphone.

The measured latency in the system can be obviated in the eventual implementation of the VSS by editing the SRIR as is described in the next section.

3.4.6 Editing SRIRs

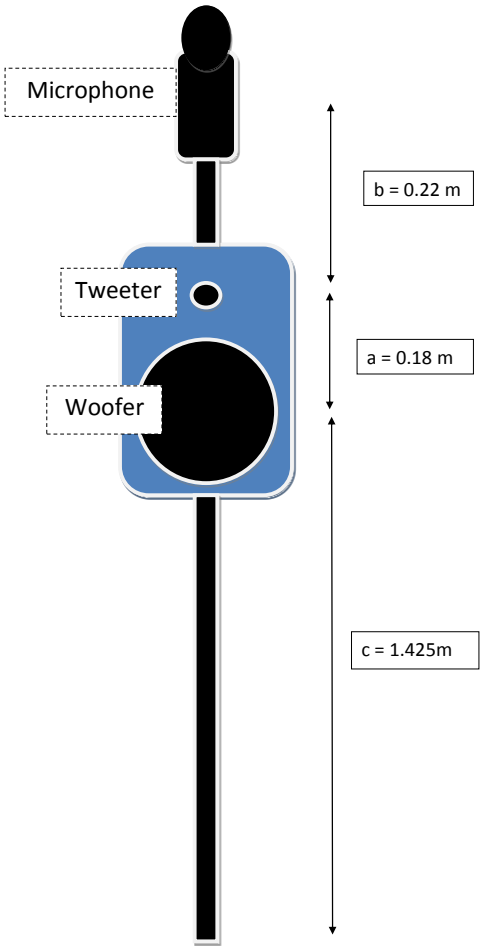


Figure 3.26: *Diagrammatic representation of loudspeaker and microphone topology showing the relative positions and distances between the Soundfield Microphone, and the loudspeaker tweeter and woofer.*

The sine wave sweeps recorded in the space were deconvolved in MATLAB and normalised across each set of 16 impulse responses (four azimuth angles $*$ 4 channels (W, X, Y, Z) for each singer position.

Inspection of the recorded impulse responses in the time domain revealed that the direct sound, and the first reflection (from the floor) consisted of two peaks: one arising from the loudspeaker tweeter and one from the loudspeaker woofer. In the direct sound

the tweeter sound arrives first followed by the woofer sound, whereas the opposite is true for the floor reflection. As can be seen in diagram 3.26 the floor reflection from the tweeter travels a longer path than that of the woofer. The tweeter floor reflection path is 2c+2a+b, whereas the woofer floor reflection path is 2c+a+b and so arrives at the microphone approximately 0.0005 s earlier than the tweeter floor-reflections.

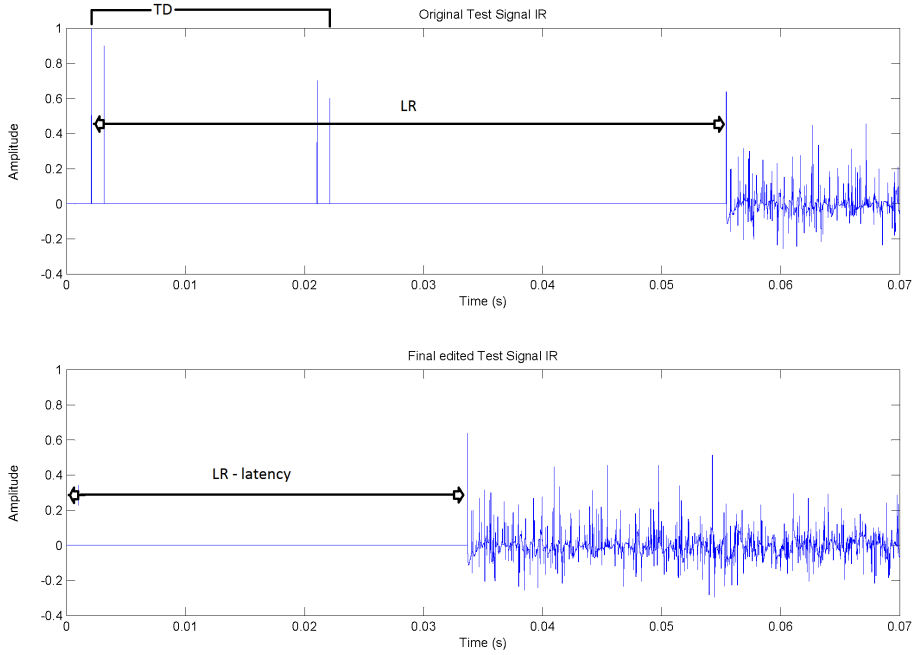


Figure 3.27: *Test signal to illustrate processing of RIRs for use in the VSS*

Since the singer in the VSS provides the direct sound, it is necessary to remove the direct sound from the SRIR and the first (floor) reflection whilst preserving the arrival time of the first lateral reflection.

In order to remove the floor reflection and preserve the correct arrival time of the first lateral reflection the following procedure was carried out in MATLAB (readings taken in the venue at the time of measuring the impulse responses averaged 23° C and 45% humidity, therefore the speed of sound used for these calculations was 345 m/s) :

- The distance to the floor from the tweeter and the woofer is noted for each singer position, as the loudspeaker height was adjusted for each singer position to replicate the singer involved
- The expected arrival times of the tweeter and the woofer floor reflections are calculated from measurements of the loudspeaker/microphone set up

- The expected time delay (marked TD) between tweeter direct sound and woofer floor reflection (thus encompassing the whole floor reflection delay time) is verified by inspecting the time domain representation of the Impulse Response (as marked on Figure 3.27)
- The overall delay between the tweeter direct sound and the first lateral reflection is measured from the waveform (marked LR)
- The initial portion of the impulse response is edited to remove the direct sound and floor reflections
- The impulse response is edited by removing the first 19.67ms (the measured latency of the system) from the beginning of the impulse response. This ensures that the first lateral reflection arrives at the singer's ear after the required delay (LR) and also removes the direct sound and the floor reflections as is required.

Using the editing procedure above, which was verified by producing a test signal of a direct sound, floor reflection and reverberant sound using ideal distances as illustrated in Figure 3.27, ensures that the first lateral reflection arrives at the singer's ears at the "correct" time when convolved through the VSS system.

3.4.7 Calibration

The proper level of the simulated soundfield relative to the direct sound (singer) needs to be determined and replicated in the *virtual performance space* via the VSS.

Method

Following a procedure developed by Laird, Murphy and Chapman [125] calculations of the energy in the early and late parts of the RIR of the VSS were made and compared to those measured in the real performance space. A 15 second long log sine sweep (as used in the original measurements) was output via the Genelec 8040a loudspeaker, captured by the DPA4066 microphone in front of the loudspeaker as pictured in 3.28, convolved with placed at the central point of the loudspeaker array.

The input sine sweep was convolved in the VSS convolution engine (running on Reaper DAW as described in Section 3.4.4) and output to the 16 loudspeakers of the array, convolved via the VSS with a SRIR measured in performer position A (See Section 3.4.1)) and subsequently captured by the Soundfield microphone placed above the loudspeaker (in a similar topology to the original measurements).

Figure 3.28: *Arrangement of source loudspeaker (Genelec 8040) with DPA4066 microphone in front and Soundfield microphone above as used in the calibration experiment*

The sound source level was measured at 10cm from the loudspeaker as 88.5 dBA. All levels on the external RME Fireface800 soundcard and Reaper DAW were noted and maintained whilst the procedure was repeated, changing only the output level of the convolution channel BUS to the Ambisonic decoder. Four different level settings (-43dB, -36dB, -29dB and -22dB) were tested in this way, and the output sine sweep captured by the Soundfield microphone was deconvolved to provide an impulse response of the *virtual performance space*.

Measures of *Support* namely ST_{early} and ST_{second} were calculated for the *real performance space* at the five different gain settings of the *virtual performance space*. ST_{second} is a new parameter, based on Support measures as outlined in Section 2.4.3 but evaluating the balance of the later arriving energy to the direct sound.

$$ST_{late} = 10log\left\{\frac{E_{100-1000ms}}{E_{0-10ms}}\right\}, dB \tag{3.1}$$

$$ST_{second} = 10log\left\{\frac{E_{1000-2000ms}}{E_{0-10ms}}\right\}, dB \tag{3.2}$$

Results

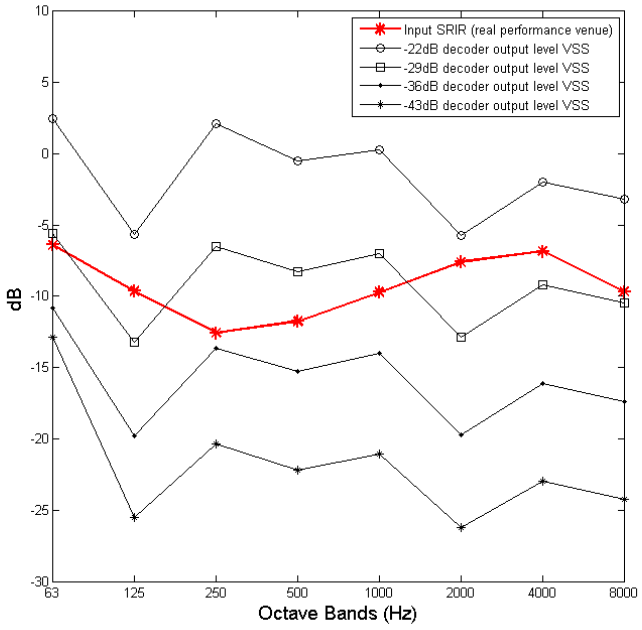


Figure 3.29: *Comparison of ST_{late} values of the real performance space and virtual performance space at different decoder output levels.*

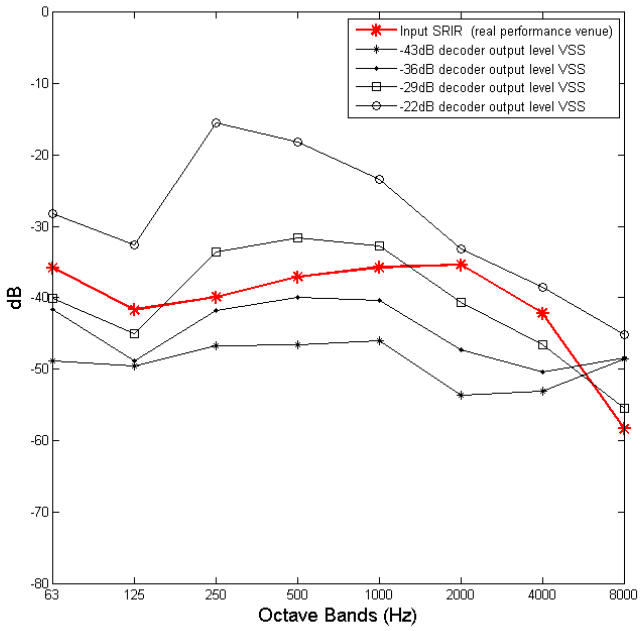


Figure 3.30: *Comparison of ST_{second} values of the real performance space and virtual performance space at different decoder output levels.*

Discussion

The ambisonic decoder output level of -29dB was chosen as the best fit of the levels tested. However, when the settings were replicated and a singer, wearing the head-mounted microphone used the VSS set up in this way the level appeared to be too high for the singer to accept as realistic or plausible. A singer who was familiar with the real performance space, and had performed there on several occasions was able to advise on the adjustment of the loudspeaker outputs to a plausible level. The level was adjusted until the singer was happy that the level replicated what she felt to be natural for the *real performance space*. This new setting for the convolution channel BUS of -45dB was furthermore maintained for all singers who used the VSS and care was taken to ensure that the head-mounted microphone was the same distance from the mouth (5 cm) for all singers

The calibration method used here suffered from a loudspeaker and microphone topology which do not sufficiently replicate the levels and directivity of the singer source signal. A more robust method of calibrating the VSS should be a priority for further development, and will involve the use of a head and torso simulator.

3.5 Verification of The Virtual Performance Space

An assessment of the perceptual relevance of the errors found is based on a method used by Favrot [95, 6, 5] which compares potential errors introduced by the room acoustics simulation system to single and double tolerance subjective limen. These subjective limen are taken as 1 and 2 times respectively the quoted "just noticeable differences" (JNDs) for each parameter [38, 126, 74].

3.5.1 Method

The objective evaluation of the VSS was carried out in order to assess any errors which were introduced due to the signal processing involved, playback methods and spatial properties of the listening room in which the VSS is located.

Using a method developed by Farina [25] the SRIR of the *virtual performance space* was captured and evaluated in the same way as the SRIRs in the *real performance space*.

A 15 second log sine sweep was convolved over four channels (W,X,Y,Z) with the SRIR measured in the *real performance space* (input SRIR), decoded with the same ambisonic decoder and output to the loudspeaker array. This output sine sweep was captured via a Soundfield SPS422B microphone at the central point of the array. "Output" SRIRs were then obtained via deconvolution implemented in MATLAB as described in Section 3.4.1.

Room Acoustic Parameters

Room acoustic parameters were evaluated across seven octave bands ranging from 125Hz to 8 kHz as outlined in the Table 3.3 with subjective limen (JND) values as stated in [126]. It should be noted that there are no published subjective limen for RR_{160} and measures of stage support so a JND of 1dB has been assumed, in line with other parameters which evaluate the relative energy of direct to late sound.

Parameter Name	Definition	Subjective Limen
EDT	energy drop from 0 to -10 dB.	5 %
Reverberation Time (T30)	energy drop from -5 to -35 dB.	5 %
Early Support (ST_{early})	ratio of direct sound (0-10 ms) to early arriving reflected energy (20-100 ms)	1 dB
Late Support (ST_{late})	ratio of direct sound (0-10 ms) to late arriving reflected energy (100-1000 ms)	1 dB
Total Support(ST_{total})	ratio of direct sound (0-10 ms) to total energy (20-1000 ms)	1 dB
Running Reverberation (RR_{160})	ratio of early energy (0-160 ms) to later energy in early part of sound (160- 320 ms)	1 dB

Table 3.3: *List of room acoustic parameters evaluated in the real performance space and the simulation (VSS values from [126])*

Room acoustic parameters were calculated using AcMus MATLAB toolbox developed by Masiero et al. [127] and available at [128]. Additional scripts were written to work with this toolbox to evaluate levels of *Support* and RR160 as described in Section 2.4.3 . Note that measures of *Support* evaluated here are not fully comparable to those used by other researchers as originally proposed by Gade [2] since the standard Support measurements specify that the microphone should placed at 1m from the source loudspeaker to replicate the topology of player and instrument [49, 39].

The room acoustic parameters listed in Table 3.3 as evaluated in the *real* and *virtual* performance spaces can be found in Appendices D and E, and are presented in graphs in Sections 3.5.2 and 3.5.3.

Section 3.5.2 presents a comparison of T30, ST_{early} and ST_{late} between the different acoustic configurations of the real space, to give a sense of the differences in settings. Section 3.5.4 compares the *real* and *virtual* performance spaces by plotting graphs of the errors arising from the room acoustic simulation.

3.5.2 Room acoustic parameters of Real Performance Space

Large Choral Setting (LC)

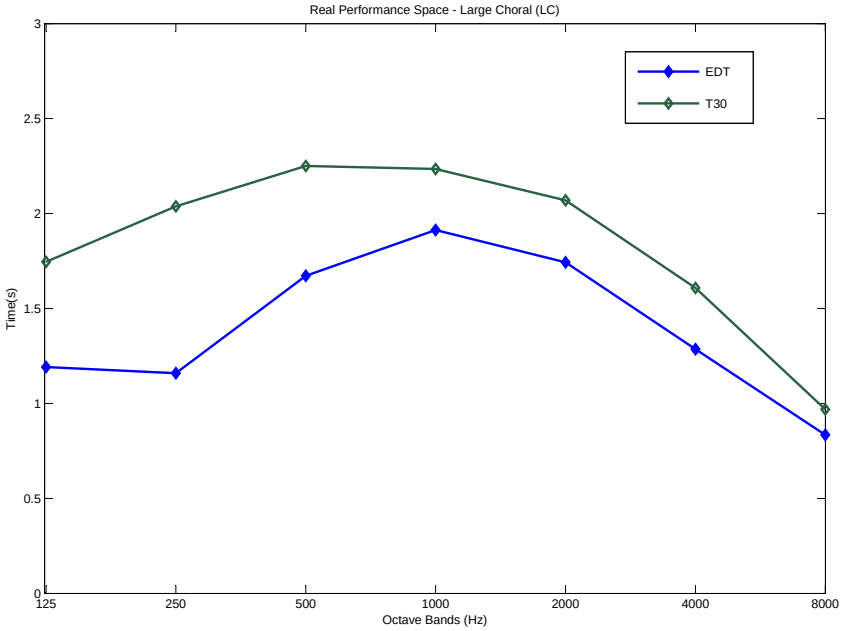


Figure 3.31: *Mean T30 and EDT values evaluated in the Large Choral setting of the Real Performance Space as as measured in the four performer positions*

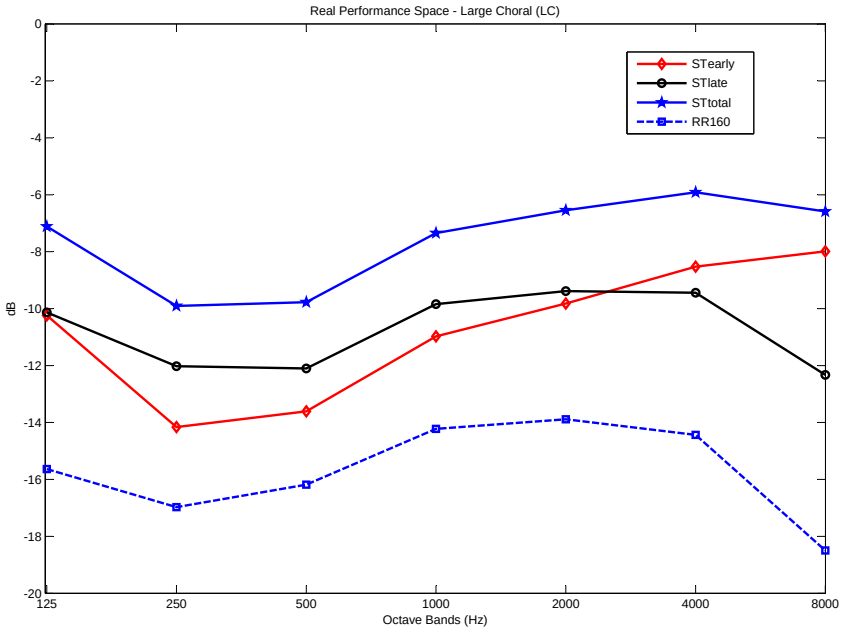


Figure 3.32: *Mean of Support and RR160 values of the four performer positions evaluated in the Large Choral (LC) setting of the Real Performance Space*

Music Recital Setting (MR)

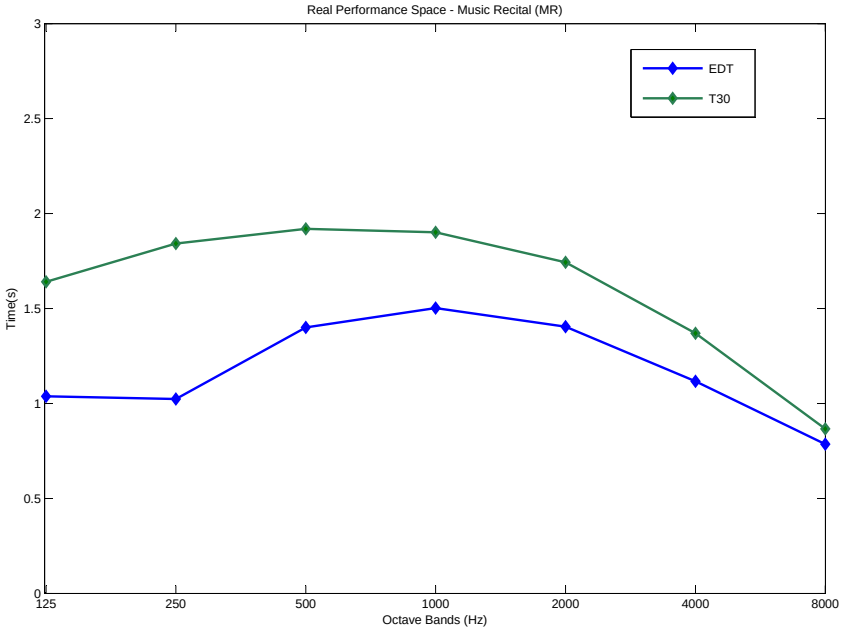


Figure 3.33: *Mean T30 and EDT values of the four performer positions evaluated in the Music Recital (MR) setting of the Real Performance Space*

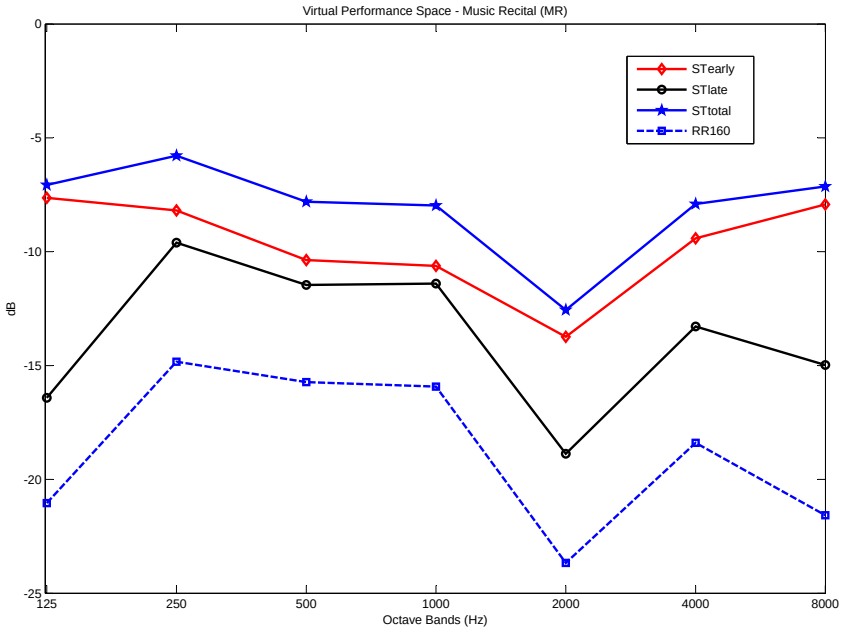


Figure 3.34: *Mean Support and RR160values of the four performer positions evaluated in the Music Recital setting of the Real Performance Space*

Speech Setting (SP)

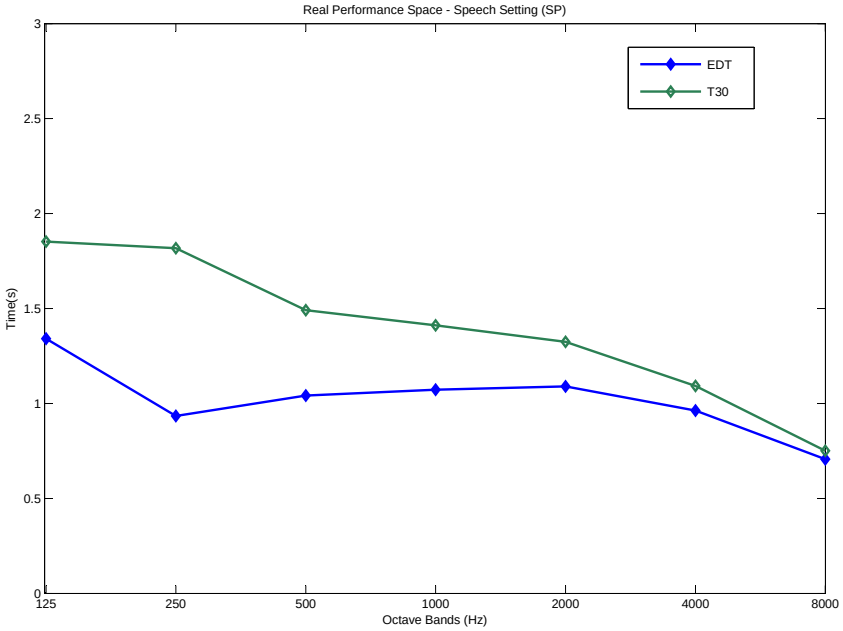


Figure 3.35: *Mean T30 and EDT values of the four performer positions evaluated in the Speech (SP) setting of the Real Performance Space*

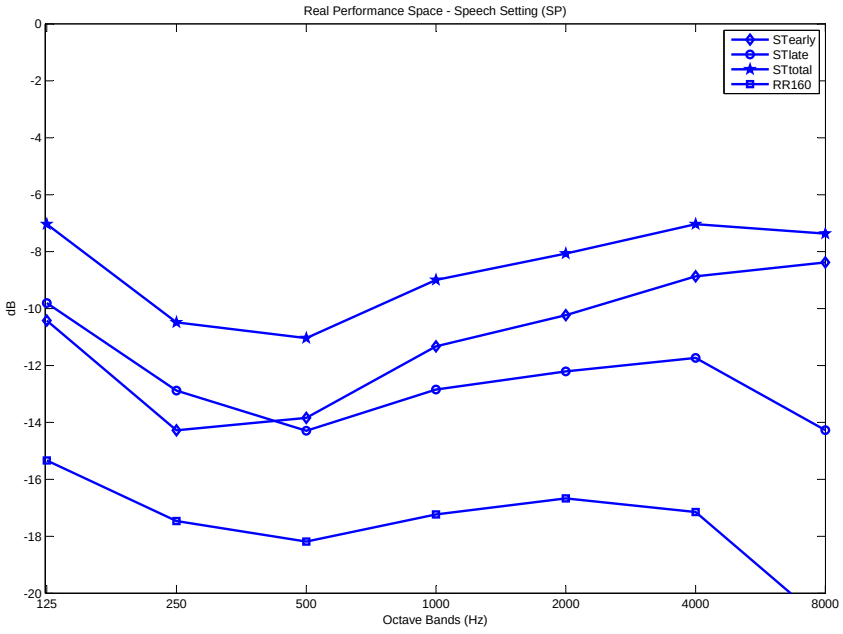


Figure 3.36: *Mean Support and RR160values of the four performer positions evaluated in the Speech (SP) setting of the Real Performance Space*

Comparison of Acoustic Configurations of the Real Performance Space

To give an indication of the differences between the three acoustic configurations chosen for use in this research, Figures 3.37 to 3.39 plot mean T30, ST_{early} and ST_{late} values for the performer positions in the Large Choral (LC), Music Recital (MR) and Speech (SP) settings.

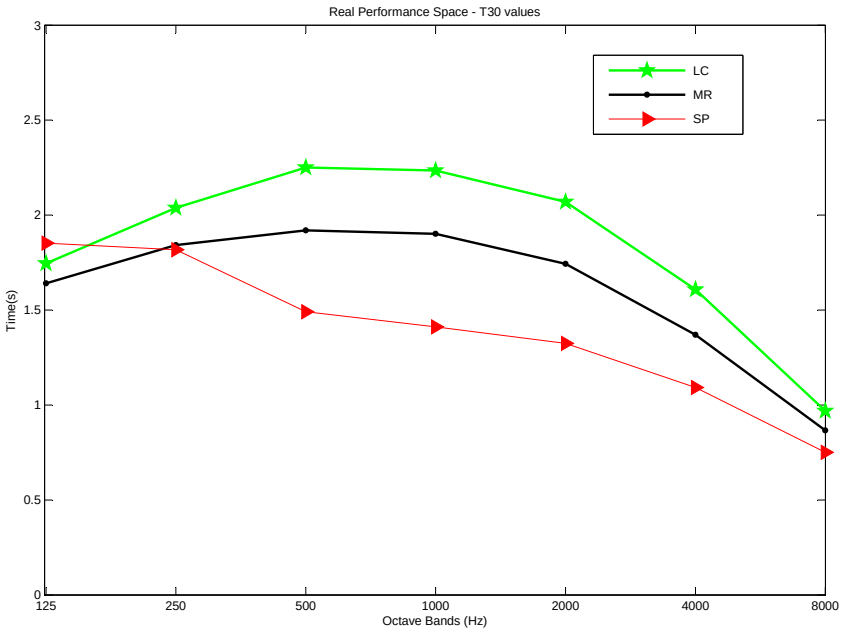


Figure 3.37: *Mean T30 values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP)*

Discussion

As is common in a mid-sized church building ([129, 32]), T30 values rise with frequency up to a peak in the 500Hz octave band (see Figures 3.37 and 3.40). In both the listener position and performer position T30 values decrease with frequency in the octave bands above 500 Hz In the Speech Setting EDT values peak at the 125Hz octave band which could be due to the location of a nearby column and early reflection from the back wall of the stage area.

Gade [2] found that "good" concert halls (those rated highly by performers) have ST_{late} values 1 to 3 dB higher than ST_{early}, which does seem to be the case in this performance venue (see for example Figure 3.32).

In larger performance spaces such as the 30,000-seat concert hall measured by Kim et al. [49], ST_{early} values between 250hz and 4kHz can be as low as -19.9dB to -11.3dB and ST_{late} of between -18.8dB and -15.4dB. It should be noted that ST values can vary

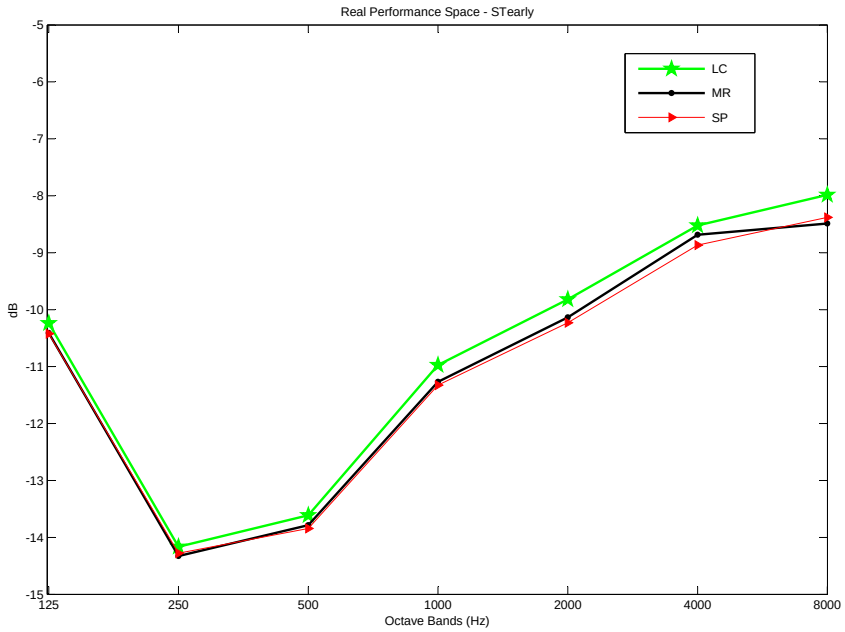


Figure 3.38: *Mean Early Support values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP)*

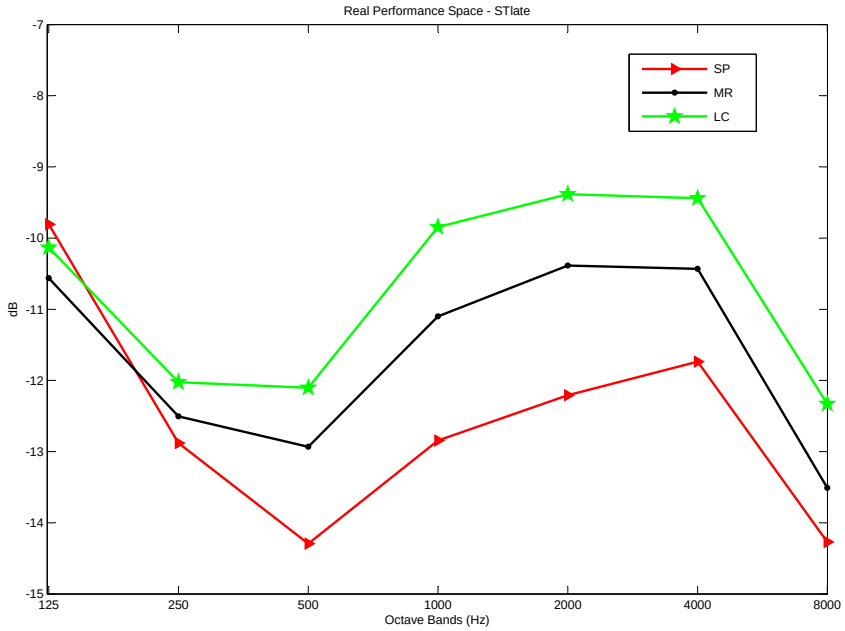


Figure 3.39: *Mean Late Support values of the four performer positions evaluated over seven octave bands in the three acoustic configurations : Large Choral (LC), Music Recital (MR) and Speech (SP)*

greatly (in the region of ± 10 dB) at different positions on the stage. Nevertheless stage Support measures here give some sense of the levels of *Support* the singer might expect in this venue.

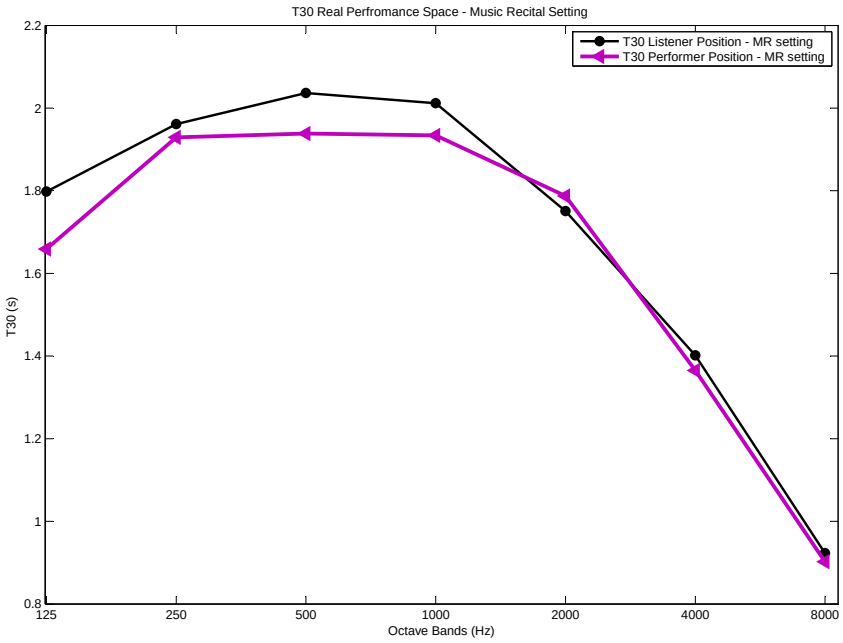


Figure 3.40: *Comparison of T30 values measured at Listener Position and Performer Position B across seven octave bands in Music Rectial Setting*

3.5.3 Room acoustic parameters of Virtual Performance Space

Discussion

In all three acoustic configurations values of T30 and EDT differ from the real space in the lower (125Hz and 250 Hz) octave bands and at the 2000 Hz octave band. Support measures and RR160 also do not match well at these octave bands. The next section presents these differences as errors produced by the implementation of the VSS and makes an assessment of their perceptual relevance.

3.5.4 Comparison of Real and Virtual Performance Space

Since EDT and T30 are measured in seconds the relative errors (%) are calculated as the difference between the output and input SRIRs (virtual - real). Errors in Support and RR160 are presented as absolute errors, again calculated as the difference between the *virtual* and *real performance space*. The single and double subjective limens for each parameter (one and two times the JND) are also plotted, as in Favrot [5], as an indication of the perceptual relevance of the errors.

Large Choral (LC)

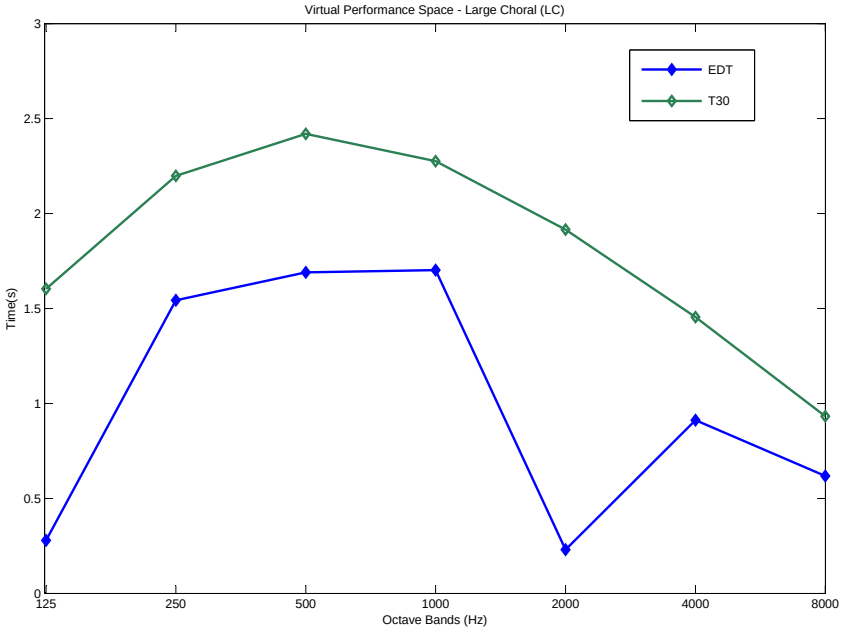


Figure 3.41: *Mean EDT and T30 values evaluated in the Large Choral setting of the Virtual Performance Space as simulated for the four performer positions*

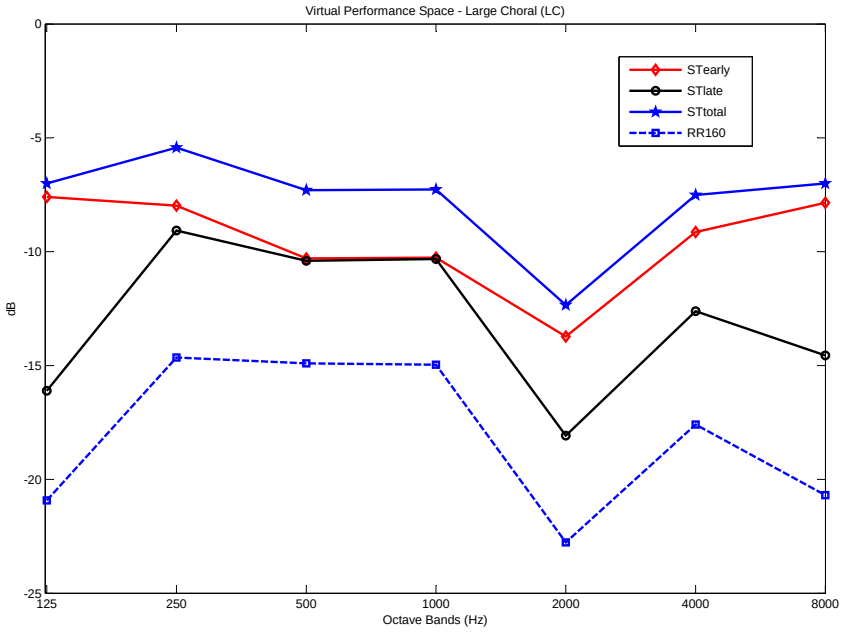


Figure 3.42: *Mean Support and RR160 values evaluated in the Large Choral setting of the Virtual Performance Space as simulated for the four performer positions*

Music Recital Setting (MR)

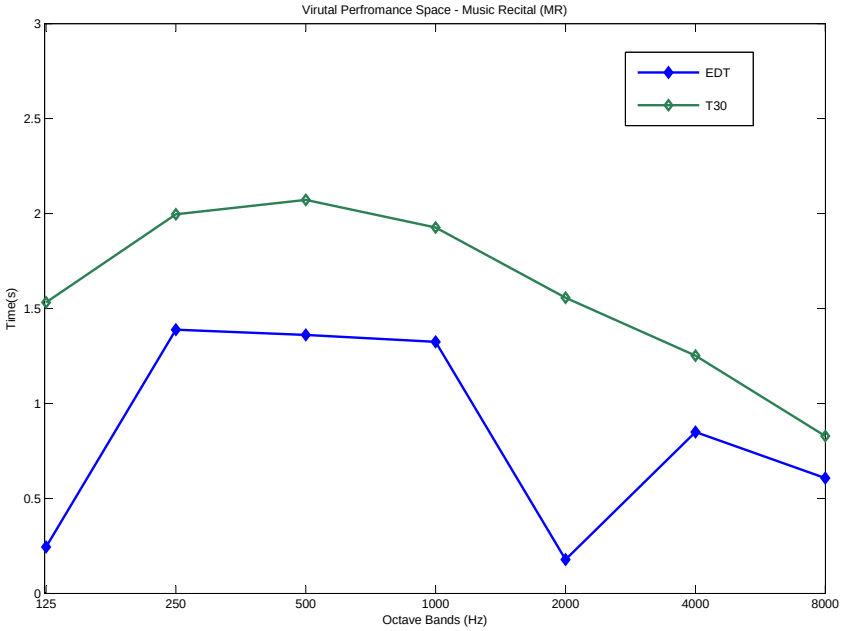


Figure 3.43: *Mean EDT and T30 values of the four performer positions evaluated in the Music Recital (MR)setting of the Virtual Performance Space*

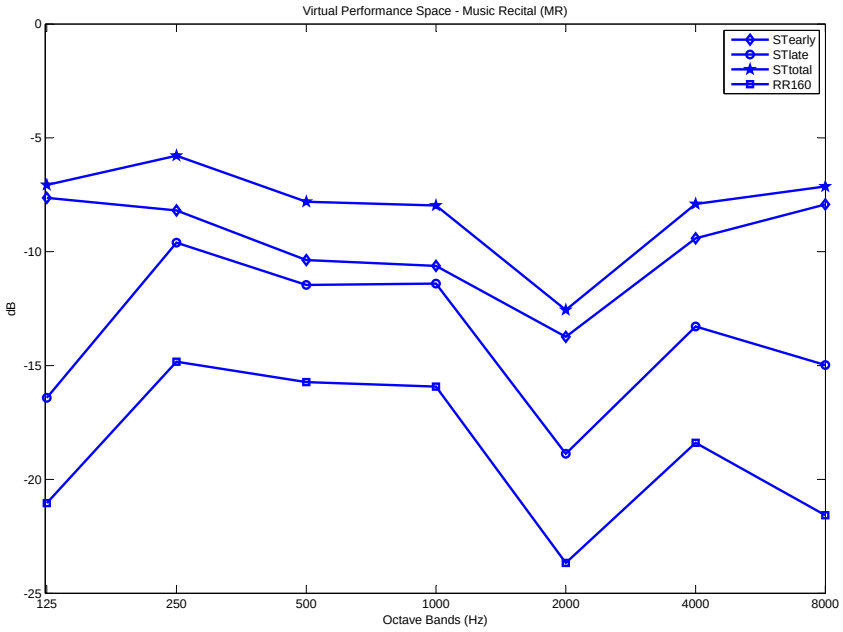


Figure 3.44: *Mean Support and RR160 values of the four performer positions evaluated in the Music Recital (MR) setting of the Virtual Performance Space*

Speech Setting (SP)

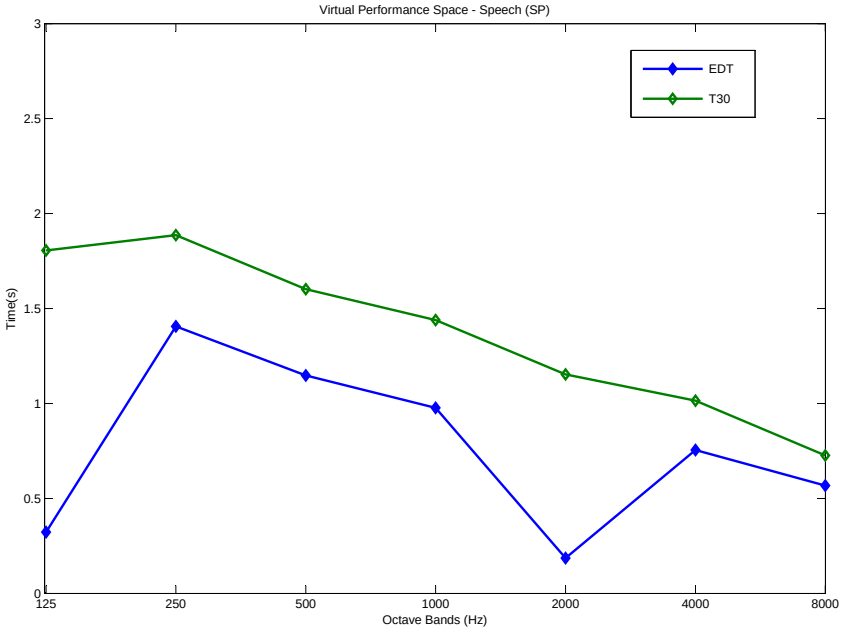


Figure 3.45: *Mean EDT and T30 values of the four performer positions evaluated in the Speech (SP) setting of the Virtual Performance Space*

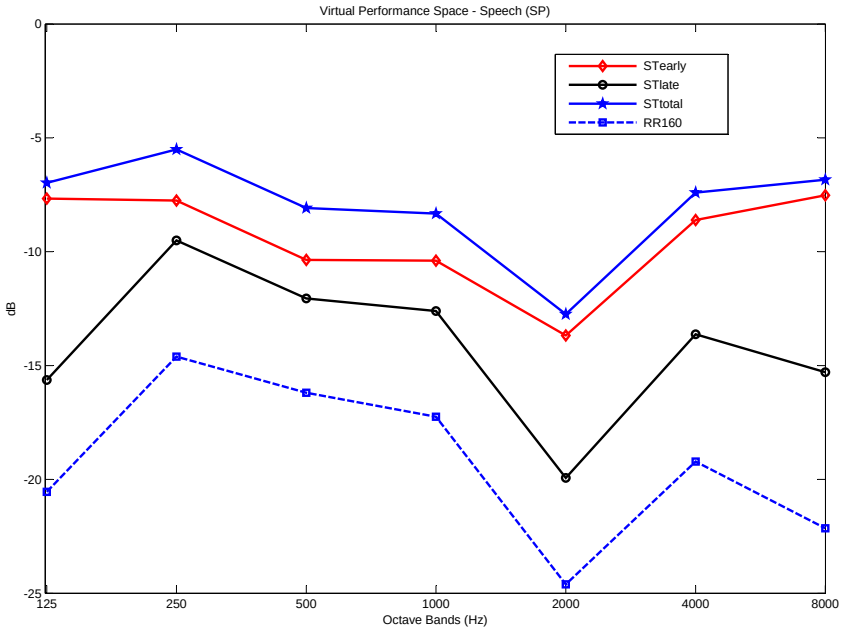


Figure 3.46: *Mean Support and RR160 values of the four performer positions evaluated in the Speech (SP) setting of the Virtual Performance Space*

Errors in time-based parameters

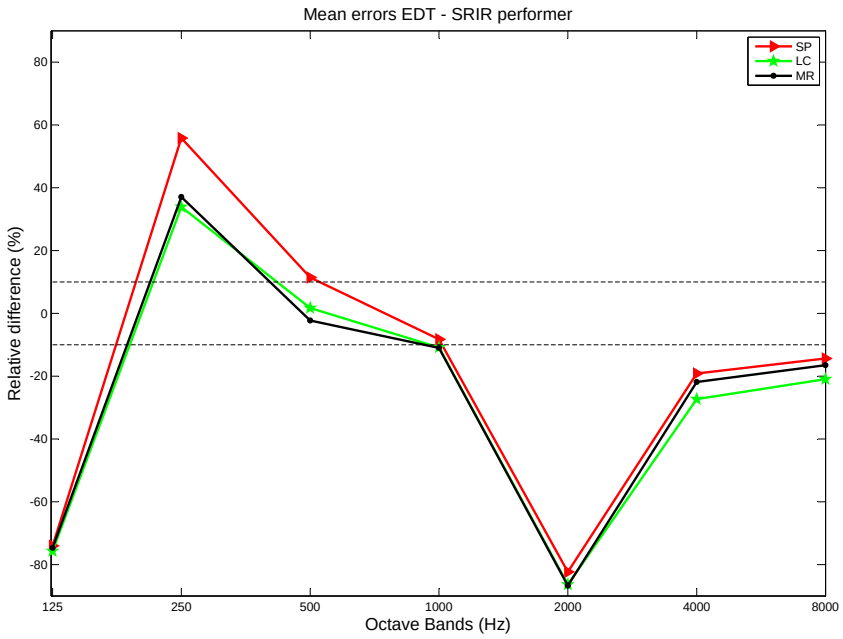


Figure 3.47: *Differences between virtual and real performance space EDT values in performer position, for octave bands 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

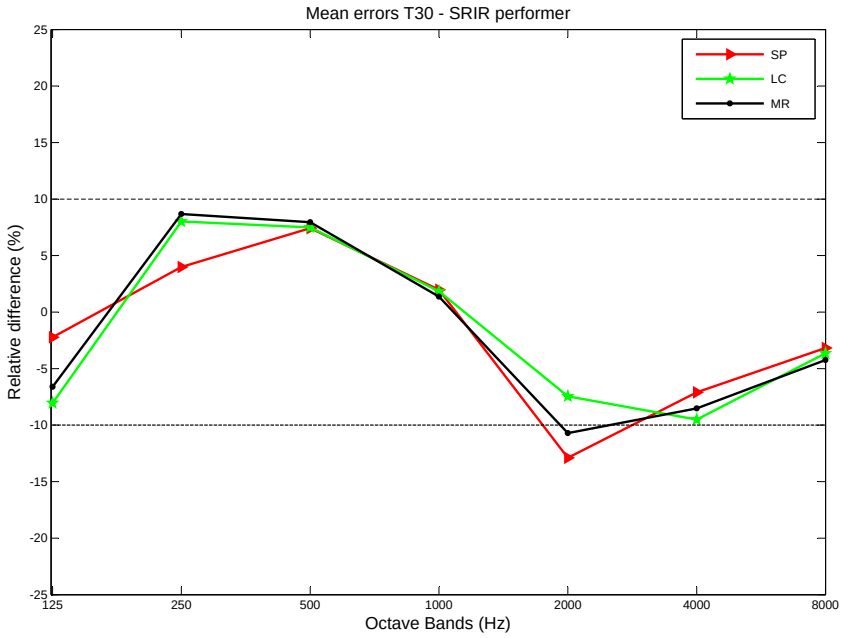


Figure 3.48: *Differences between virtual and real performance space T30 values in performer position, for octave bands 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

Errors in energy-based parameters

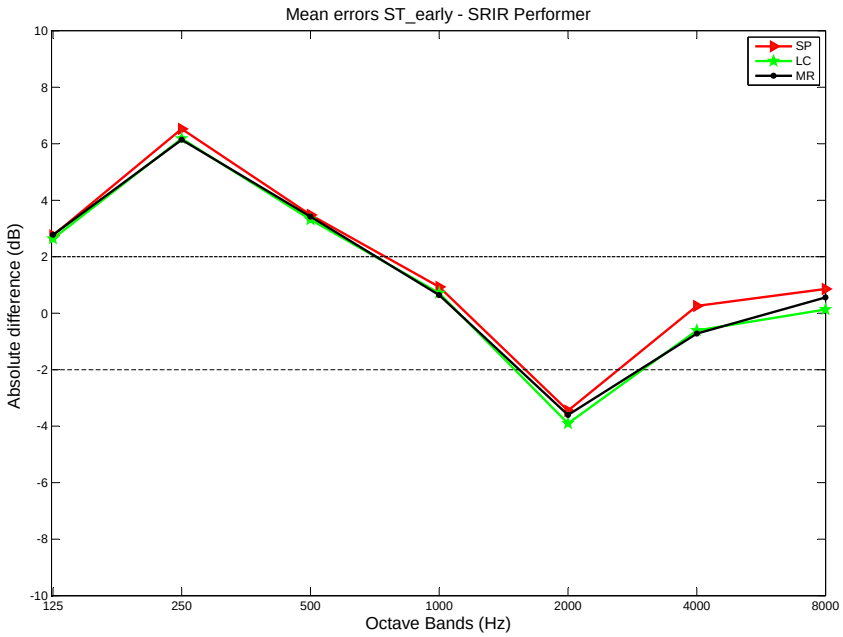


Figure 3.49: *Differences between virtual and real performance space Early Support values in performer position, for octave bands between 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

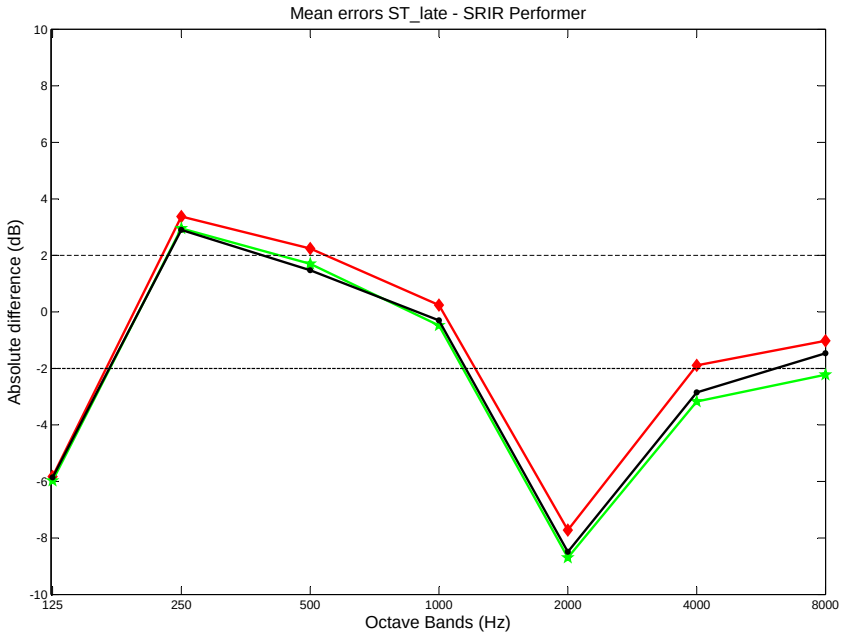


Figure 3.50: *Differences between virtual and real performance space Late Support values in performer position, for octave bands 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

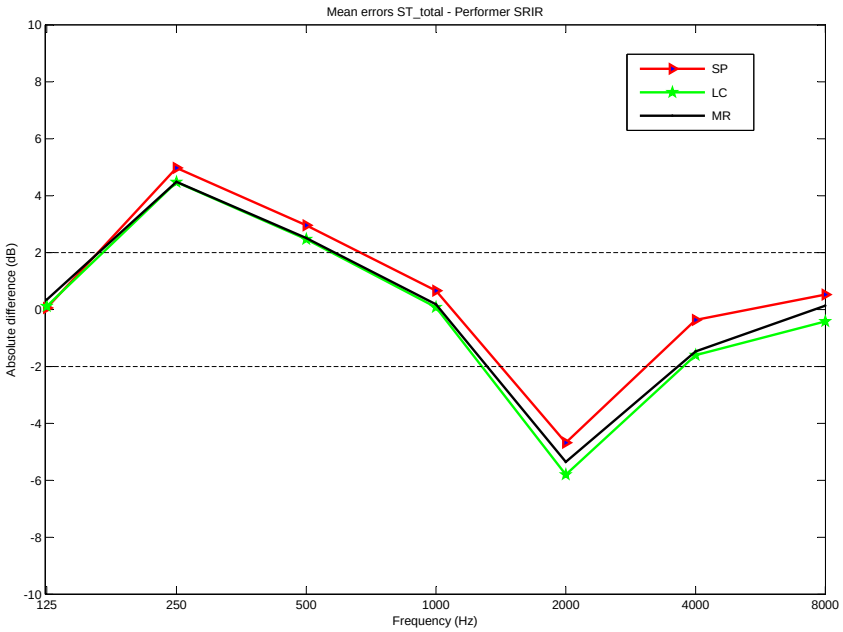


Figure 3.51: *Differences between virtual and real performance space Total Support in the performer position, for octave bands 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

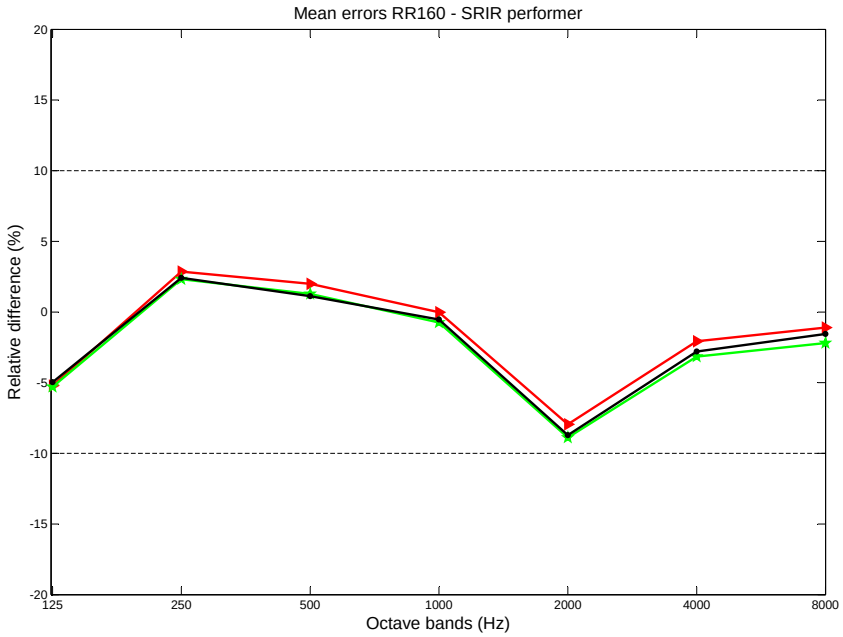


Figure 3.52: *Differences between virtual and real performance space RR160 values in the performer position, for octave bands from 125hz - 8000Hz. Dashed lines indicate double subjective limen.*

Discussion

EDT values are not well matched between input and output RIRs except at the 500Hz and 1000Hz octave bands (see Figure 3.47). In the 250 Hz octave band EDT is longer in

the VSS than in the real space, whereas above the 1000 Hz octave bands EDT is shorter.

These errors lie outside of the double subjective limen and are greatest for the "Speech" setting. As EDT calculations rely heavily on the direct sound and early part of the RIR, the positive errors in EDT under 500Hz are most probably a symptom of room modes and an early reflection (ceiling) in the listening room.

The "Speech" setting SRIR is the least well matched to the input RIR at 500Hz and below (See Figures 3.47 and 3.48). This is probably because the shorter reverberation time in this setting mean that the rate of decay is more strongly influenced by the presence of early reflections in the listening room itself.

Most of the T30 errors lie within the double tolerance subjective limen (10% - dashed line) with the greatest errors occurring in the 2000Hz octave bands in the "Speech" (SP) setting (see Figure 3.48). The "Large Choral" and "Music Recital" settings are most closely matched to the input RIR in terms of T30 with error values lying within the single tolerance subjective limen (5%) at all but 500Hz and 2000Hz octave bands.

RR160 (running reverberation) has been suggested by Griesinger [55] as a measure of reverberation the musician perceives whilst playing music (see Section 2.4.3) and as an alternative way to evaluate musician self-support. Errors in RR160 all lie within the double subjective limen (10%) indicating that in at least one measure of support for the musician the virtual and *real performance space* are well matched (see Figure 3.52).

ST_{early} errors lie within the double subjective limen only at the 1000 Hz, 4000 and 8000 Hz octave bands (See Figure 3.49). ST_{late} similarly are mostly outside of the double subjective limen, however this parameter has a smaller error than the ST_{early} in the 250 and 500 Hz octave bands, whereas the error is greater in the 2000 Hz octave band, suggesting discrepancies in the early part of the impulse response at the lower octave bands and, in contrast, in the later part of the impulse response in the upper part of the frequency spectrum.

There does seem to be a clear difference between the *real* and *virtual performance space* parameters in the 2000 Hz octave band. In order to investigate this further a spectrogram of the input (real) and output (virtual) impulse response (256 sample Hanning window, 25 % window overlap, 256 frequency bins, sampling rate 48 kHz) is plotted in Figure 3.53.

It can be seen that reverberant energy appears to decay more quickly in the 2000Hz octave band (indicated by the black arrow). The lower level of energy in this frequency band would explain the lower ST_{late} and ST_{total} values in the *virtual performance space*.

It is more difficult to explain the positive errors in the 250 Hz and 500 Hz octave bands. It could be the case that the reference level (the total energy in the first 10ms of the impulse response) is lower in the *virtual space* in comparison to the real, leading to an overall higher ST_{early} value in the virtual. Gade's original Support calculations [39]

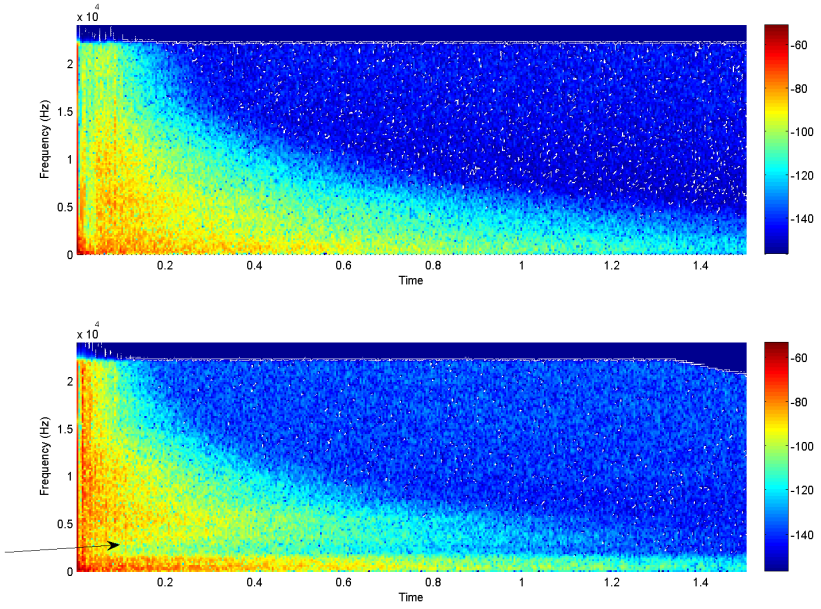


Figure 3.53: *Spectrograms of impulse responses measured in the real performance space (upper panel) and virtual performance space (lower panel)*

measure Support with the receiver at a distance of 1m from the source, to reflect the topology of musician and instrument, and the initial 0 - 10ms window is chosen to contain the direct sound of the instrument, but not the first early (floor) reflection. However, in the measurement of Support used here, where the source loudspeaker was placed directly below the receiver (sound field microphone) the floor reflection is contained in this 0 - 10 ms window. The *real performance space* has a stone floor, whereas the *virtual performance space's* floor is carpeted, suggesting that the floor reflection in the simulation might be at a lower level than the real space, leading to higher ST_{early} values.

An additional confounding factor in achieving suitable levels of *Support* for the musician, is probably the output level chosen for the implementation of the simulation system, which, due to difficulties with objective calibration, was set heuristically after feedback from one singer. As was noted earlier a more robust calibration method for the VSS is a priority for future development of the VSS and recent work (e.g. [54] and [53]) will help inform this improvement.

In the *virtual performance space* it is not possible to simulate any early reflections to arrive earlier than the time delay resulting from the distance from the loudspeaker to the singer/listener. In this case the loudspeaker-to-performer distance is 1.95 m meaning that no reflections can be simulated to arrive earlier than 5 ms. Happily, in the *real performance space* the first lateral reflection arrives, for example in tenor position C (see Figure 3.11), at 8.6 ms, after reflection off the back wall, which is at a distance of 1.48m

from the singer.

Whilst JNDs provide a useful benchmark for potential errors there is some debate as to how relevant they are to the listener. Since JNDs are established using carefully controlled laboratory experiments Bradley [42] argues that in the perception of bass reverberation in concert halls, for example, "Differences that can be detected by listeners in actual halls are probably much larger. This is partly because more than one aspect of the sound field will change when the listener moves in a concert hall, making it more difficult to identify the individual effects of each changing parameter". The increase in JNDs for the performing musician is probably similar, if not greater, and especially for the singer who is at once listening whilst vocalising.

3.5.5 Summary

All T30 errors lie within the double tolerance subjective limen (10% - dashed line) and all except the 250 Hz band are within the single tolerance subjective limen (5 %). EDT values are not as well matched, due to the acoustic presence of the listening room in which the VSS is housed.

In general the *virtual performance space* does not mimic the *real performance space* well in the lower octave bands, which is most probably due to listening room reflections or possible room modes at lower frequencies. In all parameters there is a difference in the 2000 Hz octave band (indicated by the arrow in Figure 3.53), due to a faster decay of energy in this frequency region, as can be seen in the spectrogram of the input and output SRIRs in Figure 3.53. The causes for this discrepancy in this region and their perceptual relevance need to be further investigated.

Measures of support evaluate the balance of cumulative energy, that is they do not indicate the arrival time nor direction of early and late reflections, so in practice two perceptually very different concert hall stages might have similar ST_{total} values, for example, but be highly dissimilar in terms of the performer's experience of the sound field on stage.

Further work might also seek to understand and verify the arrival time and direction of early and late reflections, which could be carried out by a SIRR analysis of the impulse responses [87] (see Section 2.5.5).

The objective evaluation of the VSS has shown that the performance space is simulated within subjective limen for T30 and RR160 values and that larger errors occur in the 250, 500 and 2000 Hz octave bands. The next section presents subjective evaluations of the VSS from a number of professional singers who were recorded in both the *real* and *virtual performance spaces*.

3.6 Singers' Evaluation of the VSS

3.6.1 Method

A number of professional singers were asked to sing in the *virtual* and *real performance spaces* and their singing was recorded via the head mounted microphones used in the VSS as described in 3.4.3.

Seven professional singers took part in the recordings (1 soprano, 1 mezzo-soprano, 2 altos, 1 tenor and 2 basses) ranging from 24 to 35 years old (Average age 30, SD 3.25 years). All were experienced singers with 6-27 years (Mean 19.43, SD 7.89) years of musical training and 6-19 (mean 12.14, SD 4.3) years of specifically vocal training. All singers sang regularly in professional solo voice ensembles, and most (5) also performed regularly as soloists in oratorios and recitals. One singer (soprano) also worked as a member of a professional opera chorus. All singers cited "early music" (music written prior to 1750) as a specialism; one singer (alto) also had expertise in extended vocal techniques.

After the singer had performed in each acoustic setting in the virtual and real spaces a questionnaire was completed using an on-line interface; informal interview was also carried out by the author to gain further insight into the singer's experience of the performance spaces. The questionnaire is available on-line at `http://tinyurl.com/o32h9c7`.

3.6.2 Results of questionnaire

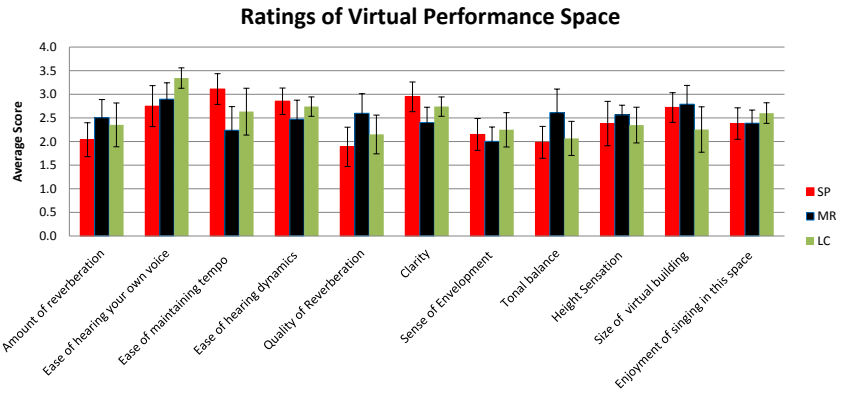


Figure 3.54: *Mean scores (and standard error) of singers' responses to questionnaire on the Virtual Singing Studio*

In general, for the *virtual space* (3.54), in the Large Choral setting (LC) "ease of hearing your own voice" and "enjoyment of singing in this space" was rated most highly. "Amount and quality of reverberation", "tonal balance" were rated most highly in the Music Recital setting in the virtual space.

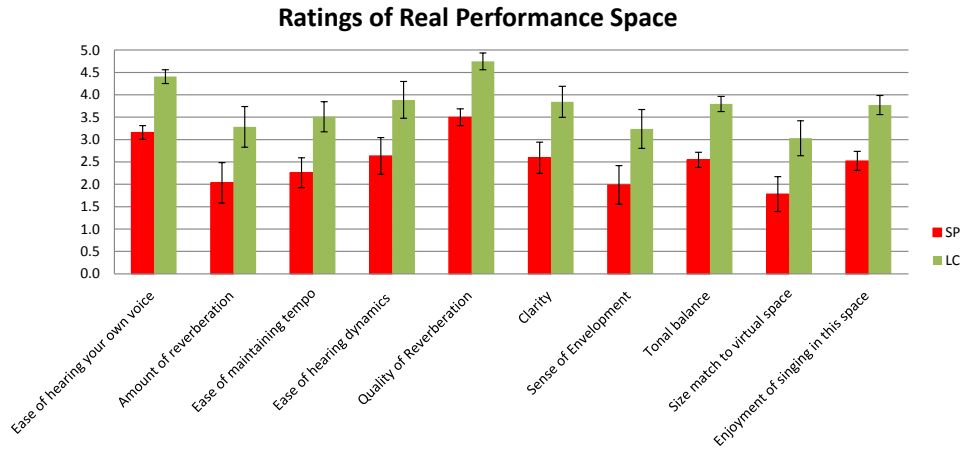


Figure 3.55: *Mean scores (and standard error) of singers' responses to questionnaire on the Real Performance Space*

For the *real performance space* (3.55) on the other hand, ratings of the SP and LC settings (the two settings which were evaluated in the real space) are much more distinct. The LC setting was rated more highly for all questions.

In the *virtual performance space* both sopranos and the bass preferred the Speech setting (SP), whereas the altos and tenor preferred the Large Choral (LC) setting. [NEW] In the *real performance space* there was no discernible pattern of preference for the different configurations between the different voice types.

[NEW]In order to examine the effect of simulation and acoustic setting on the experience of the solo singers who performed in the *virtual* and *real spaces*, a number of two-way ANOVA were run with *simulation* (real vs virtual) and *acoustic* (Large Choral vs Speech setting) as the independent variables, and singers' rating as the dependent variable. A separate ANOVA was conducted for each of the questions which were common to the questionnaires for the real and virtual spaces namely ratings of: "amount of reverberation", "ease of hearing own voice", "ease of maintaining tempo", "ease of hearing dynamics", "quality of reverberation", "clarity", "sense of envelopment", "tonal balance" and "enjoyment of singing in the space".

There was a significant effect of *acoustic* on the perceived "amount of reverberation" ($F_{(1,6)}$=4.25, $p < 0.05$) , ease of maintaining tempo ($F_{(1,6)}$=4.35, p ¡0.01) , and the perception of "clarity" ($F_{(1,6)}$=5.02 p <0.05). There was also a significant effect of *acoustic* ($F_{(1,6)}$ = 7.19, p <0.05) and of *acoustic*simulation* ($F_{(1,6)}$ = 5.69, p <0.05) on the "sense of envelopment".

There were also significant effects of *simulation* on the ease of maintaining tempo ($F_{(1,6)}$ =8.45, p <0.01) and on perception of "clarity" ($F_{(1,6)}$=4.85 p <0.05). The effect of *simulation* on perceived quality of reverberation ($F_{(1,6)}$=14.15, p<0.001) was the strongest

effect. There were no significant effects of either *acoustic* or *simulation* on the ratings of enjoyment of singing in this space .

3.6.3 Discussion of results

Singers comments about their experience of singing in the real and virtual spaces can be found in Appendix F. Singers mentioned *support*, *reverberation* and *hearing oneself* as being important aspects of room acoustic conditions for the performing musician. These perceptual qualities relate to the objective parameters of ST_{early}, ST_{late}, T30 and RR160.

[NEW] Singers' ratings of perceptual qualities vary between the different acoustic configurations as expected given the values of the corresponding objective parameters, both in the real and virtual spaces. For example singers' ratings of "amount of reverberation" and "clarity" were expected to vary between the acoustic settings, since rooms with longer reverberation times generally have lower levels of clarity, and indeed the effect of acoustic on these ratings is significant.

It is interesting that *acoustic* also has a significant effect on the "ease of maintaining tempo"; most probably this relates to the perception of the amount of reverberation and is an aspect that three of the singers remark upon as differing between the acoustic settings.

Singers' sense of "envelopment" also differed significantly between the acoustic settings with comments by the singers suggesting that the more reverberant setting (Large Choral) was the most enveloping. However, it should be noticed that longer reverberation times are not always correlated with a sense of envelopment (for the listener at least), and indeed the visual aspect of a virtual simulation can also influence a sense of envelopment. The inclusion of a visual aspect to the VSS is mentioned by the majority of singers in this study as a possible improvement to the VSS in order for the simulation to becoming more enveloping

"Tempo" and "Clarity" ratings were also affected by the simulation, but not the "amount of reverberation" which suggests that there may be another aspect of reverberation which differs between the *real* and *virtual* simulations. Indeed the strongest and most significant effect of *simulation* was on the ratings of "quality of reverberation" indicating that the spectral characteristics of the reverberated sound in the virtual simulation does not successfully match that of the real performance space.

The objective comparison of a number of room acoustic parameters measured in the real and virtual performance spaces presented in Section 3.5.4 have shown some differences, which here could have led to singers impressions of the quality of reverberation. Chapter 7.1.1 outlines further work which will be undertaken to minimise undue signal colouration in the system and to mitigate against the auditory effects of the acoustics of the listening room itself.

Nevertheless, on the whole singers were happy that the VSS was a plausible simulation of the *real performance space* and enjoyed singing in the system; in fact some singers expressed surprise when taken to the real venue at how realistic they found the simulation had been.

3.7 Conclusion

This chapter described the implementation of the VSS, covering the methods used to measure the *real performance space* SRIRs, editing them for use in the simulation and methods of rendering and reproducing the sound field for the singer.

It has presented an objective evaluation (Section 3.5) of the VSS undertaken by calculating a number of room acoustic parameters seen to be relevant to the performing musician. Section 3.6 briefly described the subjective responses to the VSS through interview and questionnaire.

Although the objective evaluation of the VSS show that some errors exist, singers' evaluations rated the simulation highly. Chapter 6 describes how recordings made in the VSS and real venue are analysed and compared in order to ascertain if singing performance changes are similarly elicited in the matching acoustic configurations of the *real* and *virtual performance spaces*.

Chapter 4

Singing in space(s)

4.1 Introduction

This chapter begins with a an overview of music performance analysis and its development throughout the 20th and 21st centuries, and describes more recent automated/semi-automated techniques. Section 4.2 outlines the parameters which can be extracted from musical performances and how they are used to calculate or analyse musical performance attributes, introducing some of the specifics of singing voice performance analysis (Section 4.3)

Music performance analysis is covered in more detail in Section 4.3 and many of the attributes which can be extracted from musical performances are described together with the analysis of music performance attributes.

Section 4.4 summarises the empirical research that has investigated the ways in which musical performances are influenced by the room acoustic characteristics of the performance environment.

Section 4.5 outlines what is known about the alterations voice users make to their spoken voice use according to the surrounding room acoustics for example, the lecture theatre or classroom. Section 4.6 examines in more detail the changes in singing voice performance which have been explored by others working in this area of research.

The last section (Section 4.7) reports the experiences of members of a vocal quartet who were asked to sing in different room acoustic conditions provided by the adjustable acoustics of the *real performance space* which forms the basis for the room acoustic simulation of the VSS.

4.2 Music Performance Research

4.2.1 History of Music Performance Research

In 1893 Max Planck, the German theoretical physicist who formulated quantum theory, wrote a contentious article on the use of "natural tuning" in modern vocal music [130]. Since recording devices were not available at that time, he relied on his own acute sense of hearing for the analysis of the size of musical intervals in vocal music. In order to train his ear to hear the small differences between intervals differing by only a small number of cents, he had built a special harmonium with 104 notes per octave, comprising 8 manuals each of which differed from its neighbour by one 'syntonic' comma. Such a laborious process of ear training was necessary as data on actual performance of music was extremely difficult to obtain at that time without easily available recording devices.

Systematic studies of music performance analysis began around the beginning of the 20th Century, since it was at that time that the necessary mechanical and technical tools became available. Such tools were able to record, reproduce and analyse music performances; performances that had previously only existed fleetingly in time and were non-reproducable and therefore difficult to capture and analyse.

In 1916 Dayton Miller invented the "phonodeik" (also called the "phonautograph"), which recorded sound waves photographically, and used it to study the acoustic properties of flutes made from different materials [131].

Other such devices began to appear in the early 20th century, and as it became easier to record and reproduce musical performances, analysis of musical performance increasingly became a topic of scientific interest. In 1937 Vernon [132] undertook a study of the timing and synchronisation of chords in piano playing, which was facilitated by the use of piano rolls used to record and reproduce performances on mechanical pianos. This device together with a specially designed piano camera provided a rich source of data about timing and tempo.

The most comprehensive of these early studies in music performance analysis were undertaken by Seashore [133, 134] during his time at the University of Iowa. His studies looked at a great number and variety of aspects of music performance analysis including: vibrato, pitch and loudness. The use of oscillographs and stroboscopes allowed him to analyse fundamental frequency, tonal, temporal and intensity-related characteristics of a number of instrumental performances (including bassoon, clarinet, cornet, trombone, violin) as well as investigating aspects of the singing voice. He was the first to describe vibrato of the singing voice in scientific terms, identifying three individual parameters - pitch, intensity and timbre - all of which vary in rate, extent and form (see 4.3.11 for more on vocal vibrato).

Seashore's pioneering work [133, 134] led him to categorise music performance attributes into four main groups: tempo/timing, loudness/intensity, pitch and timbre. Lerch [135] explains these categories in more detail as follows:

- tempo and timing: global or local tempo and its variation, rubato, or expressive timing, subtle variation of note lengths in phrases, articulation of tones, etc.
- velocity, loudness or intensity: musical dynamics, crescendo and diminuendo, accents, tremolo, etc.
- pitch: temperament, tuning frequency, expressive intonation, vibrato, glissando, etc.
- timbre: sound quality and its variation resulting from instrumentation and instrument specific properties such as bow positioning

Manual music performance analysis

Until the second half of the twentieth century, studies of music performance such as those outlined above relied on time consuming manual techniques based on inspection of the audio waveform, identification of the onsets and offsets of notes, followed by subsequent analysis of other features such as average fundamental frequency or vibrato rate and extent.

In many cases the most time consuming task in performance parameter extraction is to find the time stretches which correspond to individual notes. Some authors have managed to undertake empirical performance analysis studies despite their having to use labour intensive techniques of annotating note beginnings and endings manually after inspection of the audio waveform, counting cycles and calculating fundamental frequency values. For example, in 1961 Shackford [136] undertook a study of the sizes of musical intervals in string quartet performances by hand counting wave peaks captured on 35mm film. Automatic cycle counting could not be used due the presence of vibrato in the tones, and so the study also investigated the perceived pitch of the vibrato tones, as well as calculating the performed interval sizes in cents and discussing the use of equal temperament and other tuning systems.

When using such time-consuming techniques, usually only a small number of pieces or performances could be analysed and evaluated, and often special exercises were written in order to produce specific information about the issue in question - specially written exercises were used for example in Shackford's study [136] and in Planck's study [130] on "natural tuning")

It was not until much later, towards the end of the twentieth century, that music performance analysis gained ground, as increasingly automated methods were devel-

oped, although similar investigations continued in the 1960s and 1970s in the field of ethnomusicology and studies of acoustic properties of musical instruments.

Automated Music Performance Analysis

The standardisation of MIDI (Musical Instrument Digital Interface) technology in the early 1980s, meant that not only could electronic musical instruments be connected and communicate with each other, but also a rich seam of musical performance data was available to the researcher. MIDI carries messages which indicate note numbers together with key-velocity, as well as parameters for volume, vibrato, panning, and clock signals which are used to relay tempo and timing of notes.

MIDI-enabled pianos which appeared in the 1980s provided rich data about piano performance and gave the investigator exact details on a large number of parameters including onset velocity, duration of tone, and the pitch of the note played. For this reason keyboard performances have been studied in great detail since the 1980s. Interestingly one of the first studies of piano performance using MIDI-data [137] analysed piano performances given by the same pianist but in different room acoustic settings; the findings of this study are summarised in section 4.4.5.

The development of digital signal processing techniques towards the end of the 20th century has also been an important contributing factor to the growth of the research area of music performance analysis. Standard office/home personal computers were now able to be used for recording and storing audio signals, and meant that larger quantities of data could be recorded and analysed, allowing investigators to attempt more general descriptions of music performances of different styles and genres, different groups of performances or, indeed, performances given in different acoustic environments (see Section 4.4 for more on these).

In recent years, newer techniques for data extraction and data mining have been developed (for example [138, 139, 140, 141, 142, 143, 144, 145]) and such automated techniques have enabled large quantities of performance data to be extracted and analysed. Lerch's book "An Introduction to Audio Content Analysis: Application in Signal Processing and Music Informatics" gives a thorough exposition of the history of music performance analysis, current techniques for Music Information Retrieval and analysis, and the ever increasing number of applications for research in this field [135].

Music Information Retrieval (MIR) has grown from a need to help organise and retrieve digital music from vast collections either stored on-line or locally. Digital music files can be accompanied by meta-data describing the music/audio (metadata-based approach) which facilitates indexing, searching and storing data. Nevertheless, attaching meta-data to a music file can be a laborious process, and therefore much research concentrates on the

extraction of music information from the audio signal alone which can then be stored as meta-data to accompany the file (content-based approach). Applications for both types of MIR lie in the organisation of music collections (multiple audio tracks of different genres, types and by different performers), automated software to suggest music choices to the listener based on past preferences, or for music retrieval (finding particular tracks amongst a large collection of audio via some given input data).

There are now available a number of musical parameter (or feature) extraction software tools. Notable software applications for automatic audio content analysis include the MIDI-toolbox for MATLAB [146], which allows the user to visualise MIDI data and melody, using visualisations similar to the historical piano-rolls, and facilitates the retrieval of data on keys, meter and melodic similarity.

Researchers at Queen Mary University London have developed Sonic Visualiser [147, 148, 149] a software application for viewing and analysing the contents of music audio files, which can be augmented by VAMP plugins, many of which perform music performance analysis audio feature extraction, or MIDI file score matching.

Devaney et al. [150, 151] recently developed AMPACT (Automatic Music Performance Analysis and Comparison Toolbox) which builds on a number of existing MATLAB toolboxes including the MIDI-Toolbox [146] and is specifically optimised for analysis of the singing voice. Further information on AMPACT and its use in the present research can be found in section 6.4.

4.2.2 Musical Score and Performance

For many years musicologists relied on the consideration of the musical score for musical analysis. However, it should be noted that musical performances can vary in many complex ways from that which is indicated a in musical score, and therefore analysis and measurement of actual performances are more valid than descriptions of scored music [152].

More recently researchers have investigated the relationship between the performer's intention and the listener's perception, especially in terms of how emotion is expressed through musical performance [152]. Some of these studies are outlined in section 5.2.3 below.

4.2.3 Musical Structure and Performance

50The relationship between musical structure and musical performance attributes has been a rich area of research. Many authors, for example, identify correlations between tempo and loudness, and the overall musical structure of a piece [153, 154, 155, 156, 157].

Lerch [158] gives a good overview of the role of musical structure and performance stating:

> *Most authors agree on the close relationship between musical structure such as musical phrases or accents and performance deviations mainly in tempo and timing. In particular, larger tempo changes seem to be most common at phrase boundaries. There is a general tendency to apply ritardandi or note lengthening at the end of a phrase and moments of musical tension* [158].

Repp [159] found that musical structure played the largest role in influencing expressive timing for pianists playing the same piece at three different tempos, whereas tempo and intensity-related features changed in proportion with the global tempo differences.

For example, to emphasize a particular note in a phrase for musical effect a musician has a number of options. A note can be "stressed" by increasing its relative intensity, lengthening its duration or delaying its start - these aspects were already understood by Seashore [134]. So, the interplay between temporal and intensity related parameters must be noted, given that a change in intensity is not always needed to signify the emphasis of a particular note and subsequent shaping of a musical phrase.

In contrast, in a study of string quartet performances, Lerch [158] found that timbral aspects did not systematically vary with musical structure, concluding that, unlike tempo and loudness, variations in timbre were not clearly related to musical structure.

4.2.4 Music Performance Studies

Studies have often shown differences in performance attributes between different groups of performers: beginner learners and more experienced or trained musicians [160, 161, 162, 163]; different cultural styles such as early music style singing and operatic singing [164]; country singing and classical singing [165].

Timmers, in a study of historical and modern recordings [166], also notes that performance style and fashion can change over longer time periods. She found that performances recorded in the early 20th century included more extreme tempo fluctuations, more frequent pitch glides and less prominent vibrato than more modern singing performances. Similarly Bowen [167] measured the length of performances of Beethoven Symphonies and looked at the historical trends in tempo from a series of orchestral recordings dating back to 1912.

It should be noted that in the MIR literature the terms "parameter" (also known as "feature") and "attribute" are very frequently used interchangeably; nevertheless it is beneficial to make a distinction between the them. In this thesis, "parameter" will be used in regard to data which can be extracted (either instantaneous or over periods

of time) and "attribute" will refer to a characteristic of musical performance which can be described by the analysis of extracted parameter data. For example, fundamental frequency is one parameter which needs to be extracted in order to be able to further characterise attributes such as vibrato or intonation.

4.2.5 The Singing Voice

The vocal system can be analysed in three parts, the Power Source (lungs), the Sound Source (vocal folds) and the Resonator (or Sound Modifiers) i.e. the lips, tongue, jaw, mouth, vocal tract.

Power Source

The lungs are connected to the diaphragm and the intercostal muscles of the rib cage and behave like a set of bellows: as the diaphragm (a flat, dome-shaped muscle cross-sectioning the body at the bottom of the rib cage) contracts, the rib cage expands and the abdominal organs move downwards, which in turn means that the lungs expand, creating negative pressure in the thorax causing air to move into the lungs. As the diaphragm rises and the rib cage returns to its pre-expanded state, air leaves the lungs via the wind pipe. If the glottis is closed, or partially closed when air leaves the lungs phonation occurs and a pitched sound is produced [168, 20]

Sound Source

The vocal folds primary function is to act as a valve to protect the lungs from solids such as food stuffs and liquids. The positioning of the larynx and function of the vocal folds have evolved and are now also adapted for use in communication through speech. The vibrating motion of the vocal folds results in a complex periodic waveform. This waveform consists of a fundamental frequency and its relative partials (overtones) in the harmonic series. The creation of a pitched sound through the vibration of the vocal folds is known as phonation. The fundamental frequency (and associated perceived pitch) produced by the vocal folds corresponds to the number of times they vibrate every second (Hz).

The phonation frequency is controlled by the complex musculature of the laryngeal mechanism with prominent control from the cricoid and thryroid cartilages, which contribute to the tilting of the larynx, via a hinged mechanism, in order to stretch the vocal folds. The folds stretch and lengthen to raise the fundamental frequency of a sung note through the contraction of the cricothyroid muscle, which tilts the thyroid forwards and tenses and elongates the vocal folds. The complex musculature system of the larynx is

explained clearly in [169]. Some of the techniques used to measure production-related parameters concerning the vocal output are described in Section 4.3.5.

Sound modifiers

Garcia [170] was the first to consider and prove that the area above the vocal folds (the sound modifiers): the vocal tract, tongue, mouth, jaw, teeth, lips and nasal passages act as a resonator. As mentioned above, the speed at which the vocal folds vibrate controls the perceived pitch of the sound being sung (i.e. the phonation frequency). However, it is the manipulation of the sound modifiers which changes the quality of the sound, including the perception of different vowel sounds, overall timbral quality and perceived loudness.

The position and shape of the sound modifiers changes the relative energy of the harmonics in the frequency spectrum of the vocal sound, producing broad peaks in the spectral envelope, known as formants. Section 4.3.3 presents more detail on formants and the "singer's formant cluster", two important aspects of vocal sound production.

4.2.6 Vocal Performance Analysis

Although acoustic analysis of speech has a long tradition, objective analysis of the singing voice is a more recent field of research. Even so, there is now a good body of singing voice science and vocal performance research; thorough summaries of much of this research have been provided by Sundberg in 1981 [171], Cleveland in 1994 [172] and similarly on choral acoustics by Ternström in 2003 [173]. Kob et al. [174] also summarised the state of the art in singing voice research, highlighting areas of voice analysis which still need further investigation. Many of these challenges are being addressed by researchers from diverse backgrounds in a lively interdisciplinary international field of research.

For the singing voice, correlations between internal/external influences on the singer and the resulting acoustic output are not clear cut. Given the large number of quantitative parameters and the seemingly myriad aspects of the singing voice which can change depending on a number of inter-related factors, such as differences in singing style, modes of phonation, frequency dependent vocal function and voice classification, to name but a few, most authors focus mainly on quantifying a small number of attributes. Many of the studies undertaken seek to address the differences in vocal performance between diverse groups of singers or different means of producing vocal sound, for example:

- singing styles e.g. musical theatre and opera [175], yodellers and non-yodellers [176]
- performance styles, e.g. early music, opera and naïve [164] or solo vs chorus [177, 178, 179]

- different voice classifications e.g. tenor, baritone, bass [180]
- categories of timbre [181]
- modes of phonation/vocal register [182, 183, 184, 185, 186]
- changes over time, due to training [160]
- spoken voice and singing voice [187]
- production-related vocal function aspects [182]
- trained and untrained singers [160]
- developing voices in children and teenagers [188] [189, 190, 191]

The next section is organised according to five main categories of music performance attributes: timbral, production-related, intensity related, temporal and tonal and describes briefly some of the relevant music performance research that has been undertaken. In particular it discusses some of the quantitative voice measures/vocal parameters in general use. Focus is given to those already pinpointed by other authors as being varied under different room acoustic performance conditions. Some of these parameters are therefore used in the present analysis of vocal performances recorded in the *real performance space* and in the *virtual performance space* as described in Section 6.4.

4.3 Music Performance Analysis

This section outlines some - but by no means all - of the music performance parameters and analysed attributes which are discussed in the now considerable research literature on music performance analysis. A number of performance attributes and findings relating to their measurement and analysis are summarized below. It is by no means an exhaustive list; more comprehensive reviews of current understanding in this area and the digital signal processing techniques used to analyse features extracted from audio are given by Gabrielsson [192, 193] and Lerch [135].

A number of objective measures of performance have been used recently by authors to quantify aspects of musical performance in order to compare different performance styles (for example [194, 164]), differences between trained and un-trained musicians (for example [195]) or intonation and tuning (for example)[139, 195]).

4.3.1 Timbral Parameters

The Oxford English Dictionary defines timbre as "the character or quality of a musical sound or voice as distinct from its pitch and intensity" [196]. Timbral properties of musical sound relate first and foremost to the spectral characteristics of the signal, and indeed this interpretation of timbre echoes Helmholtz's [197] reason to coin the word "Klangfarbe" (*tone colour*) making an analogy with colour as changes in the spectrum of light. However, since timbre (also sometimes termed *tone quality*) is a multi-dimensional property, a number of additional acoustic features can contribute to the timbre of a musical note, such as its amplitude envelope and temporal characteristics.

In order to examine timbral attributes, following the narrower definition of timbre, a frequency domain representation of the audio waveform is needed, which can be gained by using a Fourier transform to identify the component frequencies and their relative magnitudes.

4.3.2 Timbral Attributes

Perceptual aspects of timbre are difficult to pin-point, but often they are best described by listeners in listening tests in terms of bi-polar scales such as soft-hard, dark-bright, lean-full. A large number of timbral attributes relating to the spectral properties of musical sounds have been posited in music instrument acoustics research and music performance analysis, and some of them are described here:

- **Spectral Flux** - a measure of the rate of change of the spectral shape, with low values signifying low roughness and steady-state signals [158, p.77].
- **Spectral Centroid** - sometimes denoted as Harmonic Spectral Centroid is the weighted mean of the frequencies or harmonics in the audio signal, or the centre of gravity of the spectral energy [158, p.78].
- **Spectral Spread** - describes the concentration of energy around the Spectral Centroid.
- **Spectral Slope/Roll-off Frequency** - a measure of the bandwidth of the audio signal which is dened as "the frequency bin below which the accumulated magnitudes of the STFT reaches 85% of the overall sum" [158, p.76].
- **Mel Frequency Cepstral Coefficients(MFCC)** - Mel Frequency Cepstrum is a perception-based analysis of the frequency spectrum using linearly spaced filters below 1000 Hz, and logarithmically spaced filters above 1KHz, in order to mimic the

way the human ear processes the sound spectrum. The first 20 coefficients of the discrete cosine MFC transform are seen to represent formant peaks in the spectrum.

4.3.3 Vocal Timbral Attributes

Vocal timbre is generally analysed by looking at the LTAS of the recorded voice signal but some authors have also studied voice source characteristics and how they contribute to voice timbre [198, 175, 199, 200, 201]

Long-term Average Spectrum

A long-term average spectrum where the spectral envelope of the voice signal is averaged over time, is often used to enable spectral characteristics of singing to be compared. It has been shown that a sample of speech or singing needs to be longer than 30 seconds in order for the spectrum to stabilise [168]. An LTAS is understood to be insensitive to the linguistic content of speech or singing (when the sample used is sufficiently long), as formant patterns arising from vowels are averaged over time, so that longer term characteristics may be examined.

A number of authors have examined LTAS of different singing styles, for example Barlow and Lovetri found that for young singers Musical Theatre singing differed from 'classical' singing in the relative strengths of the first six harmonics [189]. Monson [121] looked at the balance of high-frequency energy and low-frequency energy in speech and singing. He found that listeners were able to perceive changes in high frequency energy (HFE), especially in the the 8kHz octave band, and were more sensitive to these changes in singing rather than speech where listeners reported that HFE was important for voice quality evaluation.

Singing Power Ratio and Energy Ratio

Singing Power Ratio (SPR) has been used by some authors to characterise differences between singing voices. SPR is defined as the ratio of the peak intensities between the regions 2kHz - 4kHz and 0 - 2kHz [202, 203]. Lower SPR values indicate greater energy in the higher harmonics of the vocal sound, which are often seen to correlate with the perception of "ring" and "richness" of the voice. Watts found differences in SPR measures between untrained talented and untrained nontalented singers [203]. Similarly, the energy ratio (ER) measures the balance of the average energy in the low (0 –2 kHz)and high (2–4 kHz) ranges of the spectrum. Low ER and SPR values indicate more energy in the high range and singers formant region relative to the energy in the low range of the spectrum [177].

Formants

Formants were defined by Fant as "spectral peaks of the sound spectrum of the voice" [204]. The centre frequencies, bandwidths and relative amplitudes of formants in the vocal spectrum allow different vowels to be distinguished by the listener. Of course, the positions of formants relate to vowels, and are therefore mostly determined by the text or lyrics of a song.

For example, Figure 4.1 gives an idealised illustration of the formant peaks which make up an /ɑ/ ("ah" vowel). The first two peaks (formants 1 and 2) are close together whilst the third formant is further separated.

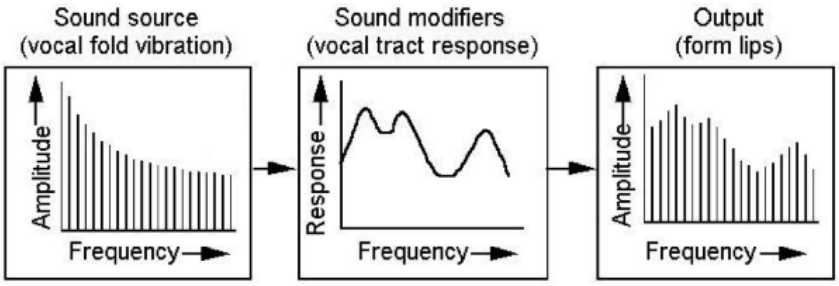


Figure 4.1: *Images showing the impact of the sound modifiers on the voice source sound spectrum to create formant peaks when producing an "ah" vowel from [20] reproduced with permission*

If the second formant is raised towards the third formant, via the movement of the middle of the tongue upwards towards the hard palate, the ear recognises /i/ (an 'ee' vowel).

A small number of authors have measured formant frequencies and bandwidths to account for the perceived differences in timbre between singing styles such as solo and choral singing [178, 205, 206, 177] or between trained and untrained singers [161].

Singer's Formant Cluster

Several techniques are employed by singers, particularly opera singers, to increase their perceived loudness and allow them to be heard over large orchestras in concert halls and opera houses.

The most commonly known of these techniques is the "*singers formant* " which applies most significantly to the operatic tenor voice. The *singer's formant* (now more usually known as the *Singer's Formant Cluster (SFC)*) was first suggested by Sundberg [168] to describe an area of increased spectral energy in the vocal sound in the region between 2.5kHz and 5.5kHz; it is generally found in operatic singing and related western Classical styles, but is usually absent in other types of singing and speech.

The Singer's Formant Cluster (SFC) manifests as a peak in the spectral envelope between around 2 kHz –4 kHz. It is usually achieved by manipulation of the sound modifiers (e.g. an increase of pharyngeal space) which results in the bunching together of formants 4,5 and 6 and is particularly related to the technique of lowering the larynx. Although the spectral envelopes of each orchestral instrument differ (giving each instrument its identifiable timbre), an idealised spectral envelope of the acoustic output from a whole orchestra, illustrated in Figure 4.2, produces a gradual decay in amplitude of harmonics, meaning relatively low amplitude in the frequency region of the singers formant cluster. The 2kHz–4kHz frequency region of the SFC is also the most sensitive frequency region of the human hearing response, adding to the auditory significance of the technique.

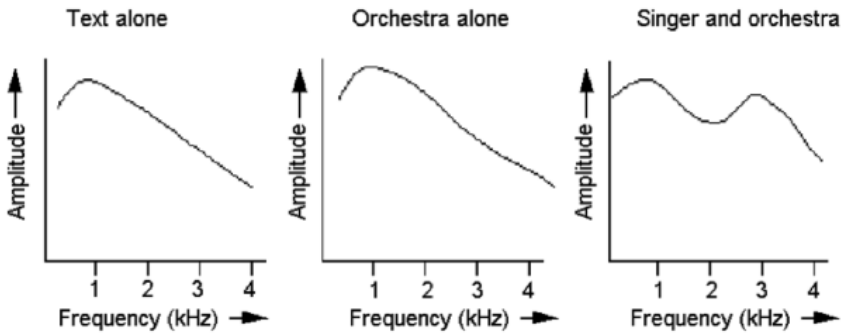


Figure 4.2: *The impact of the singers formant cluster on an idealized spectral envelope of an orchestra and singer from [20].*

The SFC increases the relative amplitude of harmonics in the spectrum between 2–4 kHz. As a complex tone includes harmonics at integer multiples of the fundamental frequency (F0), (i.e. 1st harmonic (H1) = F0∗1, 2nd harmonic (H2) = F0∗2, 3rd harmonic (H3) = F0∗3 etc.) the higher the fundamental frequency, the fewer harmonics will be present in the singers formant area of the spectrum.

The acoustic possibilities of the SFC become less useful the higher in the pitch range the voice-type of the singer is placed. Whilst low female voices may make use of the technique, soprano voices in particular make use of alternative techniques, both physiologically and acoustically. For example, in the speech of an adult female the expected frequencies of the first 3 formants for this vowel would be 850Hz, 1200Hz and 2800Hz. However if singing an A5 (fundamental frequency of 880 Hz) the first formant is redundant as there is no sound energy in this frequency range to amplify. The soprano singer therefore tunes the (now redundant) first formant (again by manipulating the placement of the sound modifiers) up to (or near to) the fundamental frequency, greatly increasing the relative intensity of the fundamental and increasing the perceived loudness of the sung tone. However, the spectral envelope no longer has distinguishable peaks, putting into question vowel recognition when higher fundamental frequencies are sung (see idealised illustration in

Figure 4.3).

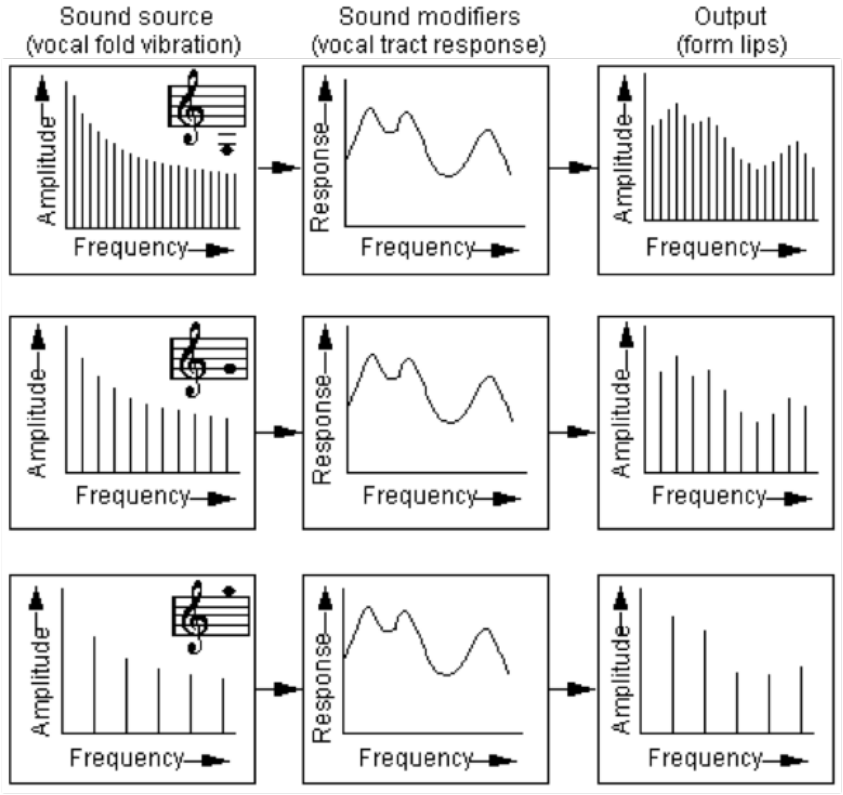


Figure 4.3: *Showing the spacing of the harmonics at different fundamental frequencies changing the impact of the formant frequencies on the spectral envelope. (Reproduced with permission from) [20]*

Once the technique to produce the SFC has been learnt, it can be difficult to avoid. Howard [207] demonstrated that a professional singer trained in Western Classical style still produced a strong spectral peak around the singer's formant region, even when performing in a non-operatic style. However, not all singing styles employ the SFC, for example, Cleveland [165] found no evidence of the singer's formant in country singers and indeed showed that country singers' vocal output spectra were similar for speech and singing.

Autocorrelation Function (ACF)

Autocorrelation is a signal processing technique which provides a measure of similarity between a signal and a time-delayed copy of itself at a given time lag (τ_e). When a reference window of the original signal matches the time-lagged window, the correlation value is large. In the case of a periodic signal, peak correlation values will be reached at lag distances which correspond to one period, or integer multiple periods, of the fundamental frequency.

A handful of authors have attempted to use ACF as a measure of the timbral quality of singing voice, for example, Noson et al. [208] investigated the effectiveness of using $\tau_e(min)$ as an evaluation criterion for singing voice. $\tau_e(min)$ is defined as the minimum value of effective duration of the running autocorrelation function (r-ACF). Noson et al. used this parameter to characterise the recorded audio signal of a singer in an opera house to quantify the "blendedness" of the singer's voice in the performance venue and argued that $\tau_e(min)$ showed close correlation with the listeners' and performers' subjective impression. They also showed that $\tau_e(min)$ varies according to performance style, singing style, vowel, relative pitch, extent of vibrato and intonation and argued further that the fine structure of the running ACF (rACF) related to the identification of vowels in the singing voice.

4.3.4 Production-related Parameters

The extraction of data on the technical mechanisms or physiological processes which are involved in playing a musical instrument is often used as another source of information on musical performance. Production parameters are not usually investigated in isolation but rather examined in relation to the corresponding timbral or temporal aspects of performance. With the standardisation of MIDI in the 1980s it became easier to extract production parameters such as key velocity for studies of piano playing; more recently other sophisticated sensors have been developed to obtain data of finger force, pedal timing etc.

For bowed instruments, motion tracking sensors have recently been developed which can be used to detect motion and other aspects such as bow force and bowing speed. For example, Chudy et al. [209] studied the relationship between gesture, tone production and perception in classical cello performance and extracted not only acoustic parameters of the music performance, but also features of the production mechanism and instrumental playing technique.

4.3.5 Vocal Production-related Parameters

Inverse filtering

A number of voice source parameters have been analysed through the use of inverse filtering, which filters the voice output signal with the inverse of an estimated vocal tract filter shape, in order to ascertain the original voice source waveform. Voice source parameters which can be calculated from the inverse filtered voice signal include glottal flow, glottal closure, opening and closing phase, open and closed quotient (see also Section 4.3.5).

Cleveland & Sundberg [210] examined the voice source properties of three different male voice categories by measuring the closed phase of the glottal cycle from an acoustic glottogram achieved by inverse filtering the acoustic voice signal. They found that the fundamental frequency amplitudes differed between the bass, baritone and tenor singers, with the bass singer demonstrating the strongest fundamental.

The present author used inverse filtering techniques to study the differences between male opera singers' chest and head voice registers [198, 183]. Other researchers have used inverse filtering techniques to examine differences between vocal registers [211, 198, 183], or to analyse and synthesise the voice source in speech, for example [201, 212].

Electrolaryngograph

The output signal from an electrolaryngograph enables not only very accurate analysis of fundamental frequency of the singing voice, but also analysis of the shape and regularity of vocal fold vibration. An electrolaryngograph consists of two disk electrodes which are worn externally by the singer on either side of the neck in line with the thyroid cartilage. A small constant radio frequency (RF) voltage flows between the electrodes and the resulting output signal (Lx) plots the current flow against time. When the vocal folds are in full contact the current flow is high, and as the vocal folds move apart the current flow decreases [213, 194].

Open and Closed Quotient

The percentage of each cycle when the vocal folds are open can be measured from the Lx waveform. A number of different methods exist to determine the exact start and end points of the closed phase and open phase (see [213, p5]).

Authors have also found differences in closed quotient measures in the changing voices of young singers [189, 188] and changes in closed quotient values as a function of singing training [160].

4.3.6 Intensity-related Parameters

A number of intensity-related parameters can be extracted and calculated from the audio signal and are generally correlated to the perception of musical dynamics.

Sound Pressure Level

Average sound pressure levels across bars, notes or phrases can be calculated if the reference level is known. Relative sound pressure levels can be analysed where the absolute dB SPL is not available. Root Mean Square measures are generally used for this purpose -

the mean root value of the squared sound pressure level of a signal over time - and denotes the intensity of a signal.

For example, in a study of historical and modern solo singer recordings, Timmers calculated the average sound level of each bar [214] and found that the correlation with perceived dynamics was not very high.

4.3.7 Intensity-related Attributes

Loudness and Musical Dynamics

Some authors have employed perceptually motivated measures of loudness in order to characterise differences in musical dynamics or overall loudness levels.

Overall loudness has been show to correlate with tempo in piano performances [215] which is probably due to the large amplitude of vertical finger movement in the pianists' playing technique.

Sound pressure level readings of singing performance have been found to differ between performance styles. For example, Howard et al. found a difference of 17dB between three different singing styles [164].

Loudness characteristics correspond generally to changes in musical dynamics such as "crescendo" (increasing loudness over time) and "diminuendo" (decreasing loudness over time), but other attributes also contribute to the perception of musical dynamics, such as timing and timbre.

Intensity Vibrato

Some musical instruments such as woodwind instruments and the singing voice can exhibit an intensity vibrato, which consists of (quasi-)periodic changes in intensity over time during notes. Intensity vibrato (IV) very often co-exists with fundamental frequency vibrato (quasi-periodic changes in fundamental frequency over time (see section 4.3.11).

4.3.8 Temporal Parameters

Temporal parameters are some of the most fully researched in music performance analysis studies since they are relatively easy to extract using simple methods such as timing a performance, or counting the number of beats in a performance of a certain time length to obtain a tempo rating. Identifying note onsets and offsets is crucial to much performance analysis, since not only can tempo related attributes such as variations in tempo/rubato be calculated once these are found, but many other performance attributes such intonation

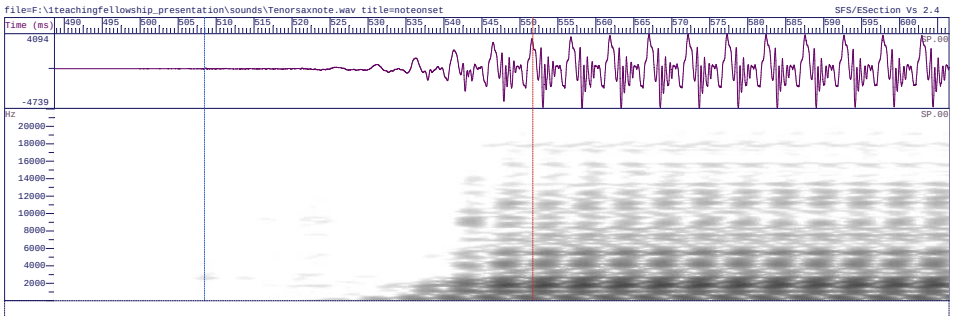


Figure 4.4: *Illustration of the note onset (initial portion of note between vertical lines) of a saxophone note*

(see Section 4.3.11) and vibrato (see Section 4.3.11), rely on successful identification of note beginnings and endings.

Note Onsets

Pinpointing the position of individual notes in a musical performance is an important first step before a number of other parameters can be calculated. Once note onset times have been found other parameters can be subsequently calculated, such as note durations, note-on ratio and inter-onset intervals. This is relatively easy for MIDI data taken from keyboards, and indeed much research has been facilitated by using MIDI data which includes note-on and note-off "messages" (e.g. [216, 137, 217, 153, 154, 218]). Note beginnings and endings must also be determined in order that tone-related attributes such as pitch, intonation, and vibrato can be analysed (see section 4.3.11).

Several authors have proposed algorithms to automatically identify the time instants of note onsets. Most of these are based on peak-picking algorithms which detect local maxima in the audio waveform and/or a sudden change in the amplitude of the waveform. Bello et al. have produced a useful tutorial on onset detection in musical signals [219].

Very often the start of a tone includes a short time span from the start point of the instrumental sound until the point where a quasi-periodic state is reached (see Figure 4.4) This onset time span is also referred to as *rise time* or *initial transient time*. Different musical instruments can exhibit varying note onset envelopes - indeed it has been shown that the ability of listeners to distinguish between different instrument sounds relates to a large extent on recognition of the note onsets [220]).

Nevertheless, it should be noted that the human singing voice and some musical instruments, such as bowed string instruments, do not always demonstrate impulsive note onsets, and in these cases different techniques have to be used.

Vocal onsets and offsets

Accurate data on the temporal aspects of singing voice performance can be difficult to obtain. Most note onsets in vocal music are non-percussive and so the standard peak-picking algorithms used for automating the process of finding onset times and temporal note positions are not adequately accurate. Vocal onsets can also take a number of different forms primarily determined by the text of the song. For instance, vocal onsets might comprise a glottal onset (where the vocal folds are closed prior to the onset of voicing) or glide onset (where vocal fold vibration starts from an open glottis), whilst notes that begin with consonants demonstrate many different characteristics according to the acoustic properties of the consonants involved. Transition times between vowels and consonant sounds, relate to the overall articulation of the text and can be used as a quantitative measure of articulation in vocal performance.

Note offset times

A decision must be made by the researcher to identify the point at which one can measure the note offset. For example, in a recent study of drum and speech sounds, Patel & Iverson [221] defined the note decay time as the 'duration between envelope peak and point at which the envelope decays to 50% of the peak value'.

A number of quantitative measures can be derived once note onset and offset times have been determined, such as those outlined below.

Note durations

Note durations can be easily calculated as the difference between note onset and offset times. Variations in note lengths and phrases contribute to *agogics* - see section 4.2.3. Note durations are denoted by some authors as *note-on ratio*. For example, in a study of piano performances in different room acoustic conditions Bolzinger and Risset [137] found that notes were played more "staccato" i.e. the *note-on ratio* was smaller, when reverberation time in the room increased.

Inter-onset Intervals (IOIs)

Inter-onset intervals (also sometimes called inter-tone onset intervals) can also be calculated from the identified note onset times.

$$IOI(i) = \frac{t_o(i+1) - t_o(i)}{fs} \tag{4.1}$$

Where IOI(i) represents the inter onset interval (in seconds), $t_o(i)$ is the sample number of the detected onset and $t_o(i+1)$ the sample number of the next onset, and *fs* is the sample frequency.

In a highly detailed analysis of 28 recorded piano performances by a number of performers, Repp extracted Inter-tone Onset Intervals (IOIs) and undertook a series of statistical analyses of the data [222]. He found that, within the constraints of global and local timing patterns which appeared to be determined by musical structure of the piece (e.g. *ritardandi* at the end of major sections), pianists displayed a huge amount of individual variation at the lower levels of the structural hierarchy, such as in the timing of individual notes and phrases.

Beats

In order to ascertain the tempo of a performance the position of musical beats must first be identified. This can either be achieved by annotating the onset times in a waveform editor e.g. [153] or using one of the beat tapping systems designed to allow a listener to tap along to the beat whilst listening to the audio signal [223]. Once accurate beat positions are known, temporal attributes such as global or local tempo can be calculated. Others have used automatic or semi-automatic beat tracking systems, for example [224], or more recently developed automatic alignment algorithms which use information from the score of a MIDI file to guide the detection of beats [225, 150].

4.3.9 Temporal Attributes

Tempo

Extracted timing parameters allow the changes in tempo in a musical performance to be plotted by means of a tempo curve, which simply plots the tempo as a function of time for the duration of a piece. A number of authors have used this method in a descriptive approach to musical performance analysis.

Global tempo

Global tempo is the average tempo calculated across the whole of a piece, song or substantial section of piece. In an investigation of recorded performances of keyboard pieces by J.S. Bach, Schulenberg [226] calculated the proportions between overall average tempo of the piece in certain key sections. He found that tempo relationships between the opening section and the fugue showed a broad continuum, which did not seem to relate to age of the recording, nor the instrument on which the piece was played (harpsichord,

piano). The study focussed on average tempos across large sections of the piece, rather than on local tempo variations (musically known as "ritardandi", or "rubato") - such local variations in tempo are often characterized by calculating local tempo rates as described in the following section.

The average tempo of a sung passage can be found by measuring its overall length and dividing by the number of musical beats it contains, giving a value for average beats per minute (BPM)- equivalent to musical metronome markings. However, care needs to be taken when there are silences between musical phrases (musical rests), or between verses of a song, as these may not be of the "correct" duration, especially if a singer is singing unaccompanied.

Local tempo

Measures of local tempo are evaluated by the length of individual local beats. The variation in this local tempo can be used to identify "rubato" or expressive timing. For example, Repp [222] studied the variations in the timing microstructure in several performances of the same piano piece given by a number of pianists. Local beats are calculated as follows:

$$BPM_{local}(i) = \frac{60s}{t_b(i+1) - t_b(i)} \tag{4.2}$$

Where $BPM_{local}(i)$ represents the local tempo (measured in Beats Per Minute), $t_b(i)$ is the time point of the detected beat measured in seconds, and $t_b(i+1)$ the time point of the next beat measured in seconds.

A bar-level measure of local tempo can also be calculated by measuring the duration of each bar of a piece, and dividing by the number of beats in the bar.

Rubato - variation in tempo

Many of the numerous studies on variation in tempo and timing in musical performances have examined performances of solo piano music. For example, Todd looked at the rubato (and dynamics) in piano playing [156, 227] and produced a model which related these attributes to musical structure.

Clarke [157] also related rhythmic tendencies in piano performances to the structural hierarchy of the piece and note-level expressive gestures. Repp [222] similarly investigated the timing in performances of piano music by Beethoven and Schumann, and found that instances of "ritardando" related to the overall structural hierarchy of phrases, e.g. the higher the structural level of the phrase the more pronounced the "ritardandi".

Friberg and Sundberg [228] undertook an experiment where listeners were asked to adjust the onset time of one note in a sequence and found that the JND for tempo (at

least in a regular monotonic phrase) was fixed at 6ms for tones that were shorter than 250ms. For tones longer than 250ms there was a constant relative JND of 2.5 %.

Timing has also been shown to be a crucial factor in the synthesis of musical performances; as Sundberg et al. demonstrated, [217] a computer based MIDI-controlled musical performance will sound more human-like if the timing and variations in timing (rubato) are correctly mimicked [217].

Agogics

Agogics concerns the emphasis of particular notes in a phrase for musical effect by lengthening, increasing loudness, or by other means. For example, the well known technique used in harpsichord and organ playing of delaying the arrival of a note in order to add emphasis was also found by Sundberg et al. [217].

Thorough surveys of timing and temporal attributes of musical performance, as well as all other music performance attributes, are to be found in the comprehensive review articles completed by Palmer in 1997 [229] and Gabrielsson in 1999 and 2003 [230, 192].

4.3.10 Tonal Parameters

In contrast to timbral parameters, the "tonal" category denotes aspects relating to the **notes** (tones) of a musical performance, rather than referring to "tone-colour" which relates to the spectral characteristics of a sound. Tonal parameters rely first of all on the accurate extraction of the fundamental frequency of individual notes. Where the instrument involved is a keyboard instrument this is easily facilitated by the use of MIDI, or by detecting which keys are in use. However, the human voice and other instruments, e.g. stringed instruments, with 'fluid' pitch capabilities, call for more elaborate processing techniques. Some of the more recent digital signal processing techniques used for this purpose are outlined below.

Fundamental Frequency

A number of methods have been developed to facilitate the extraction of fundamental frequency data from the audio waveform. An early automated method is Linear Predictive Coding (LPC), which was developed for speech coding purposes, and relies on the assumption that the speech signal consists of a sound source, and a (vocal tract) filter whose resonances correspond to vocal formants. The position (centre frequency) and level of the formants are predicted by the model, and the errors in the model (the residual) equate to the voice source, the intensity and frequency of which can be calculated. Thus,

LPC can be used to estimate the fundamental frequency of a signal. A good summary of LPC and its uses is given by Makhoul in his tutorial review [231].

Measuring F0 in singing

Whilst average fundamental frequency measures can be useful for speech analysis, in singing the fundamental frequency is generally dictated by the musical score. Nevertheless, the actual fundamental frequencies produced by singers in performance, and their relation to tuning and temperament is a growing area of interest [139, 232, 233, 234, 235].

Most audio analysis software packages now include pitch estimation functionality and many of these are used in singing as well as speech analysis. For example, PRAAT [236] implements an autocorrelation function (as used in recent studies on intonation [233, 234, 235]). The recently developed robust fundamental frequency estimation algorithm YIN [138] has been incorporated into toolboxes such as AMPACT [150]. (See Section 6.4.1 for more on AMPACT and its use in this research)

Whilst pitch estimation techniques such as those based on Linear Predictive Coding or autocorrelation analyses of the audio waveform have been used for fundamental frequency detection in singing, for vocal performances often the most accurate *F0* measurement can be obtained from the electrolaryngographic signal (see Section 4.3.5) if this is available.

4.3.11 Tonal Attributes

Pitch glides

Pitch glides between notes in vocal music are also referred to as *glissandi* or *portamenti*. These have been well studied by many authors and were originally thoroughly described by Seashore [134] as the use of portamento was prevalent at the time of his studies in the 1930s. Indeed, Timmers [214] found that the number of pitch glides up and down in each bar was one of the distinguishing features of historical versus modern vocal performances.

Intonation and Musical Temperament

Tuning a musical instrument raises the problem of which tuning system or *temperament* to use. The ratio of two notes which sound an octave apart is 2:1, and the ratio of two notes which sound a perfect fifth apart (e.g. C and G) is 3:2. If notes are tuned to perfect integer ratios around the circle of fifths (C to G, G to D, D to A etc), after twelve instances of tuning this interval the original note is reached again. However, the fundamental frequency of this note will be slightly higher the starting point. This difference is known as the "pythagorean comma". Several temperaments (tuning systems to work around

this problem) have been suggested and all involve "tempering" certain notes in order to distribute the Pythagorean comma amongst notes of the scale. The modern piano, for example, is tuned in "equal temperament" where each note in the octave is "tempered" to be an equal distance apart, leaving only the octaves tuned to an integer ratio of 2:1.

The debate about which temperament (tuning system) is used by musicians in performance continues to be an area of much discussion. Singers, wind players and string instrumentalists have more freedom over the intonation of the notes they perform, meaning that they can adapt and adjust tuning strategies at will during a performance. Keyboard instruments in contrast are subject to a tuning system/temperament chosen in advance. Early work on intonation has concentrated on whether singers sang in equal temperament, just intonation or Pythagorean tuning [130, 133, 134, 197].

In a study on the intonation of wind instrumentalists Karrick [195] analysed the performance of duets performed with a synthesized harmony line. He analysed the deviation of the produced tones from those expected in equal temperament and just tuning and found that the performed tones deviated least from equal tempered tuning, although there was also variation in the intervals; thirds and sixths were produced slightly less in-tune than fourths, fifths, unisons and octaves. It should be noted that the participants in this study were asked to perform to a synthesized line replayed over headphones and it must be also presumed that the synthesized line was recorded in equal temperament, which might have influenced the tuning strategy of the singers involved.

Examining the intonation of thirds by professional flautists, Leukel [237] found that players on the whole tuned thirds in different harmonic contexts to just intervals rather than equal temperament or Pythagorean tuning.

Temperament in Singing

It has been argued by many that singers in particular maintain just intonation - the system of tuning intervals based on integer ratios e.g. the notes of a perfect fifth are in the ratio 3:2, similarly the perfect fourth (4:3), the major third (5:4). Table 4.1 gives tuning frequencies of an equally tempered and just tuned scale in C major. The largest differences between the two temperaments in the major scale are in the tuning of the major sixth (C to A - difference of 16 cents) and the major third (C to E - difference of 14 cents).

At the end of the 19th century Helmholtz [197] and Planck [130] both argued that a capella choirs sang in just-tuned intervals (intervals with whole number ratios). However, Barbour [238] reported the empirical research undertaken by Guthrie et al. [239] at the University of Washington, which asked participants to state preferences for just tuned and tempered intervals, and drawing conclusions from this work, Barbour stated:

Note	Just Scale (Hz)	Equal Tempered Scale (Hz)	Difference (Hz) (ET - Just)	Difference (Cents) (ET - Just)
C4	261.63	261.63	0	0
D4	294.33	293.66	-0.67	-4
E4	327.03	329.63	2.6	14
F4	348.83	349.23	0.4	2
G4	392.44	392	-0.44	-2
A4	436.05	440	3.94	16
B4	490.55	493.88	3.33	12
C5	523.25	523.25	0	0

Table 4.1: *Table of fundamental frequency values and the difference between notes of the C major scale in equal temperament and just intonation (reference to A=440Hz) presented in Hz and Cents)*

> *These experiments prove conclusively that Helmholtz and his followers are wrong, that singers have no predilection for the so-called natural or just intervals, not even the major third (5/4), the interval which most surely distinguishes just intonation from equal temperament [238, p.53].*

He also cited evidence from Seashore's studies at the University of Iowa [134] on the use of pitch glides, portamento and vibrato, concluding that:

> *Even if the omnipresent vibrato be disregarded, it is highly improbable that just intervals can be sung within a reasonable margin of error. Singers show no natural preference for these intervals [238, p.55].*

He also concluded that string players' intonation was closest to Pythagorean tuning, where fifths are tuned perfect (3:2 ratio) and major thirds are tuned sharp (a slightly wider interval than just-tuned 9:8).

Lloyd added to the argument in 1940 [240] explaining the difficulty in playing just "off the note", that is to say, to achieve the mis-tuning required to sing or play a stringed instrument in equal temperament. He concluded that a flexibility of tuning existed for instruments and voices not restricted by fixed intonation.

The debate over the uses of "just tuning", "Pythagorean tuning" and "equal temperament" has continued throughout the 20th Century and still continues today. With the increase of empirical data obtainable from real performance contexts (rather than under laboratory conditions) more recent authors have added to the debate.

Howard [241, 242] found that in performances involving a capella quartets, singers did tend towards non-equal-tempered tuning, rather than singing in equal temperament. Prame [243] also found that the 10 singers in his study of vibrato and tuning deviated substantially from equal temperament, with the largest differences between the mean F0 of tones and the expected equal tempered fundamental frequencies being ±44cents.

Hagerman and Sundberg [244] studied chord intonation in barbershop singing and found several intervals were narrower than in an equally tempered scale.

In any discussion of vocal intonation it is worth noting that the difference, for example, between an equal tempered major 3rd (400cents) and a just tuned (ratio of 5:4) major third (386 cents) is only 14 cents. This discrepancy is equal to the standard deviation measured across singers in fundamental frequency pitching found by Ternström & Sundberg [245] in choral singers. For example, Vurma and Ross [246], in a performance experiment to ascertain how singers produced intervals and how they were perceived, found that the professional singers in the study sang major seconds more narrowly , but perfect fifths more widely, than equal temperament.

In the quantitative rule-system developed by Friberg [247] for computer-based musical performance *melodic intonation* is defined with narrower minor seconds than equal temperament (so that the leading note is sharper than it would be in equal temperament), for use in performance with single voice line only. These rules for melodic intonation accord with performance attributes examined by Sundberg [248] and Prame [243]. *Mixed intonation* rules combine the wish to keep minor seconds narrower than equal temperament, whilst also trying to maintain beat-free (just tuned) intervals and chords.

In a study of intonation practice in two-part, singing Vurma [249] found that singers' interval deviation from equal tempered (perfect 5th) intervals ranged between 14-24 cents and that singers tended to remain true to their own 'melodic tuning' even when the accompanying part was mis-tuned by up to 40 cents.

Horizontal intonation can be used for expressive purposes, whereas vertical intonation is most highly influenced by the need or desire to tune a chord with other singers and instrumentalists. However, in polyphonic choral music, Devaney & Ellis [139] noted the conflict between the harmonic (vertical) and melodic (horizontal) tuning requirements.

Intonation Metrics

Outside the arguments about which temperament singers use when singing solo, with accompaniment or in a choral setting, researchers have recently become interested in measuring accuracy and precision of intonation in singing. In these studies a clear distinction is made between pitch accuracy and precision.

Accuracy measures the error against a target reference, in the case of singing this

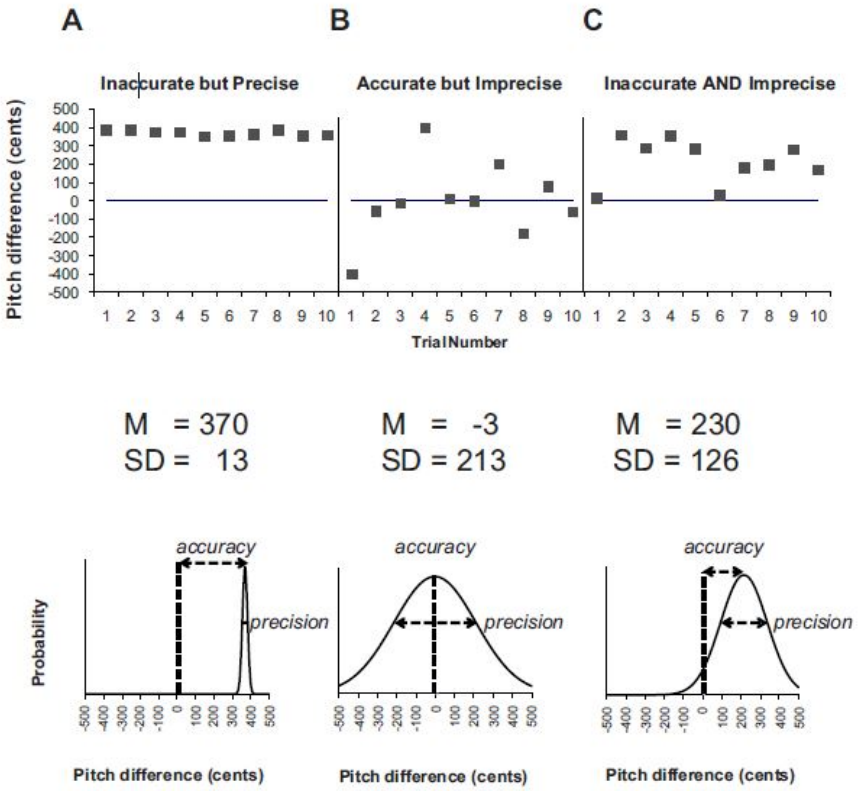


Figure 4.5: *Graphical representation of hypothetical examples of repeated attempts to sing a single pitch class taken from [233]*

might be the expected target pitch of the sung note as denoted by the score. **Precision** on the other hand is a measure of the "consistency" of re-production of a note (or interval). Similarly, interval accuracy is measured against the expected size of the target interval, and interval precision is the consistency or stability of reproduction of the interval in question. It should be remembered that larger values for accuracy and precision relate to "less accurate" and "less precise" singing respectively.

Figure 4.5 taken from Pfordresher et al. [233] serves well to illustrate the differences between precise and accurate singing of a single pitch class.

Intonation Precision in Choirs Ternström and Sundberg [245] examined the acoustical factors that influenced intonation difficulty in choirs, which they defined as pitch precision amongst members of a choir. Using a synthetic sung vowel as a reference they asked individual members of the choir to sing intervals in just intonation, and used the standard deviation in fundamental frequency amongst the ensemble (S$SF0$) as a measure of the difficulty of producing correct intonation. In a second experiment they found that intonation for choral singers was easiest when the synthetic vowel contained no vibrato, but included harmonics common to the two tones in question (reference and

target interval pitch), or if the stimulus tone included higher partials. This study has important implications for intonation practice in different acoustic environments, as the frequency response of a performance space influences the spectral characteristics of the reflected sound a singer receives as aural feedback.

In a later paper the same authors [250], in an experiment to gauge intonation difficulty, measured the geometric mean ($MF0$) and standard deviation of a tone, and the difference between the MFO and the stimulus tone was calculated (in cents). In this study they used the standard deviation of F0 across the tone, referred to as the F0 fluctuation indicator ($F0FI$), as well as the absolute error in MFO, referred to as F0 discrepancy ($F0$D), as measures of intonation difficulty. Mean $F0$D for a unison tone was between 8-12 cents. In order to gain a measure of intonation difficulty where a target frequency is not known another metric was proposed; the standard deviation, across subjects, of the singers' MF0 values, referred to as SF0. SF0 values across eight notes in a phrase were found to lie between 10.3 cents and 15.8 cents, with a mean value of 13 cents.

Intonation in Solo Singing A new paper by Mauch, Frieler and Dixon [251] (in press) investigates intonation and intonation drift in solo singing. Two measures of intonation accuracy are calculated - mean absolute pitch error (MAPE), equivalent to $F0$D used by Ternström and Sundberg as described above, and mean absolute interval error (MAIE).

Mean Absolute Pitch Error(MAPE) Pitch errors are measured with respect to the target pitch in equal temperament with a reference to A=440Hz. Overall mean MAPE across 72 recordings of "Happy Birthday" was 18.9 cents.

$$MAPE = \frac{1}{M}\sum_{i=1}^{M}|e_i| \tag{4.3}$$

Mean Absolute Interval Error(MAIE) The interval leading to the i^{th} pitch is expressed as the distance between the pitch of one tone and that of its predecessor in semitones.

$$\triangle p_i = p_i - p_{i-1} \tag{4.4}$$

Interval error (e_i^{int}) is then calculated as the difference between this distance ($\triangle p_i$) and the nominal interval distance ($\triangle p_i^0$) as determined by the musical score using equal temperament interval sizes (where each semitone is 100 cents).

$$e_i^{int} = \triangle p_i - \triangle p_i^0 \tag{4.5}$$

Mean absolute values of interval error are then calculated:

$$MAIE = \frac{1}{M-1}\sum_{i=2}^{M}|e_i^{int}| \tag{4.6}$$

This means that information about direction of the error, whether the interval was larger or smaller than expected, is lost. (Equations 4.3, 4.4, 4.5 and 4.6 are taken from Mauch et al [251]).

Dalla Bella, Giguiere and Peretz [234] in an investigation of singing proficiency in the general population measure several attributes of intonation and tempo, including *Interval Deviation* (equivalent to MAIE) which is defined as the average absolute difference (in semitones) between the interval produced and the target interval (see Section 6.4.3 for more on interval accuracy). Values of interval deviation in professional singers ranged from 20 - 40 cents with an overall average of 30 cents.

Pitch and Interval Precision Dalla Bella et al also proposed a measure of *pitch stability* (pitch precision) as the mean of the absolute difference in semitones between each corresponding note in two repetitions of the same melody phrase and report values between 10 - 40 cents.

Pfordresher, Brown, Meier, Blyck and Liotti [233] use several metrics to study the topic of poor pitch singing including *note precision* which is a measure of the precision of repetitions of notes within each pitch class (i.e. note name, C,D,E F etc within one octave). They also found, for singers singing familiar melodies, mean values for *Interval Accuracy* (equivalent to MAIE) of 90.8 cents (SD 15.9) and for *Interval Precision* of 155 cents (SD 16.8).

In a study of 50 occasional singers, Berkowska and Dalla Bella [235] report mean absolute pitch accuracy (equivalent to MAPE above) in singing familiar melodies of up to 170.5 cents, mean pitch precision values of 41.2 cents (SD 19.9), mean absolute interval precision (MAIP) of 49.4 cents (20.2) and mean interval accuracy (MAIE) of 19.8 cents (SD 13.5).

Vibrato

Vibrato is defined as quasi-periodic fundamental frequency or intensity modulation over time. Prame [243] undertook extensive investigations into vibrato in professional Western lyric singing tradition and describes vibrato as follows:

> *Basically it corresponds to a frequency modulation of F0 characterized by its rate and extent. This modulation, in turn, causes amplitude modulations of*

the individual spectrum partials which result in a modulation of the overall amplitude. This amplitude modulation may be both in phase and out of phase with the original frequency modulation.

Fundamental frequency vibrato stemming from alteration of the fundamental frequency is denoted F0V [243, 252] and comprises three main measures:

- Vibrato rate (F0VR)- the rate of fundamental frequency modulation (measured in Hz)
- Vibrato extent (F0VE) - distance between the peaks of each vibrato cycle (measured in cents or Hz)
- Mean fundamental frequency - the running average of the fundamental frequency of each successive vibrato cycle plotted over a moving time window for the duration of each individual tone

Accurate fundamental frequency vibrato rate (F0VR) and extent (F0VE) analysis relies on accurate measurement of fundamental frequency from the audio signal which must first be performed by automatic pitch detection algorithms based on Autocorrelation Function (ACF), Linear Predictive Coding (LPC) or taken from the electrolaryngograph signal (see section 4.3.5) Often vibrato in singing manifests itself as a modulation in intensity rather than phonation frequency, and similar measures of intensity vibrato rate and extent have been made, for example [14] (see Section 4.3.7).

In his extensive studies, Prame [243] analysed performances of Schubert's *Ave Maria* by 10 singers and found that the mean extent of the vibrato of individual tones was between ±34 and ± 123 cents. He also found a correlation between the length of tone and the vibrato extent with tones of shorter duration displaying greater vibrato extent.

Bowman Macleod [253], in a study of vibrato rate and width (extent) in string players, found that pitch height significantly affected the vibrato rate which the performer produced, with higher-pitched tones showing a faster vibrato (higher vibrato rate) than lower pitched tones; similarly she found that higher-pitched string tones had a larger vibrato extent than lower pitched notes.

Prame [254], in an earlier study of the vibrato rate of ten singers, found that the rate increased towards the end of notes, in the order of 15%. He found that average (mean) vibrato extent in individual tones varied between ±34 cents and ±123 cents and the average vibrato extent amongst singers was ±71 cents.

Perception of Intonation and Vibrato

Lynch, Eilers, Oller, Urbano and Wilson [255] estimated the Just Noticeable Difference (JND) for musically experienced listeners of mis-tuning within a melodic context to be around 10 cents, which corresponds to the often quoted JND of 10 cents for pitch-matching two tones [256]. Vurma & Ross [246] showed that melodic intervals can be more than 20-25 cents 'out of tune' (i.e. larger or smaller than the expected equal temperament interval size) before they are judged to be incorrectly tuned by expert listeners.

Intonation and the use of vibrato are closely interlinked, and many have argued that vibrato can be used primarily as a means to disguise errors in tuning. Yoo et al. [257] explored the effect of vibrato on the length of time required to determine the pitch relationship between two violin tones presented in succession and reported that listeners took longer to determine the pitch relationship of the tones when the second tone was presented with vibrato. They concluded that their data "may explain the commonly held belief among musicians that vibrato can be used to mask poor intonation" [257, p. 211].

Stowell [258] argued that vibrato used in violin playing is considered one of the most effective ways of adjusting tuning to the accompaniment, or other strings, in an orchestra.

Van Besouw et al. [252] investigated the range of acceptable tuning, as judged by listeners, for intervals comprising tones with and without vibrato, and found that for vibrato tones, the range of acceptable tuning was 10 cents greater than for tones without vibrato; for unmodulated tones the range of acceptable tuning was 24 cents in contrast to modulated tones where the range of acceptable tuning was 34 cents.

Van Besouw, Brereton and Howard [252] found that the range of acceptable tuning (the region where listeners judged tones to be "in tune") was 10 cents greater for vibrato tones in comparison to unmodulated tones. The range of acceptable tuning of tones in an arpeggio also differed according to whether the arpeggio was heard as a rising or falling sequence of notes.

Since the majority of vocal performance includes vibrato tones, the researcher must decide on a method of determining the principal pitch of any given vibrato tone. The "perceived principal pitch" is defined by Iwamiya, et al. as the overall pitch of the vibrato tone, which "has enough stability for cognition of a musical melody" [259, p.73].

Sundberg [260, 261] looked at vocal vibrato and its effect on perceived pitch using synthesized sung vowels. Participants were asked to adjust the pitch of an unmodulated tone to match that of the vibrato tone. He observed that most matches were within 5 cents of the linear average of the variation of the fundamental frequency over time.

Shonle & Horan [262] carried out a thorough study on the perceived pitch of vibrato tones, using a variety of stimuli with differing rates, extents, carrier frequencies and carrier types. They concluded that the principal pitch of a vibrato tone consisting of a sine wave

modulated by a triangular wave corresponded to the geometric mean between the extreme frequencies.

D'Allessandro & Castellengo [263, 264] studied the perceived pitch of synthesized short-duration vibrato tones, again using an adjustment procedure, paying particular attention to the initial phase characteristics of the modulating tone. They concluded that the perceived principal pitch reflected a weighted time average of the F0, but also that changes in frequency at the end of the tone had an influence on the perceived pitch. However, van Besouw & Howard [265] found the opposite effect of phase on the perceived pitch. Nevertheless, both of these studies show a perceived pitch of vibrato tones within six cents of the carrier frequency as well as the influence that the modulation tone phase can have.

Determining the principal pitch of a vibrato tone can also be achieved by using a perception based model of vibrato, such as that used by Devaney et al.[150] who incorporated an estimation of the perceived pitch of vibrato tones by using a weighted mean of the rate of change of the fundamental, which is based on perceptual studies by Gockel et al. [266].

4.3.12 Summary of Music Performance Analysis

Although quantitative analysis of music performance is now a rich research endeavour it should be remembered that performance itself is only one link in the chain between the composer's intentions, and listeners' perception. Indeed Gabrielson [193] in his first review of Music Performance Research reminds the reader that:

> *measurements of performance should, as much as possible, be conducted and considered in relation to the composer's and/or the performer's intentions and the listener's experience... After all, music is a means for communication and expression, and the characteristics of different performances may be easier to understand given this self-evident frame of reference [193, p. 550]*

In this thesis the quantitative analysis of the recorded vocal performances are examined in conjunction with the performers' own reported experienced differences between the performances, and analysed together with listeners' evaluation of the similarity between the recorded performances.

4.4 Room Acoustics and Musical Performance

Although already identified at the end of the sixteenth century as a key element of a performance, the effect of the acoustic environment on musical performance is only begin-

ning to be objectively studied in some detail. Nevertheless, for many styles of music the acoustic characteristics of the performance space form one of the most important aspects of musical performance which informing not only the choice of music to be performed in a venue, but also the creation of music composed with a particular performance space in mind.

The room acoustics of a performance venue not only affect the audience perception of a performance but also the performance itself. Modern day musicians are required to adapt their musical performance to suit a number of venues, from reverberant cathedral, carefully designed concert hall to the dry acoustics of the recording studio.

Recent investigations of concert hall stage acoustics from the musicians perspective have demonstrated that musicians adjust a number of performance attributes according to the acoustic environment in which they are performing [2, 39, 45, 36].

4.4.1 Room acoustics and musical style

In addition to the performance alterations made according to acoustic characteristics of a performance space, particular styles of music are suited more generally to different room acoustic conditions. For example, highly contrapuntal music with many independently moving musical lines will be blurred and muddied in a performance space with long reverberation times. The same long reverberation time, however, would suit slow moving polyphonic choral music or plainchant.

Throughout history, composers have been aware of the effect of the performance surroundings on the music which is performed, and as Thurstan Dart argued, shaped their music accordingly:

> *Plainsong is resonant music; so is the harmonic style of Leonin and Perotin .. Perotins music, in fact, is perfectly adapted to the acoustics of the highly resonant cathedral (Notre Dame Paris) for which it was written Gabrielis music for brass concert is resonant, written for the Cathedral of St. Marks; music for brass concert by Hassler or Mathew Locke is open-air music, using quite a different style from the same composers music for stringed instruments designed to be played indoors. Purcell distinguished in style between the music he wrote for Westminster Abbey and the music he wrote for the Chapel Royal; both styles differ from that of his theatre music, written for performance in completely dead surroundings. The forms used by Mozart and Haydn in their chamber and orchestral music are identical; but the details of style (counterpoint, ornamentation, rhythm, the layout of chords and the rate at which harmonies change) will vary according to whether they are writing room-music, concert-*

music or street-music. [267]

Blesser argues that Christian buildings such as monastries and cathedrals "determined the nature of music for a thousand years" [268, p.92] and this music was stylized and constrained by the unique acoustics of the spaces for which it was conceived and in which it was performed. Indeed, Bagenal reflects this hypothesis in his assertion that the acoustic characteristics of church buildings in themselves played a vital role in the development of German church music during Bach's time, since the building of galleries and boxes in St Thomas' Leipzig not only reflected the importance of the church as a building but also "created the acoustic conditions that made possible the seventeenth century development of Cantata and Passion" [269].

In similar vein, in a study of eleven Venetian churches, Howard & Moretti relate specific musical styles to different acoustic surroundings. They found the "largest churches poor for performance of complex choral music involving advanced polyphony and/or multiple choirs" whereas in the larger churches higher frequencies were strongly dampened, leading to low values for clarity and brilliance, which meant they were perhaps best suited to plainchant singing [32].

4.4.2 Importance of aural feedback for musical performance

Musicians react to the aural feedback provided by the acoustic conditions of the environment and alter their performance in accordance with their perception of how their own sound is being affected by the acoustics of the space [3].

Studies of ensemble playing have shown that *ease of hearing each other* is of prime importance to players in groups. Early work by Marshall demonstrated the importance of early reflections for effective ensemble playing [34]. In addition, not only for ensemble musicians but also for soloists, *hearing oneself* is also hugely important, since a circular auditory feedback loop exists which helps to coordinate perception and action in music performance [270].

In general three main aspects of the acoustic conditions on stage are important for musicians: the balance of direct sound to reflected sound on the stage, the level of "support" from early reflections, and the angles of projection of reflecting surfaces in the stage area [13].

A number of authors have examined the circular loop between the performer and the auditory feedback of sound they hear back from the room acoustic. Often it is the absence of such auditory feedback which alerts the musician to its existence. For example, trombonist Will Kimball was recorded for a study of instrument directivity [271] in an anechoic chamber and reports that:

> *When we play trombone, the majority of the sound radiates in a very directional way out of the bell (which is itself relatively far away from our ears), directly away from us, such that the sound we actually hear and adjust to when we are playing is primarily sound that has reflected off of some surface and then returned to our ears. So when there is nothing for the sound to bounce off of, as in an anechoic chamber, it is challenging to hear what you are doing! ... The tendency is to play louder and louder in order to hear yourself and try to create some kind or resonance. Dynamic shadings (part of what they measured in the study) are difficult because of the lack of aural feedback, and you end up going as much by the feel of your embouchure as by sound. [272]*

4.4.3 Musicians' Preferences for Room Acoustic Conditions

The acoustic conditions on stage (see Section 2.4.2) play an important role in influencing musicians' preferences for concert halls. A number of studies since Gade's early investigations have looked at musicians' preferences for stage acoustics. Gade [2] originally found that the *ease of hearing oneself*, due to the presence and make up of early reflections and the ease of *hearing others* were both highly correlated with musicians' preferences of acoustic support which resulted in the development of two widely used parameters for stage acoustic measurement: Support (ST) and Early Ensemble Level (Early Ensemble Level (EEL)) (see Section 2.4.3). Ueno et al. [48] have also found that the ratio of direct sound to early reflections had more effect on musicians' preferences for concert halls than overall reverberation time (RT60).

Nakayama [273] found that a solo recorder player preferred a single simulated early reflection with different delay times according to the tempo of the music being performed e.g. an early reflection at 35ms was preferred when playing at a faster tempo, but a reflection at 50ms was preferred when playing at a slower tempo. Overall a shorter delay time for the early reflection correlated with the musicians' impression of a lack of stage support.

Dammerud & Barron found that musicians preferred early reflections to arrive at 20ms and that musical performances were degraded when the early reflection was longer than 60ms [52]. Their investigations into concert hall preferences of orchestral musicians, involving a number of concert hall venues, showed the orchestra members' preferences corresponded to acoustical measurements associated with clarity. Similarly Ko et al. [274] found that measures of ST and C_{80} explained 68% of the preference ratings of musicians who played in enhanced stage acoustics.

Ueno & Tachibana [8] found that the magnitude of early reflections contributed to musicians' subjective impression of the size of the concert hall; the stronger the early

reflection the smaller the room was perceived to be, especially for wind players. Weak early reflections, on the other hand, led musicians to perceive the sound field as more reverberant.

Marshall & Meyer [275] investigated the acoustic preferences of a vocal quartet by synthesizing varied room acoustics and playing them back to singers in a hemi-anechoic room. The amplitude level of the early reflections seemed to have the most important influence on choir singers' preferences, although, in contrast the reverberation time did not seem significant. Singers disliked early reflections which arrived at 40ms delay but preferred early reflections between 15 - 35 ms and lateral early reflections were better liked than vertical ones.

Chiang et al. [108] looked at the correlations between the subjective parameters of *hearing oneself, hearing others, ease of ensemble*, and *overall impression* with the early-to-direct energy ratio and measures of ST_{late} and ST_{early}. Their findings show that performers' impressions of the concert hall correlated more strongly with late rather than early reflections and concluded also that chamber groups (small ensembles) might prefer stronger early energy than orchestra players.

In a study which asked performing musicians to compare eight different positions on a concert hall stage, Kim et al. [49] found for instrumentalists and singers that ST_late and RT were the most dominant factors in predicting preference ratings. Similarly Ueno & Tachibana [7] found that musicians' preferences for early and late reflections and reverberation time differed according to whether they were wind players, string players or singers. On the other hand Ko et al. [274] found no significant influence of gender, age, experience or instrument in the 20 professional musicians who rated enhanced stage acoustics provided by virtual acoustic technology, although all of these musicians were members of string quartets, so it is possible that brass or wind players may have had different responses.

Overall, musicians' preferences for stage acoustics seem to be mainly correlated with the balance between direct sound, early reflections, and later reverberant energy relating to overall reverberation time. This was articulated by Tom Beghin's assessment of virtual acoustic stage support offered to him as a player, who suggested "There's a triangle of listening: listening locally to the instrument, listening to the sound in the room, and listening to what the observer will hear" [13]. These three aspects of stage acoustics could be seen to correspond to direct sound, early reflections and reverberation time respectively.

Indeed a later study by the same team [274] which investigated string quartet members' preferences for virtual stage acoustics found that *stage support* and *clarity* explained most of the variance in subjective ratings. A principal component analysis of the ratings of a number of subjective parameters (such as *ease of hearing oneself, ease of hearing*

others, amount of reverberation, tonal balance, enjoyment of playing) found three main underlying dimensions to the data, namely: "tonal quality", "stage support" and "spatial attributes".

4.4.4 Influence of room acoustics on musical performance

Although there is a now a good body of work into musicians' preferences for stage (concert hall) acoustics, there is still much less objective research into the influences of room acoustic conditions on music performance attributes. Gade's seminal work on acoustic conditions as perceived by performers in concert halls [39] made an undeniably pivotal contribution to the field, demonstrating that the stage acoustics of the concert hall had an important effect on a musician's performance:

> *the room can be regarded as an extension of their instruments, through which they perceive the sound and quality of their own and co-players performance. Musicians adjust level, tempo, phrasing, timbre and intonation i.e. their means of musical expression – according to what they hear. [2]*

Studies in this area have until more recently been difficult to carry out due to the time and cost involved in asking musicians to perform in a number of different concert hall venues. What is more, the amount of data collected when recording a number of musical performances is time consuming to analyse.

However, there is now a growing research field of Music Information Retrieval, which is developing automated techniques for extracting data about recordings of music. Part of this work involves the development of new techniques and algorithms for Audio Content Analysis which help to automate the process of extracting music performance parameters and analysing performance attributes [158, 135].

This together with developing capability in the area of collecting and analysing "big data", facilitating the analysis of large sets of complex data, through growing use of machine learning techniques means that music performance analysis can be more easily achieved on larger data sets than was previously possible.

At the same time investigations into how music performance changes in different acoustic environments have been increasingly facilitated by the use of virtual auditory environments and/or room acoustics simulations. For example recent work by Ueno, Kato and colleagues [48, 107, 3, 15, 9, 15, 14] have utilised a virtual acoustic simulation of a number of performance venues in order to investigate how musical performance changes according to room acoustic conditions.

4.4.5 Evaluating room acoustics through analysis of performance

Brunskog et al. [53] examined the effect of different room acoustic conditions (in real rooms) through the analysis of vocal output and they related changes in voice sound power to the objective room acoustic parameters and the speakers' subjective impression of the room. Ternström [206] has analysed choral performances by different choirs in different room acoustic conditions and found that performance parameters changed in the different environments. These findings and similar studies will be discussed further in Section 4.6.

Kato et al., [15, 14] have analysed performance parameters of musicians playing in real-time room auralizations, and correlated this with feedback from the musicians about how their playing varied according to acoustics of the simulated performance venue. Through listening tests and statistical analysis they concluded that the differences in performances in the various simulated room acoustic conditions were significant and perceivable. Changes were found in terms of tempo, vibrato rate and extent, and sound pressure level between the simulations. They also found spectral variations, in that flute and oboe players suppressed the higher harmonics of the instrument tones when playing in a simulated reverberant room. In addition, in reverberant conditions the violin and oboe players considerably decreased the length of notes, whilst increasing the length of silence between notes (i.e., playing more *staccato*).

Whilst Kato, Ueno and colleagues have undertaken thorough investigations of musical performance and room acoustics using novel VAE techniques, they conclude that further work still needs to be carried out in this field.

> *A parametric investigation into acoustic factors that have a dominant effect on the musicians performance remains to be done. This could be in the form of part of a future study that utilizes parametrically synthesized room impulse responses. [3]*

Two recent studies by Kalkanjiev & Weinzierl [10, 276, 11] looked at the effect of room acoustic conditions on a solo cellist's performance. One study [10] analysed recordings made in a number of (real) concert halls and through statistical analysis of the correlation between the numerous room acoustic parameters measured, they identified four main room acoustic parameters (RT60, ST_{late}, G_e and Br) (See Section 2.4.2 for definitions) which could predict changes to seven performance parameters; tempo, agogic, loudness, long-term dynamics, short-term dynamics and timbre.

In a later laboratory-based study [11] which used binaurally rendered room acoustic simulations of real concert hall venues, the authors investigated the same room acoustic parameters and their correlation with performance attributes of recordings made by two

cellists. Again they found that four main acoustical parameters (EDT, RT60, G_e and Br) characterized 97.5 % of the acoustic variance between the concert hall simulations.

Although there is still only a small amount of research seeking to quantify the ways in which musicians alter their musical performance according to the acoustic characteristics of the performance space, it is an area of increasing interest. A number of musical performance attributes have been found to alter between performances undertaken in different acoustic environments as summarized below.

- **Intensity-related attributes**
  - sound pressure level [14, 137, 11]
  - intensity vibrato extent [14]
  - long-term and short-term dynamics [14, 10]
  - loudness [10, 12]
  - hammer velocity and sustain pedal [137]
  - dynamic bandwidth [11, 10]
- **Timbral attributes**
  - strength of higher harmonics [15, 277]
  - timbral bandwidth [10]
  - timbre (soft-hard, dark-bright, lean-full) [11, 10]
- **Temporal attributes**
  - note-on-ratio [15, 137]
  - tempo [137, 14, 10, 12, 11]
  - synchronicity (temporal alignment in ensemble playing) [12]
- **Tonal attributes**
  - fundamental frequency vibrato rate [14]
  - fundamental frequency vibrato extent[14, 3]

The following sections summarises the main work in this area focussing on findings relating to Reverberation Time and Early and Late Reflections.

Reverberation time

Kalkandjiev and Weinzierl found that reverberation time influenced many of the seven performance parameters of solo cello playing which they investigated [10] (See also section 5.3 for more on this study).

In a study facilitated by the recent availability of the MIDI Yamaha Disklavier, which allowed MIDI data (such as note played and hammer velocity) to be recorded during the performance, Bolzinger and Risset [137] analysed piano performances in a room with adjustable acoustics. They found that with an increase in reverberation time, pianists played softer (lower hammer velocity), more staccato (lower *note-on ratio*) and more slowly, with less use of the sustain pedal. It was hypothesised that these changes were due to the room acoustic conditions, concluding that acoustic feedback in a performance venue clearly influenced the playing intensity, so that "in most cases, a dull concert hall will require a greater physical effort from the interpreter" [137, p.136].

> *We then observed that all these above parameters, except the average tempo, are directly correlated to the sound intensity produced. As the change in reverberation relates to these changes, we can assume that they serve a purpose of compensating the changes in the room acoustics* [137].

This study found that for pianists in an acoustically dry environment the greater physical effort required of the performer means that hammer velocity, note duration and intensity level are all increased.

Playing more slowly in reverberant performance spaces is a recognised strategy for musicians, stemming from a desire that the individual notes of the pieces are not"blurred together" by the long reverberation time. On the other hand, when the reverberation time of the venue is shorter the musician may attempt to lengthen notes (higher *note-on ratio*) since the tones are not 'carried' to the listener by the reverberation characteristics of the room.

Perhaps counter-intuitively Kalkandjiev and Weinzierl [10] found that both short and long reverberation times led to a slower performance tempo, with moderate reverberation times leading to the fastest tempi.

Ueno et al.[3] also found that there was no direct linear relationship between Reverberation Time and normalized tempo (n-Tempo). n-Tempo was slower for the most reverberant condition, but results varied according to the musical motif played. For one excerpt ("Ave Maria" by Schubert) tempo was quicker in the smaller simulated halls, but slower in the anechoic chamber, where tempo was similar also to performances made in the longest simulated reverberant condition of CH (Church). The authors suggest that in both the large church and the anechoic chamber musicians played more carefully,

perhaps because of the lack of early reflections lending a sense of "support". More careful playing could lead to a slower tempo. The results of this study, whilst not showing a simple relationship between RT and music tempo, have been shown to be statistically significant and a complementary listening test also verified that listeners could perceive the variations that musicians claimed they made in performance.

A violinist and oboist reduced the *note-on ratio* (played more staccato) under more reverberant conditions [15] in simulated acoustics of different concert halls. Kato et al. [15] found that higher harmonics of a flute and oboe players output were suppressed in reverberant conditions; for the flute player this agreed with her assertion that she played more softly in the simulated reverberant hall which acoustically leads to a decrease in the level of higher harmonics.

In a further analysis of the same performance recordings Ueno et al. [3] found that fundamental frequency vibrato extent (F0VE) was larger in the anechoic room and medium hall, but smaller in the small and large hall simulations. Differences between fundamental frequency vibrato extent were found to be statistically significant between the majority of simulated room acoustic conditions.

Early and Late Reflections

Early reflections produce, in effect, a time delay in auditory feedback which can lead to problems in timing and maintaining steady tempo. Chafe et al. [114] investigated the influence of time delay in auditory feedback by manipulating the time delay between source and listener sounds in pairs of musicians who were asked to clap a rhythm in synchrony. Longer delays produced a deceleration of tempo, as each performer waited for the other, whereas moderate amounts of delay were beneficial to the musicians in terms of keeping a stable tempo.

Similarly, Kalkandjiev and Weinzierl [10] found that lower values of G_e, related to a decreased level of early reflections, leading to increases in performance tempo in their study of a solo cellist.

In contrast Woszczyk et al. [12] found that moderate levels of early reflections led to increased tempo in a violin duet. Similarly the presence of simulated early reflections improved synchronicity between the two players and also lead sometimes to increase sound pressure levels. Additionally, late reflections helped violin duet players to maintain good intonation. Ueno et al. [48] also found that early reflections increase the ease of hearing other members of the musical ensemble and strong early reflections increase timing accuracy between musicians.

Although two studies [137, 14] found a negative relationship between loudness and reverberation time, Kalkandjiev & Wienzierl [10] found instead a more complex picture.

The cello soloist in their study tended to reduce loudness in room acoustic conditions which offered good early and late acoustical support, i.e. higher values for G_e and ST_{late}, which was shown to lead to a perception of *increased reverberation*. Whilst it makes sense that a soloist can play more quietly when the stage support is good, it was also found that increased reverberation time resulted in production significantly higher in level. Since this finding is in contrast to other investigations, Kalkandjiev and Wienzierl surmise that responding and adjusting to different levels of stage support may be a learnt technique which varies between individual musicians with different levels of training or experience.

In the same study it was found that short term dynamics were not influenced by levels of stage support, but the bandwidth of long-term dynamics was reduced when early support (G_e) was high. It is interesting to note however that Kalkandjiev & Weinzierl found conflicting results in their two studies involving cello soloists: in the real concert halls [10], they found that G_e was negatively correlated with tempo, whereas in the study using a virtual room acoustic simulation [11] the opposite correlation was found.

They also found that, in the regression model used to analyse relationships between subjective responses and objective parameters, *timbre* was predicted by values of ST_{late} with an increased amount of perceived reverberance leading to a "harder and brighter" tone with more defined attack in articulation. The acoustic parameter Bass Ratio (BR) was also found to influence the timbre of the cellists' playing with a higher BR value leading to "darker" and "softer" playing for both performers in all pieces [11].

4.4.6 Conceptual models of performance

Woszczyk & Martens provided virtual stage acoustics and undertook a performance case study with harpsichordist Tom Beghin, where the virtual acoustics simulated the real acoustics of a performance space known to the player. The harpsichordist reported that " I am playing the room, just as much as I am playing the instrument, the room is attached to my instrument, they are one and respond together" [13].

Ueno & Tachibana [107] posited a model of the relationship between a musician and the acoustic environment and how it impacts on the musical performance produced. Through interviews with musicians, they applied the theory of "tacit knowing" to a musician's experience of playing in the concert hall.

Ueno et al. state that they have determined a "circulative system of feedback between performer and room acoustic conditions" [3, p.513] and similarly Ueno & Tachibana [107] hypothesize that there are indeed two types of circular feedback systems in operation when a musician plays. These two feedback systems are part of the conceptual model of performer and listener during music performance in a performance space, reproduced as Figure 4.6.

One feedback loop is observed as an automatic response system common to human action, perception and response. The other is described as an "acquired feedback system" which relates more directly to the musician's skill, background and experience. In this system the musician perceives the concert hall acoustic and musical expression, and forms an image of how the performance will be perceived by the listener in the audience area, which is then subsequently used to control his/her performing actions.

This is similar to the "triangle of listening" described by harpsichordist Tom Beghin in [13] where the performer identified three aspects of listening: listening to the sound of the instrument (direct sound) to judge the articulation of the performance; listening to the early reflected sound to judge the effect of the room; listening to the later reverberation to judge what the audience will hear. This performer felt strongly that in real or simulated room acoustics "the instrument, performer and room become triune entity (sic)" [13, p.1044].

4.5 Room Acoustics and Speech

Perhaps one of the most important aspects of speech as perceived by listeners in a room is the intelligibility of what is being said. Speakers must be aware of the distance between them and their listeners and adjust their vocal output accordingly. Pelegrín-García et al. [278] found that speakers not only adjusted their vocal effort according to the listener distance but also changed other vocal parameters according to the room acoustic environment. For example, average fundamental frequency was increased by 4Hz on average in an anechoic room and vocal intensity was increased as the distance to listener increased, but was lowered as Room Gain (Room Gain) was increased. (For details of Room Gain and related calculations refer to Section 2.4.3).

Pelegrín-García et al. [278] found that speakers lowered their own voice level by 3.6dB per 1 dB increase of Room Gain. It is similar to the traditional measurement of Strength) (see Section 2.3.3) which is calculated as the ratio between the sound energy measured in a performance space and in a free field (anechoic) measured with the same impulse response measurement apparatus of loudspeaker and microphone, rather than head and torso simulator.

Poor room acoustics have been shown to impair the vocal function of those who have to perform or speak in them for long periods of time, e.g school teachers or college lecturers [279, 280]. Similar problems are also reported by professional and amateur singers alike and often problems with vocal function can be exacerbated for singers as they adapt to a wide variety of room acoustic conditions on a regular basis [281].

Kob et al. [279] investigated the effect of poor room acoustic characteristics in teaching

rooms. The room acoustic parameters of four classrooms were measured and the voice quality of teachers teaching in these classrooms was evaluated over the course of a working day. They found significant differences in many of the voice parameters recorded. For example, voice pitch decreased to a lower level after teaching in the acoustically more favourable rooms, whereas the voice pitch of a read text became less variable in the acoustically bad classrooms, indicating a more monotonic vocal delivery. Voice quality improved after teaching in the favourable acoustic conditions. Voice output level was not found to have any statistically significant correlation to the acoustics of the classrooms.

Howard & Angus [282] offer voice users advice on how to avoid vocal impairment when speaking in detrimental room acoustic conditions. They lay out the aspects of room acoustics that are important if a speaker is to avoid vocal problems which include having local diffuse acoustic support with no strong discrete reflections, which might affect speech comprehensibility. They also recommend that the room has sufficient diffuse early sound and long enough reverberation time to enhance the perceived loudness of the voice.

More recently, in a study on teachers' voice use in different sized class-rooms, Åhlander, Pelegrín, Whitling, Rydell and Löfqvist [280] remarked that the large classrooms (sports halls) in their study had lower "Voice Support" (STv) values, but also that the frequency responses of these large halls were different to the small and medium-sized classrooms in that low frequency energy predominated and high frequencies were not reflected well enough for the teachers' speaking comfort

4.6 Room Acoustics and Singing Performance

Whilst the previous section has given an overview of the changes in musical performance when influenced by concert hall acoustic parameters, this section focusses in more detail on the variations in vocal performances in particular. It has been well known since renaissance times that singers modulate the use of their voice according to the surroundings in which they perform. Indeed in the sixteenth century Zarlino wrote that:

> *one sings in one way in churches and public chapels and another way in private rooms. In [church] one sings in a full voice and in private rooms one sings with a lower and gentler voice, without any shouting* [4].

In more modern times, professional singers, asked to perform in a variety of venues, have to constantly adapt many aspects of their singing, both during and between performances. However, there has as yet not been any systematic or substantial research to quantify aspects of singing performance and how they relate to the room acoustic conditions of performance venues.

4.6.1 Singers are special

It has been shown that musicians alter their musical performance according to the auditory feedback they receive, which has in effect been "coloured" or "processed" in different respects by the volume, shape and fabric of the concert hall or performance space. Auditory (aural) feedback is important for a singer to maintain control of the voice and will be discussed further in 4.6.2. Furthermore, for singers, the link between performance and room acoustic is further enhanced, not only due to the role that the room plays in shaping the auditory feedback, but also because the instrument (the voice) is an integral part of the singer. Indeed, Blesser refers to singers as "aural detectives" as they constantly, both consciously and sub-consciously adapt their vocal performance to the instantaneous auditory feedback they receive, adding that "singers investigate the acoustics of a room the way a child investigates a toy" [268].

4.6.2 Aural Feedback For Phonatory Control

A singer (or speaker) receives continuous feedback in order to regulate phonation and singing/speech production through three main means: kinaesthetic feedback from the larynx, head and chest; auditory feedback from bone conducted sound and auditory feedback from air conducted sound.

When auditory feedback is reduced for some reason, singers must rely on the other feedback mechanisms for phonatory control. Mürbe et al [162] have shown that when auditory feedback is reduced, singers are more likely to experience difficulties with pitch control especially when there are large intervals between the sung notes, singing staccato notes or singing at a quick tempo.

Chang et al. [283] recently investigated the sensorimotor cortical network which underlies the control of vocal pitch as part of the auditory-feedback loop pinpointing the neural mechanisms involved in the control of pitch. When the spoken pitch of the subject's vocal output was perturbed (lowered by 200 cents) the subject was found to increase pitch rapidly, after a delay of only 170ms, to compensate for the lowered pitch of the auditory feedback. Similar vocal control mechanisms have also been shown to be at work during singing tasks [284], although it is hypothesized that singers use different pitch control strategies compared to non-musicians due to their training and experience of close monitoring of auditory feedback, leading to increased auditory activity during the pitch-compensation tasks.

In an anechoic chamber the aural feedback for the musician is greatly reduced, since there is no reflected sound from the room. This means that the singer has to rely more heavily on the direct sound (mouth to ear) produced, as well as bone conducted sound

and kinaesthetic feedback from the ability to sense the position of ones own body position, movement and muscular tensions (proprioception). Using these other types of feedback mechanisms can help a singer maintain intonation accuracy when auditory feedback is masked or at low level, but the singer needs to adjust to the new balance of aural and physical feedback. Bone and tissue conduct low-mid frequencies in the voice most easily to the inner ear, whilst direct air-borne sound from the mouth to the ear is diffracted, meaning that higher frequencies are attenuated - both these phenomena result in the singer hearing a low-pass filtered version of their own singing voice. With the absence of any reflected sound 'filling' in the higher frequency components, the effect of hearing one's own voice in the anechoic chamber in this way will influence the resulting vocal production.

4.6.3 Singing and Listening

It should be noted that there is a masking effect associated with producing sound and listening at the same time. Although few empirical studies have been carried out in this area, recently Borg et al. [285] found that the threshold of masked noise was lower at higher frequencies than lower frequencies. For example, whilst vocalising at 70dB SPL, the threshold for the speaker's own sound to mask narrow-band noise at 250 Hz was 20dB below the vocalisation level, whereas when vocalising at 80dB SPL the threshold was increased to 30dB below. Differences in masking thresholds were found between female and male speakers, and masking was also found to be dependent on the central frequency of the narrow-band noise.

The masking effect of a singer's own voice is stronger for lower frequencies, due to the effects of bone conduction, which acts rather like a low-pass filter and also because higher frequencies are radiated from the mouth in a highly directional manner. So, there may be differences in a singer's perception of stage acoustics in comparison to instrumentalists, due to the masking effect of the singer's own vocalization, and the relative levels of auditory feedback from direct sound, room reflected sound and bone conducted sound.

When choral singers cannot hear the sound of their own voice sufficiently, either because of the room acoustic conditions, or because of the masking effect of other choir members, they tend to raise the loudness level of their voice in order to gain a satisfactory amount of aural feedback. Ternström [286] investigated the ability of choir singers to hear their own voice, and allowed them to adjust the ratio between their own sound and that of other members of the choir in order to understand preferences for Self-to-Other Ratios (SOR) amongst choral singers with the average SOR of +3.9 dB, and preferred values ranging from +1.5dB to +7.3dB.

In an earlier study Ternström found that when there were large differences between the

sound level of self and other, or the spectral properties of "Other" were unfavourable, the intonation of the singing ensemble suffered [250]. Of course, both of these characteristics can be influenced by the room acoustic conditions of the performance venue.

4.6.4 Singing Performance in Different Room Acoustic Conditions

This next section summarises the small amount of empirical research that has been undertaken to quantify the effect of room acoustics on singing voice performance.

Kim et al. [49] found that ST_{late} and RT correlated strongly with vocalists preferences for concert hall acoustics, but also that visual impression of the space had an influence on vocalists subjective ratings of concert halls.

Temporal Attributes

Kato et al. [14] had a number of different instrumentalists and one singer perform in a real-time room acoustic simulation of different concert hall and church venues. Analysing the acoustic signals of the musical performance they found that n-Tempo (normalized tempo) for the baritone was slower than the average across all performances when singing in the larger concert halls and church. On average the performers in this study played 4.4 % faster than average in the medium hall, and 7.1 % more slowly in the reverberant church condition.

Tonal Attributes

Vibrato Kato et al. [14] found that fundamental frequency vibrato rate (F0VR) and extent (F0VE) differed for the baritone singer in the five different room acoustic conditions investigated. Although F0VR seemed to be less affected by the room acoustics, with no significant adjustments made, F0VE on the other hand was altered significantly between the different (simulated) venues with a range of ±93 cents between venues and vibrato extent reducing with more reverberant surroundings.

Intonation Ternström & Sundberg [245] found that the acoustic characteristics of the musical note to which a singer is trying to tune, can affect the ease of intonation. The overall characteristic of sung sound can be altered and affected by the room acoustic conditions, for example, the stage acoustics of the performance venue might have longer reverberation time in the mid to high frequency region of the spectrum. The aural feedback that a singer receives, and the interrelation of self-to-other sound level will lead the singer to alter their sound level and voice usage accordingly.

Portamento Blesser [287] argues that portamento became fashionable during the 1930s at the same time as concert halls were designed to be more and more acoustically absorptive. He argues that the increased use of portamento ensued because of a desire of singers to "join" together notes which had started to sound separate due to the dry conditions of fashionable concert halls or acoustically treated recording studios. Indeed, in a study of historical recordings, Timmers [214] found that the number of pitch glides in a recording of the same piece in different years increased towards the 1930s and then decreased again after the 1940s.

Intensity-related Vocal Attributes

Kato et al. [14] found that a baritone singer made statistically significant adjustments to intensity vibrato extent in five different room acoustic conditions, with larger intensity vibrato extent in the more reverberant space.

Timbral Attributes

Long-term Average Spectrum Ternström [206] studied the long-term average spectral characteristics of choirs singing in rooms with different room acoustics. He found that singers use less "power" in a reverberant acoustic, such as a concert hall. He also found some changes in the spectral content of singing in different room acoustics. In an acoustically absorbent room, the boys' choir in particular reduced the amplitude of the fundamental frequency partial in relation to the rest of the vocal spectrum. In more absorptive rooms, singers adjusted the formation of vowel sounds so that in general formant frequencies were higher than when singing in the more reverberant spaces. Such differences in formant frequencies are seen to be unattributable to increased sub-glottal pressure as found in pressed phonation, and so must be due to singers' alteration of vocal tract configuration. The choir identity, musical nuance, piece of music, and room acoustic were the independent variables in the study, and room acoustics and musical nuance influenced the LTAS of the choral sound more than the other variables. Ternström concluded that the choirs in the study adapted their voice usage according to the room acoustics of the three different locations.

Vocal Production-related Attributes

An investigation into the use of HearFones [288] (headphone-like devices with plastic reflectors to reflect some of singer's own sound to their ears) suggested that the increased perception of sound energy around the range of the first formant and at frequencies above

4kHz lead to an increase in closed-quotient, as well as a lower Lx signal amplitude, both of which may have stemmed from more complete and regular glottal closure.

The next section describes a case study undertaken with a vocal quartet who were asked to sing in three different acoustic configurations of the *real performance space* and the anechoic chamber, in order to provide source material for the pilot listening tests described in Section 5.4 and 5.5. Subjective responses of the singers to singing in the different acoustic conditions were also collated in order to help inform the acoustic analysis of solo singing performances recorded later in the VSS (See Section 6.2)

4.7 Case Study I: Quartet singing in the *Real Performance Space*

A quartet of singers performed in the *real performance space* (The National Centre for Early Music, York [289], see Section 3.3. The quartet (Soprano, Alto, Tenor, Bass) sang three pieces by Thomas Tallis (c.1505 - 1585); *Audivi Vocem* - an anthem in Latin, *Remember not, O Lord God* - an anthem in English and *Fond youth is a bubble* - a secular piece in English. (For reference, copies of the scores are found on the data CD accompanying this thesis, see A). Each piece was recorded by the quartet in each of the three different configurations described in Section 3.3.1 above.

The singers each wore a head-mounted DPA 4066 microphone, at a distance of 10cm from the mouth, away from the air stream in order to capture the singer's individual sound (although with some acoustic leakage from nearby singers). A B-format recording was also made with the Soundfield microphone positioned in the same location as the listener receiver (Figure 3.10) used to record the Spatial Room Impulse Response of the venue (see Section 3.4.1).

In order to allow "virtual acoustic recordings" [44] for later evaluation (see Section 5.5) where the direct sound of the singer is convolved with the previously recorded SRIR of the performance venue, recordings of the pieces were also made by the same singers in the anechoic chamber using the same microphones and recording equipment.

Number	Activity	Acoustic Boxes	Drapes	Acoustic Qualities of Space	Reverberation Time (s)
1	Large choral (LC)	All boxes closed	All drawn back	Highest reverberation, warmth, spaciousness	2 to 2.6
2	Music recitals (MR)	All boxes closed	All drawn out	Even balance between clarity and reverberation, discrete sounds stand apart clearly, but ample reverberation	1.5 to 1.8
3	Lectures and speech (SP)	All fully open	All drawn out	Sound absorbent space, giving maximum clarity for speech	1.0

Table 4.2: *Summary of three acoustic configurations used in the real performance space*

4.7.1 Quartet Singers' subjective responses

The members of the vocal quartet were asked for their subjective impressions of singing in the three different room acoustic configurations in the *real performance venue* as well as in the anechoic chamber. Three of the four singers responded to this request and their responses are summarised below:

Impression of the space

- dry acoustic (SP) more intimate - Sop
- MR supported sound but allowed *hearing oneself* and *hearing others* - Alto
- MR allowed best blending of the sound - Alto, Bass

Intonation

- tuning best in the dry acoustic (SP) - Alto and Soprano
- tuning best in the most reverberant acoustic (LC) - Bass
- intonation very difficult/impossible in the anechoic chamber - Alto/Bass

Timing/synchrony between voices

- most reverberant setting (LC) was easiest - more "space" for error - Alto
- no difference - Bass

Pieces

- "Remember not" (homophonic) worked best in MR and LC - but lacked in the dry setting - Alto
- "Remember not" (homophonic) better in reverberant LC - Bass
- "Fond Youth" better in dry acoustic (SP) - Bass

Overall preferences

- SP - Soprano
- MR - Alto
- LC - Bass

Summary

This section outlined the subjective impressions of a quartet of singers (soprano, alto, tenor, bass) who sang in the adjustable room acoustics of the *real performance venue*. All singers reported changes to the ease of synchronizing between voice parts, maintaining stable intonation and noted the differing levels of *support* offered by the different configurations. There was no obvious pattern of overall preference for the dry, medium or reverberant conditions, with singers disagreeing on the most preferred acoustic apart from their unanimous dislike of singing in the anechoic chamber.

4.8 Summary

Although there is now a reasonable body of research on the ways that musical performance changes in different acoustic environments, there is still little empirical research which investigates vocal performance in different acoustics in particular. However, there is growing interest in this area, which is facilitated by the use of audio content analysis techniques, and virtual acoustic technology to recreate room acoustic conditions in the "lab". A number of authors have found statistically significant correlations between room acoustics parameters and the analysed music performance attributes of music performances.

Findings of previous authors can be summarized as follows:

Reverberation Time When RT60 is longer:

- Timbre is "lean" (fewer higher harmonics) [10, 14]
- Tempo is slower [14, 10, 8].
- Vocal power is decreased
- Relative amplitude level of fundamental frequency partial reduced [206]
- Loudness is decreased [14, 137, 137].
- Fundamental frequency vibrato extent (F0VE) is smaller [7, 14]
- Notes are shorter (lower *note-on-ratio*)[137]

Early Strength (G_e) When Early Strength is higher

- Timbre is "fuller" (increased energy in higher harmonics) [10].
- Tempo can be faster or slower [11, 10].

- Loudness is decreased [10].
- [CORRECTION] Range of long-term variation in dynamics reduced [10]

Support ($\mathbf{ST}_{late}$ When late Support is higher

- Tempo is slower [10].
- Timbre is "harder" [10]

Musicians' preferences and adjustments to room acoustics are seen to be affected by the balance of early to late arriving sound energy, which has been shown to correlate with measures of early and late Stage Support.

Research undertaken by others in this area (e.g.[48, 3, 11, 10]) has shown that there are variations between musicians, and alterations in performance are more subtle between venues with smaller differences in room acoustic characteristics. Nevertheless, there are a number of vocal performance attributes that have been either reported by singers themselves, or have been analysed by researchers and shown to change between performances in different acoustic environments.

Chapter 5 describes methods for the subjective analysis of audio which are also useful for the subjective analysis of musical performance. It outlines some of the methods available for statistical analysis of perceptual data on preferences and similarity ratings and presents the results of two pilot listening tests which were undertaken to test methodology and listening test design in order to inform the main listening test as described in Chapter 6.

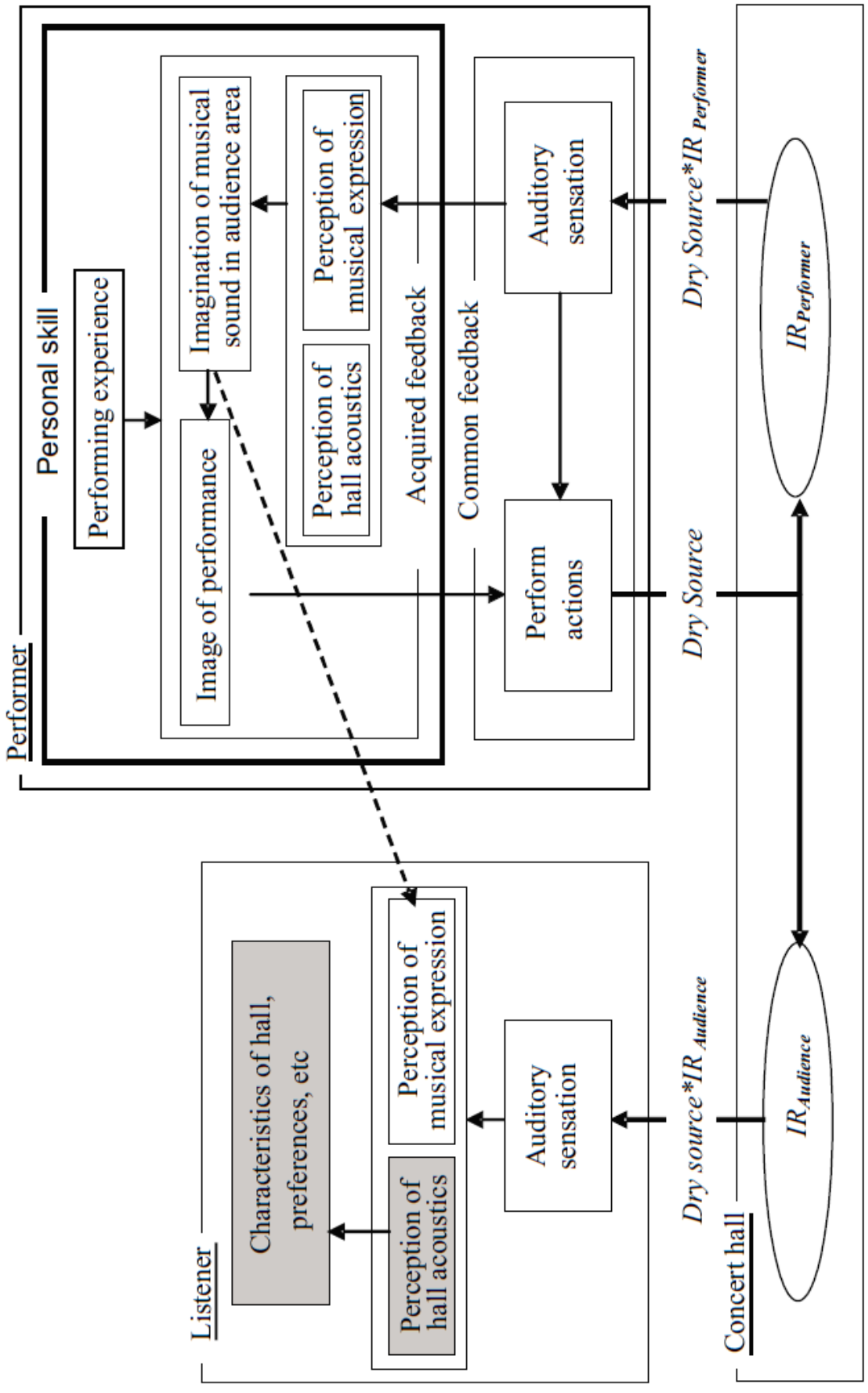


Figure 4.6: *Schematic model of a performer and a listener in a concert hall, from [3]*

Chapter 5

Singing Performance Analysis and Evaluation

5.1 Introduction

Section 5.2 examines how audio in general, and musical performances in particular, can be evaluated by listeners using a variety of perceptual testing methodologies such as rankings, similarity or preference ratings, and gives a brief overview of the statistical methods regularly used to validate such testing methods.

The next section (5.3) outlines some of the recent music performance analysis research, which has tried to identify correlations between music performance features and listener's evaluations using statistical techniques such as linear regression and correspondence analysis.

Two pilot listening tests which were undertaken to test the design and methodology of the main listening test (Chapter 6) are described in Sections 5.4 and 5.5.

5.2 Perceptual Evaluation of Audio and Music

Perceptual listening tests complement any objective analyses of audio or music performance and aim to quantify perceptual parameters or confirm objective results. A number of methodologies for evaluating audio and musical performances are encountered in the literature, and the techniques, application areas and statistical analysis methods differ between them.

Bech & Zacharov provide a thorough handbook on the theory of, and practical techniques for, evaluation of perceptual audio qualities in their book "Perceptual Audio Evaluation" [290], which includes a comparison of some of the different methodologies used

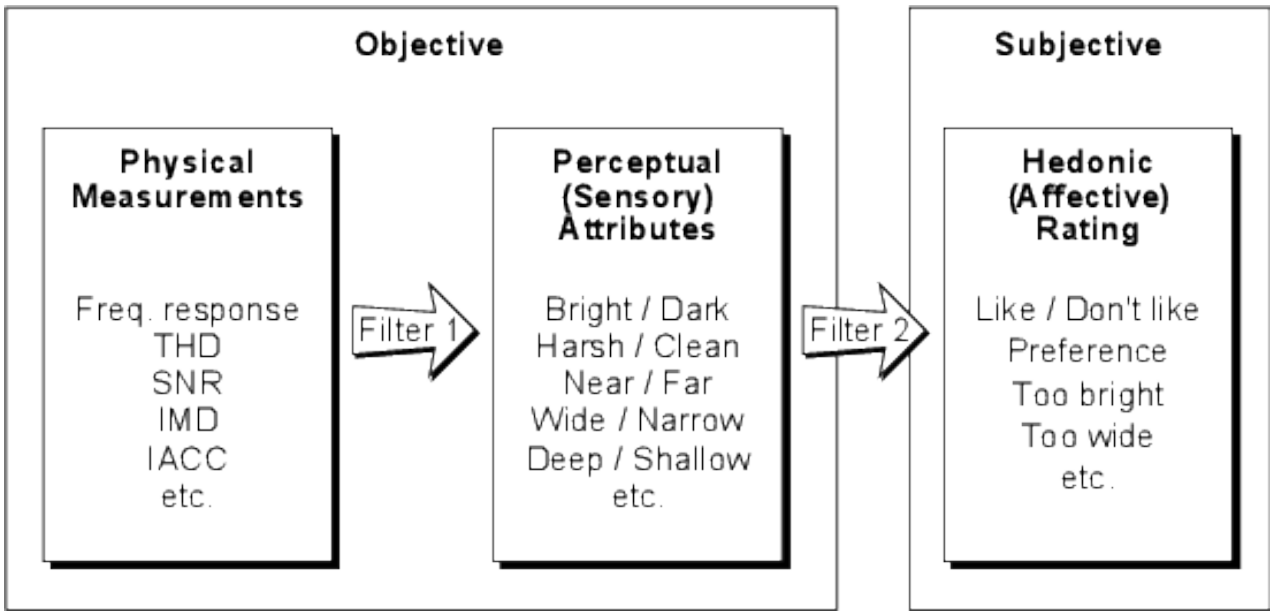


Figure 5.1: *The filter model of the relationship between objective and subjective aspects of sound: from [291] adapted by Martin, original by Fog & Pederson [292]*

in perceptual evaluation tests in general, and in psychoacoustic evaluation in particular. The next section outlines some of the testing strategies which were considered for the present study, comparing at their relative advantages and disadvantages.

5.2.1 Psychoacoustic Evaluation Methods

There is now a strong tradition of using listening tests to evaluate audio and many studies of this kind are concerned with timbral qualities of audio arising from the use of various codecs, processing techniques, or differences in equipment for audio presentation or storage.

The "filter model" illustrated in Figure 5.1 was first posited by Fog & Pederson [292] who draw a distinction between "Objective Measures" and "Subjective Measures", although "Objective Measures" are further broken down into *physical measurements* and *perceptual (sensory) attributes*.

The filters between these sets of measurements describe the way that sound is translated by the human nervous system from physical attributes (e.g. Signal-to-Noise ratio, RT60, Frequency Response etc.) into perceptual attributes such as "bright/dark", "wide/narrow" by Filter 1, which refers to sensory sensitivity and selectivity. Filter 2, in turn, determines how these perceptual attributes are translated by human cognition into a judgement of quality or preference, and is heavily influenced by a person's background, expectations, interests, emotions, and mood.

The distinction between these two sets of measurements has already been encountered in Section 2.4 where objective physical room acoustic parameters (such as RT60) were seen to have corresponding "subjective" room acoustic parameters. Confusingly, many authors denote "perceptual attributes" as "subjective attributes/parameters" whereas Fog & Pederson are clear that subjective measures are those which concern hedonic, or affective, ratings and are couched in terms of preference and liking.

A small number of suitable perceptual audio evaluation methods appropriate for this research are outlined below followed by an overview of the statistical methods used in conjunction with these methods.

Scaling methods

Direct scaling A listening test can simply ask listeners to rate a particular audio sample on a rating scale (e.g 1 to 7) in terms of a specific attribute e.g. *"timbral brightness"*. Very often perceptual attributes are presented in bi-polar constructs, for example "dark - bright", "narrow - wide", or "full - lean".

Indirect Scaling Indirect scaling methods ask listeners to compare two sounds, which are usually played successively, and then rate one against the other e.g. "which is the louder of the two sounds?". Similarity/dissimilarity ratings can be subsequently be elicited through the use of such paired comparisons.

Paired comparisons

Pairwise (or paired) comparisons are most thorough if each sample is presented in a pair with every other sample, and to ensure that no effect of ordering within the pair might bias the results, each paired sample should be presented twice with the ordering of the samples reversed in each presentation. Such rigorous treatment in a pairwise comparison test often leads to a very lengthy test procedure, dependent on the number of samples presented, so again care must be taken to balance careful testing, with provision for avoiding listener fatigue.

Similarity ratings

If the goal is to have the listener assess the similarity/dissimilarity between two pieces of audio, paired comparisons may be undertaken where the participant is asked directly to rate the similarity between a pair of stimuli, for example on a scale from 1 ("Not at all similar") to 10 ("Very similar").

"Subjective Clustering" is another possible technique, where the participant is asked to sort the stimuli into groups according to similarity. Alternatively "derived measures" can be used where participants are asked to evaluate stimuli on bi-polar scales. However, where the experimenter wishes to stay within the "spirit" of using a decompositional approach where attributes are not specified (see Section 5.2.2), this latter technique is not appropriate as the use of bi-ploar scale will already guide the participant to attend to specific aspects of the sound.

Clustering methods

Instead of asking a participant to rate preference or similarity on individual attributes of audio samples, it is often more effective and less time consuming to use a subjective clustering technique, especially when a large number of stimuli are involved, as traditional pairwise comparison methods can become lengthy and lead to fatigue of the participants. Such techniques allow the participant to audition all stimuli, in any order, and place icons representing the audio samples at relative positions on a 2-dimensional grid.

Scavone developed a freely available software [293, 294] - "Sonic Mapper - which allows multiple scaling of audio fragments directly onto a 2-dimensional representation grid. Icons represent the auditory stimuli which participants can position anywhere in the 2-D space to indicate the similarity of a particular stimulus to other stimuli in the test, i.e. when stimuli are rated highly similar then they are placed close together on the grid. Sonic Mapper then produces a dissimilarity matrix which can be used for multidimensional scaling analysis methods. Scavone [293] used Sonic Mapper software in a multidimensional scaling analysis of synthesized sound effects and demonstrated that a large number of stimuli can be assessed effectively using the software. He also notes that because the participant is engaged in an interactive activity, rather than just passively listening, longer tests are possible, with the authors noting that "we were surprised by the enthusiasm that participants showed for the mapper task, working for several hours at a stretch" [294, p.4].

One drawback of Sonic Mapper (and other similar software which offer 2-dimensional representations of perceptual space) is that it is questionable whether it can capture situations that actually require more than two dimensions.

Comparison of perceptual testing methods

Parizet, Hamzaoui et al. [295] made a comparison of listening test designs in order to test the effectiveness of different paradigms and experimental procedures on known audio test samples. They asked listeners to evaluate or compare noise pleasantness and ran the experiment using different testing methodologies: direct scaling, paired comparisons and

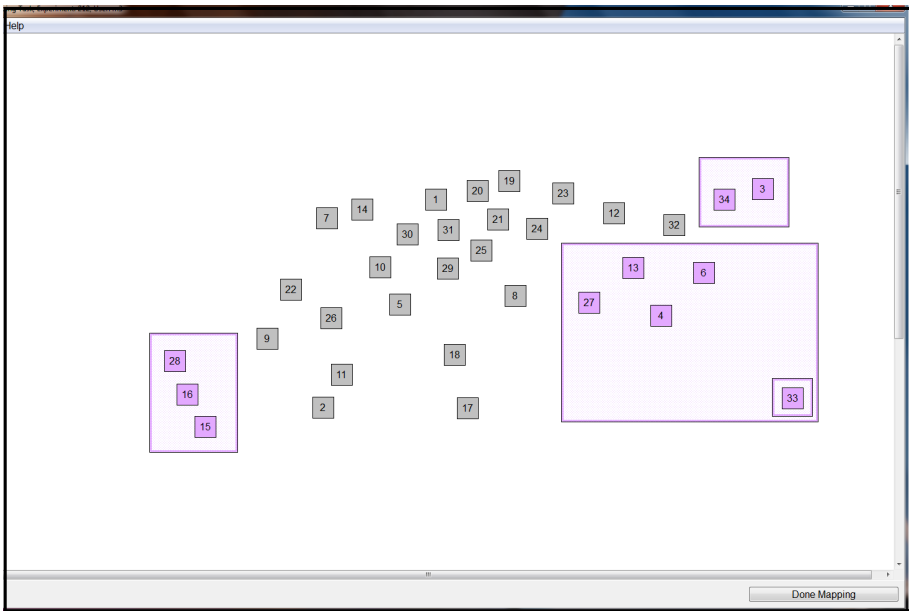


Figure 5.2: *Example of 2-D representation grid in Sonic Mapper*

similarity ratings using MDS. Although the different testing strategies brought similar results, they found that the discrimination power was greater for a paired comparison test than for those using direct ratings, but that the perceptual spaces obtained from all the different tests demonstrated very strong similarities.

5.2.2 Multi-variate Data Analysis

This section will not attempt an in-depth explanation of statistical methods for perceptual audio evaluation tests; a thorough exposition of such is provided by Bech & Zacharov [290].

For the perceptual evaluation of a particular audio system such as a loudspeaker or spatial sound rendering system, it is often possible to isolate a particular variable in order to test its perceptual attribute. However, it is widely acknowledged that perceptual evaluation of overall sound quality is a multidimensional problem arising from the multi-variate combination of individual auditory attributes. Similarly, musical performances by their very nature are inherently multi-variate sound objects, since single variables of a musical performance cannot be individually and separately controlled by the performing musician.

Multi-variate analysis techniques have become more popular over the last few decades as higher computational processing power has enabled larger amounts of data to be analysed efficiently [296, p.3]. A number of multi-variate techniques have been developed in the realm of market-research and product design and testing, such as multiple regression

analysis, multiple discriminant analysis, multivariate analysis of variance (MANOVA), factor analysis and cluster analysis.

Sometimes it is possible to identify and elicit particular attributes of the sounds under investigation. Determining the list of attributes used is often achieved by a mix of interviewing and drawing on the researcher's own experience, but more recently researchers have used dimensionality reduction techniques such as PCA or MDS.

Many of these sensory evaluation techniques have been developed by the food and drink industries. Indeed Lokki et al [19, 18, 102, 101] have recently used sensory techniques originally developed for the evaluation of wine.

In order to assess different attributes of sound, one must first identify the individual percepts to be assessed, which can be achieved either by direct elicitation - where a common or individual vocabulary of verbal descriptors is agreed upon in advance - or by indirect elicitation, where the attributes are not verbalised as such, but inferred after dimensionality reduction of the multi-variate data.

Dimensions of the perceptual data can be reduced using MDS or other techniques such as PCA and Perceptual Structure Analysis (PSA). PCA seeks to identify and interpret axes of dimensional space in terms of perceptual and/or objective attributes. PSA has been recently developed by Choisel & Wicklemaier [297], where the listener is ask to discriminate between triads of stimuli, identifying whether there is one attribute which two share that the other does not, and is used to good effect in examining auditory features of multichannel sound.

Multi-dimensional scaling

Whilst the above techniques attempt to define a list of attributes to be evaluated, often the experimenter does not want to guide the listener to specific individual auditory attributes (compositional approach), but rather make a more holistic evaluation of the audio sample *in toto* (decompositional approach).

Multi-dimensional scaling (MDS) is a now well-known series of multi-variate analysis techniques which allow the experimenter "to identify key dimensions underlying respondents' evaluations of objects" [296, p.485] without requiring the participant to use verbal descriptions, or indeed being made aware of attributes to which they should attend [290]. To this extent MDS is very useful in obtaining comparative evaluations of objects when the specific bases of comparison are unknown or undefined.

A well-known use of multidimensional scaling was undertaken by Grey [298] in his investigation of musical timbres. Choisel & Wickelmaier [297] also used MDS techniques to evaluate the auditory attributes of spatial sound in multichannel reproductions. Cerda et al. [33] have recently used MDS to relate objective acoustic parameters measured in

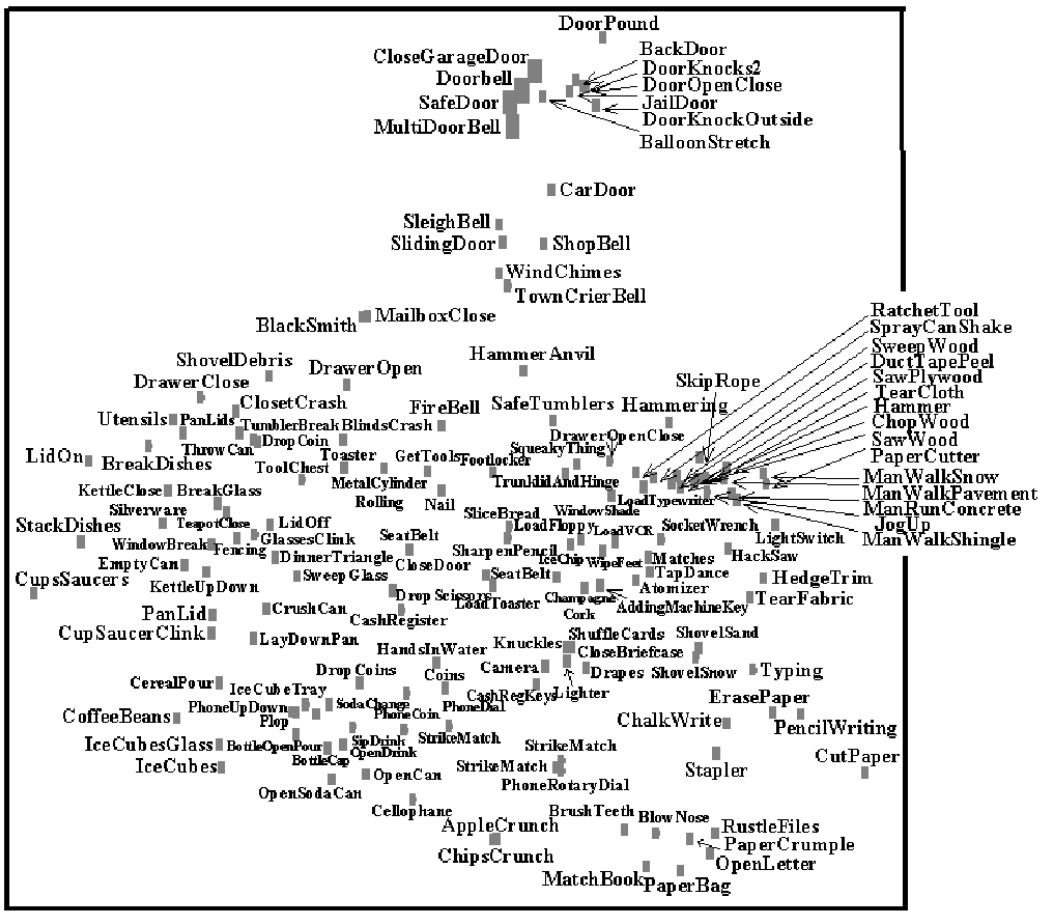


Figure 5.3: *MDS solution derived from dissimilarity matrices obtained using Sonic Mapper taken from [294]*

concert halls with measurements of the perceived quality of the room acoustics by a panel of listeners. They outline a number of reasons to use MDS in such an analysis over other techniques such as factor analysis or cluster analysis, which are also of importance to the present study. These reasons can be summarized as follows:

- Data can be measured on any scale
- Solutions are provided for each participant
- The experimenter does not specify the variables for comparison, which prevents researcher influence on the results
- MDS solutions have smaller dimensionality than factor analysis solutions
- Distances between all points are fully interpreted

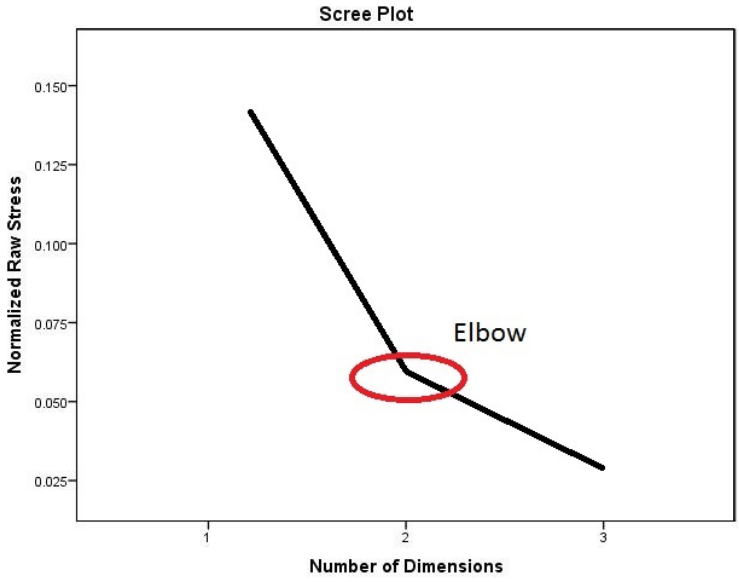


Figure 5.4: *A scree plot of the stress measure and number of perceptual dimensions used in the MDS analysis. The elbow indicates the point where increasing the number of dimensions does not yield significantly better results*

Judging the goodness of fit

Once all participants' similarity ratings have been obtained, MDS provides an aggregate analysis of the spatial configuration of similarity ratings across all the respondents - a common perceptual map. The experimenter must determine how good a fit this configuration is and how well it represents all the data collected from the individual respondents. A subjective evaluation can be made by visual inspection of the individual maps and the resulting common map and an assessment made of whether the aggregate map looks reasonable. A more robust approach is to implement a "stress measure" which gives an indication of the proportion of the variance of the disparities not accounted for the by the MDS model [296, p505], with a smaller stress value indicating a better fit. However, since a better fit can always be achieved by utilising more dimensions to describe the data, a trade-off between stress and increased number of dimensions must be observed. SPSS allows a scree plot to be drawn which enables the experimenter to visualise the improvement in goodness of fit when the number of dimensions is increased. Figure 5.3 shows an example of such a scree plot. The "elbow" of the scree plot indicates where a substantial improvement in fit is to be found.

A perceptual map of the stimuli proximity (relating to similarity) is also acquired for each participant, which has been achieved either through paired comparisons, or via direct mapping procedures such as those found in Sonic Mapper software (see above).

Disaggregate analysis considers these maps on a participant-by-participant basis, whereas as aggregate analysis combines the results of the respondents maps into a "common perceptual space".

Perceptual mapping techniques can either adopt a compositional (attribute-based) or decompositional (attribute-free) approach. Compositional approaches are based on attributes specified by the experimenter, and may use other multivariate techniques to acquire ratings based on these attributes. MDS is a decompositional approach which measures only the overall impression or evaluation and is useful when the experimenter wants to avoid undue influence on the participants evaluations and observations. Linear regression methods can supplement MDS analysis in order to correlate objective attributes with the perceptual dimensions identified in the MDS solution.

5.2.3 Subjective Evaluation of Musical Performance

As Gabrielsson urged in his review of music performance research in 1999 [193] music performance analysis should always be considered within the context of listeners' perception. A number of studies have attempted to categorise listeners' perception of music performance both in terms of perceptual attributes/features and also in terms of the expressivity or emotional content of performed music.

Perceptual Features

Many studies on the perceptual evaluation of audio or musical performance by listeners have relied on ratings of perceptual features formed in bipolar constructs. For example, "Speed: slow-fast", "Articulation: legato-staccato", "Dynamics: soft-loud" and "Brightness: dark-bright" and similar scaling methods have been successfully used in compositional approaches to perceptual audio evaluation.

Ratings of Emotion and Expressivity

For Seashore [134] expressivity could be conveyed through deviations in timing from the norm; the reference norm he used was a mechanical non-expressive performance with timings that related strictly to a metronomic version of the score.

More recently expressivity is investigated using perceptual evaluation methods, sometimes in an attempt to relate music performance features to the expressive intentions of the performer.

Ratings of Similarity between Musical Performances

A number of authors have undertaken perceptual tests to investigate how listeners rate the similarity of different musical performances.

In a study of how tempo and loudness are perceived to contribute to expressive performance, Timmers [299] looked at similarity ratings of three fragments of piano music - one fragment of a piece by Chopin and two fragments of a piece by Mozart - played by different pianists. One fragment was presented as a reference version, and four other versions were also presented and the listener was asked to judge the similarity/dissimilarity of the performances according to a seven point scale (1 = very dissimilar, 7 = very similar). Listeners were also asked to comment on which attributes of performance they had attended to. Multiple linear regression analysis was used to correlate the objective parameters with the perceived distances between the fragments. Local and global tempo measures explained most of the variance in the similarity data with measures which combined loudness and tempo being particularly effective in explaining the perceived similarity ratings.

In a study of musical performances produced in different room acoustic environments, Uneo et al. [3] asked listeners to rate the similarity between twelve pairs of samples which were presented for pairwise comparison. They concluded that the quantitative differences in the recorded performances were perceivable by the listeners, and furthermore the attributes that the listeners identified as varying between performances were in accordance with those described by the musicians who participated in the experiment.

5.3 Correlating Objective and Perceptual Attributes

As was discussed in Sections 2.4.1 and 2.4.3 much work has been undertaken in the perceptual evaluation of concert halls by listeners and stage acoustics by performers (eg. [103, 108, 26, 33, 98, 32]).

A number of techniques have been used to correlate objective to perceptual parameters have been used such as Principle Component Analysis, Linear Regression, Factor Analysis, Cluster Analysis, Multiple Analyses of Variance (MANOVA) and Correspondance Analysis.

In the growing field of research into the correlation between listeners' perceptual evaluation of music performance and objective measures of performance attributes, similar techniques are used and some of these are outlined below.

5.3.1 Regression Analysis

In general regression analysis is used to estimate the relationship between independent and dependent variables and in particular tries to predict the effect of one or more independent variables on observed dependent variables. Many forms of regression analysis have been developed, from a simple linear regression model to more complex multiple linear and non-linear regression models. Multiple linear regression can be used in conjunction with decompositional MDS techniques, in order to attempt to explain the perceptual dimensions identified in relation to characteristics or attributes of the stimuli involved.

In a recent case study on the influence of room acoustics on solo cello performance, Schaere Kalkandjiev & Weinzierl used a multivariate hierarchical linear model (HLM) - a special form of linear regression model where a nested structure of conditions is present - to investigate the correlation between numerous musical performance attributes and the room acoustical parameters of the real concert venues in which different performances were produced. They found that over half of the variance in performance parameters could be explained by the different room acoustic conditions [10, 276, 300].

In order to reduce the number of potential room acoustic parameters in the linear regression model they first undertook a Principal Component Analysis, which reduced the number of components to four, namely RT60, ST_{late}, G_e and Brilliance, which they interpret as "perceived duration of reverberation", "reverberant energy", "early acoustical support" and "timbre of reverberation" respectively; further details on this recent study are given in Section 4.4.5.

5.3.2 Principal Component Analysis

Principal Component Analysis can be used to convert a set of observational data, which may include several variables that are highly correlated with each other, into a number of uncorrelated variables called "principal components". The number of principal components is always less than the number of original variables, and hence PCA is a dimensionality reduction technique. The first principal component is the the one which accounts for the largest proportion of variance in the data as possible, and other components follow in order of proportion of variance accounted for. Lokki et al. have recently used PCA in subjective evalatuions of concert hall acoustics [100, 102, 18] and others have used PCA in evaluations of other audio and music performance which seek to correlate subjective ratings with objective measures (e.g. [301, 101, 274]).

5.3.3 Music Performance Attributes and Perceptual Correlates

Chudy [209] found significant differences in the spectral centroid and spectral deviation between string instrument performers, and related these timbral measures to perceptual parameters of tone described as "pinched, harsh and strong". These spectral aspects in turn were correlated to production parameters of bow force and motion.

Timmers [166] noted that listeners' perception of loudness might be influenced by a tendency to focus on the voice part in accompanied vocal music when judging dynamics.

Hedblad [302] correlated perceptual features with low-level features extracted by means of MIR extraction tools, using step-wise multiple linear regression and found that many of the expected predictors were highly correlated with the corresponding perceptual rating. For example,the extracted feature 'MT_Pulse_Clarity_1' was correlated with rhythmic complexity (r=0.73) and the extracted feature 'QM_onsets' with energy (r=0.75).

Alluri & Toivianen [303] investigated perception of timbre in a polyphonic context using a number of bi-polar scales. They posited three main perceptual dimensions - Activity, Brightness, and Fullness - and found these correlated well to the quantitative measurements of timbre, with spectrotemporal features being the most effective predictor, whereas Mel Frequency Cepstral Coefficient (MFCC)s did not appear to correlate well to the perceptual dimensions.

Many studies in this area have tried to relate ratings of expression or emotion in musical performances with extracted performance parameters. For example, Gabrielsson & Juslin [152] compared listeners' ratings of expressive content of musical performances such as "happy", "sad", "angry", "fearful", "tender" (which the performers had been instructed to portray), with the extracted performance parameters such as tempo, dynamics, timing and spectral aspects. They found that although some characteristics were easier to communicate than others, listeners were generally good at discerning the intended emotional content of the performance, even though individual performers encoded the emotional content in a variety of ways.

Timmers [166] looked at the perception of musical performances on historical and modern commercial recordings, and found that evaluations of the perceived emotion were independent of the date the recording was made. However, she found that performance parameters were an important factor in listeners' perception of emotion and dynamics. Furthermore, significant correlations between measured performance parameters and listeners' perceptual judgements suggested that performance parameters highly influenced the perception of emotion in the performances.

5.4 Pilot Listening Test I - Producing Stimuli

Pilot listening test I was undertaken to check whether room-reflected sound present in the dry (but not anechoic) source recordings might be perceptually relevant, even after the source recordings had been convolved with the previously measured SRIR from the listening position in the real venue. Any colouration in the recordings due to the presence of even low levels of room-reflected sound might lead to listeners being able to distinguish between source recordings originally made in the different acoustic settings. For example, it might lead a listener to be able to rate a difference between a sample originally recorded in the Music Recital (MR) setting from one made originally in the Large Choral (LC) setting, purely on the basis of a difference of the reverberant tail audible at the ends of phrases and in gaps in the singing, rather than on the (wished for) basis of differences in the musical performance itself.

Close-mic recordings of all quartet singers were recorded via head-mounted DPA4066 microphones as described in Section 4.7. Despite a large ratio of direct sound to room reverberance, at the end of phrases and in longer silences between sung notes the reverberant tail of the room acoustic is audible in careful listening (mostly after the end of louder passages of singing), although at a very low level. For example, in Fragment A the average RMS level during the sung phrase is -25.13 dB, whereas the average RMS level of the room-reflected sound measured from the end of the sung phrase until the room-reflected sound has decayed fully is -65.03dB, giving a difference of slightly under 40dB. Similar values were calculated for other fragments recorded in the same acoustic setting.

The pilot listening test was designed to test the method of producing audio stimuli for the main listening test (described in Section 6.3) and to determine whether listeners could distinguish between: a) stimuli produced by simply convolving close-mic recordings with an appropriate SRIR (*untrimmed*) and b) stimuli where the close-mic recordings had first been edited in order to remove room-reflected sound at the end of sung phrases (*trimmed*)

Complete removal of all effects of the original room acoustic of the recording venue is not possible, as this involves use of complex de-reverberation techniques, which are still very much in development and not yet fully robust. A compromise approach is to trim the source recordings at the ends of phrases to attempt to avoid the problems outlined above. However, this is not wholly reliable since it does not remove the room-reflected sound present whilst the singing is ongoing but indeed some authors argue that phrase endings are perceptually more salient than running-reverberation [55]

Method

Stimuli Fragments of the SATB quartet recorded in the real venue were edited into shorter phrases (and were also used in Pilot Listening test 2 described in Section 5.5. Each separate channel of the four vocal lines (Soprano, Alto, Tenor and Bass) was convolved with the "Music Recital" SRIR previously recorded in the *real performance venue* (Section3).

Each separate vocal track was convolved with the measured SRIR corresponding to the performance position of the relevant singer, i.e. the soprano recording was convolved with the SRIR measured at performer position A, the tenor recording was convolved with performer position B and so on. This resulted in a set of 16 audio files - four B-format files for each vocal part - which were then summed in order to produce a full convolved set of quartet excerpts. For the purposes of this listening test, W-channel files were mixed down to a stereo file in Adobe Audition (sampling rate 48kHz, 24 bit) and normalized so that there were no level differences between the two versions (*trimmed source* and *untrimmed source*) of the file.

Two fragments were used in the listening test: Fragment 1 was taken from "Remember Not" (bars 53-55) and Fragment 2 was taken from "Audivi Vocem" (bar 10 last note only). Two versions of each fragment were produced, a trimmed and un-trimmed, which were taken from recordings made in the *real space* in LC (Large Choral) conditions (see Table 4.2 for details). For reference, musical scores of the pieces performed by the quartet are to be found on the data CD accompanying this thesis and the recordings used are found on the data CD (for listing please see Appendix A).

Procedure An ABX test was carried out using a graphical user interface in MATLAB in order to determine if subjects could reliably identify the *trimmed source* versions from the *untrimmed source* versions. Each participant was presented with a simple user interface (as pictured in Figure 5.5) - samples A and B and X could be auditioned any number of times by the subject, and then the participant was asked to determine whether sample X was the same sample as A or B.

Participants 8 participants took the test - all were staff/students in the AudioLab, Dept of Electronics, University of York and had experience in critical listening skills. The headphones used for presentation were calibrated so that the signals were reproduced at a reasonable listening level of 65dB SPL.

Results and Discussion

Mean average score for correct identification of the hidden sample X was 66.66 % (St dev. 18) and overall 33 out of 48 samples were correctly identified. Although these results show

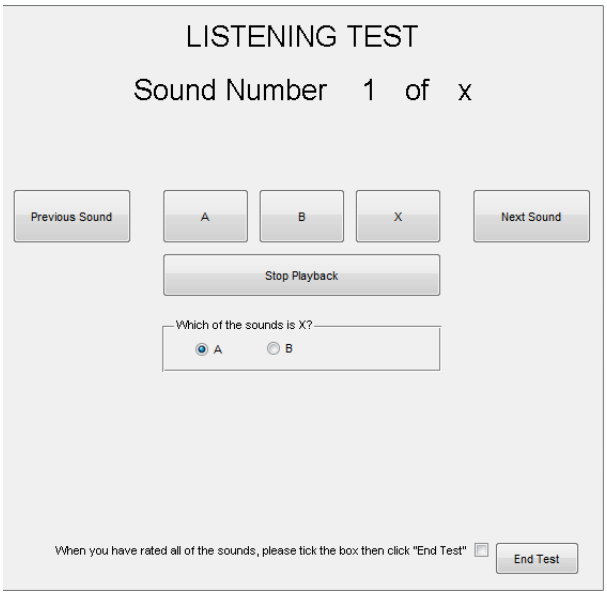


Figure 5.5: *Pilot Test user interface of ABX test to determine if listeners could distinguish between trimmed and untrimmed source material*

more than a "by chance" ability to identify the correct X sample, only one participant correctly identified the hidden sample on all occasions, with the majority of participants (5 of 8) correctly identifying the repeated sample for only 4 or fewer of the 6 tests undertaken. On the whole participants were not able to reliably distinguish the two samples presented in the ABX test and so it was decided that *untrimmed* versions of recorded singing could be used as source material in the main listening test (See Section 6.3).

5.5 Pilot Listening Test II - Quartet Performances

The present study seeks to discover listeners' ratings of similarity/dissimilarity between musical performances, and therefore techniques will be used which enable the investigation of objective perceptual attributes of musical excerpts rather than affective measurements of liking or preference.

In the design of any such listening test two factors have to be taken into account - the overall length of the test (to avoid participant fatigue) and the possibility of comparing each musical sample with every other in the test barrage - and these two factors must be balanced in the final design.

A second pilot listening test was carried out which had two main objectives: 1) to determine the most suitable method for carrying out the main listening test, in particular to evaluate the use of a pairwise comparison test or a multiple scaling test, and 2) to investigate how listeners would rate the similarity of separate performances of the same

pieces sung by the vocal quartet in the three different acoustic configurations of the real performance venue, and in the anechoic chamber.

5.5.1 Method

Two different procedures were tested in order to gain insight into the best method for the main listening test described in 6.3: a pairwise comparison task and, using the same stimuli, a clustering task.

Stimuli

A number of example fragments of the same bars of one piece recorded by the vocal quartet were chosen for the listening test. Fragments were chosen from those excepts which allowed a good number of suitable versions to be identified. Each separate channel (Soprano, Alto, Tenor, Bass) taken from the close-mic recorded vocal signals were convolved individually with the relevant listener position impulse responses (performer A, B, C and D respectively) which had been previously recorded in the *real performance venue* (see Section 3.3).

The convolution was performed in the frequency domain using a short MATLAB script. The resulting four B-format files (Singer $*$ Listener IR) where then summed according to a method outlined by Farina [28], who states that:

> *Provided that a set of B-format impulse responses has been measured or computed for a given receiver position in an acoustic space and several different sound sources positions, it is possible to place a sound track in the virtual sound space simply by convolving the original (dry) signal with the proper B-format IR. Adding the results of the convolution of different sound tracks with IRs relative to different source positions, a complete "soundscape" can be created: for example, it is possible to place in their proper positions the single instruments of a virtual orchestra, starting from multi-miked, multitrack studio recordings (which are almost perfectly anechoic) or from separately synthesised MIDI sequences.*

The resulting four-performer "soundscape" B-format files were then processed using the Visual Virtual Microphone [304] to produce a stereo file which replicates the signal that would have been recorded in the same soundfield by two cardioid microphones. In each set of four fragments from the different acoustic environments, version 3 and version 4 (for example, labelled *mrfbv3* and *mrfbv4*) are exact copies of the same fragment.

Procedure

Two different procedures were tested in order to investigate the user interaction, length of testing procedure and reliability of results using the two methods. The participants were randomly assigned to two groups - each group undertook either the comparison or the sorting task and then went on to complete the other testing procedure.

Both tests were carried out using SonicMapper software [294]. In the "comparison" task participants were presented with pairs of fragments and asked to rate on a sliding scale how similar they perceived the sung fragments to be. The user interface for this test is illustrated in Figure 5.6.

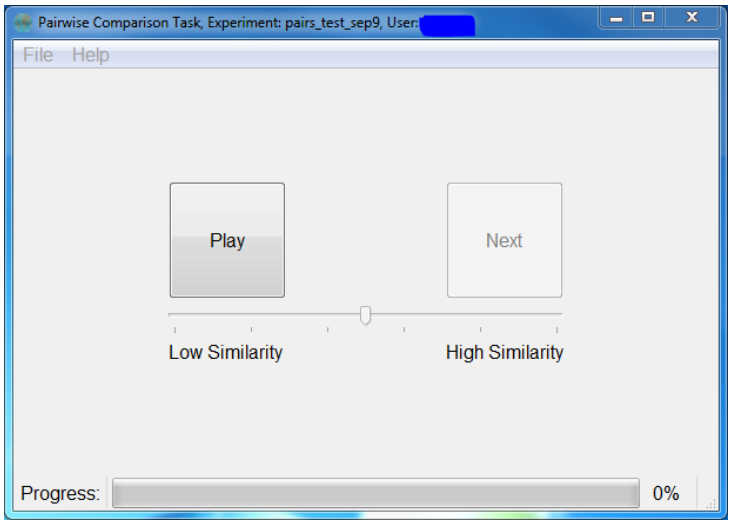


Figure 5.6: *Example of user interface in Sonicmapper of COMPARISON task*

The "sorting" task allowed participants to move icons representing sound objects around a 2-D space. Participants were asked to group sung fragments (represented by square numbered boxes) which they thought were similar and to place fragments which they found to be dissimilar at a distance from each other. If groupings of fragments revealed themselves, participants were asked to group and label such fragments, using a free choice of vocabulary. The user interface for this test is illustrated in Figure 5.6.

All sound files presented were recorded at 48KHz, 24 bit, and presented in stereo on a pair of Sennheiser HD250 linear II headphones. The playback system was calibrated to a suitable listening level over headphones, which was confirmed by participants to be comfortable and this level was consistent throughout and between tests.

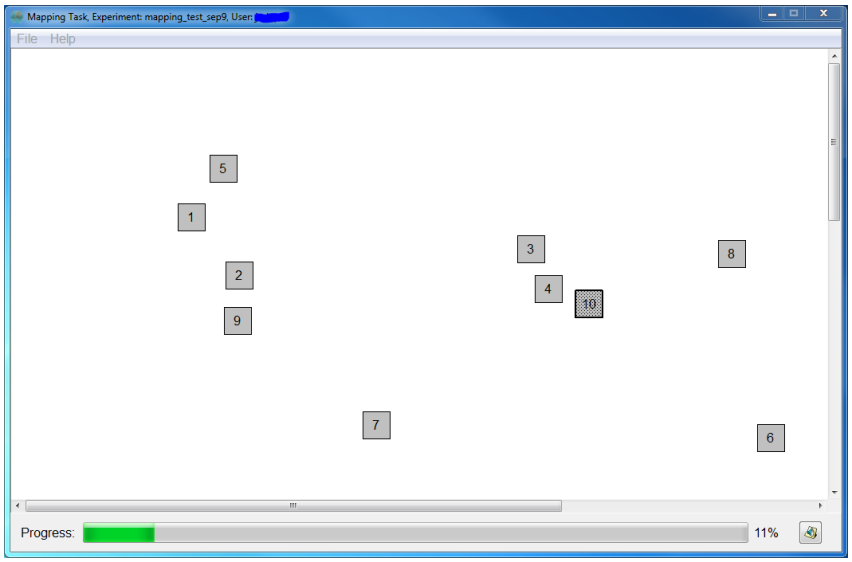


Figure 5.7: *Example of user interface in Sonicmapper of SORTING task - audio fragments are represented by numbered boxes*

5.5.2 Results

Four participants took part in the trial, but it became obvious during the course of the listening test that the pair-wise comparison method resulted in a test that was too long, meaning that only 2 participants completed the pair-wise comparison task fully. On the other hand the sorting task was managed by all participants. Therefore, results reported here are those obtained from the "sorting" half only.

The dissimilarity ratings obtained from the sorting task were aggregated across the four participants. A non-metric Multi-Dimensional Scaling (MDS) analysis (as outlined in Section 5.2.2) was carried out in MATLAB using the *mdscale* function included in the Statistics Toolbox ^{TM} [305].

Multi-dimensional scaling is an exploratory data reduction technique which seeks to identify unrecognized dimensions in the perceptual evaluation and comparison of objects, in this case audio fragments. MDS is often used when the specific attributes for comparison, or bases of subjects' comparitive evaluations, are not known or unidentified [296, p.492].

Assessing the fit of the MDS solution

The MDS technique attempts to model the similarities as distances between points in a geometric space, where each point represents one of the audio fragments. Non-metric MDS retains only ordinal information in the proximities which is used to construct the geometric space (also known as a perceptual map) through a monotonic transformation of the original dissimilarities. These optimally scaled transformed proximities are often

Stress	Goodness of fit
over 0.20	poor
0.10	fair
0.05	good
0.025	excellent
0.00	perfect

Table 5.1: *Stress and goodness of fit using Kruskal's Stress measure [306]*

referred to as *disparities*.

The goodness of fit of the model can be evaluated by calculating stress values for modelling in an increasing number of dimensions. The standard method is that originally proposed by Kruskal [306] as computed by equation 5.1.

$$STRESS = \frac{\sqrt{\sum (f(p) - d)^2}}{\sum d^2} \tag{5.1}$$

where p is the vector of proximities, $f(p)$ is the monotonic transformation of p, and d is the vector of point distances, in order that the stress value is minimized.

Small stress values indicate a better fitting solution; Kruskal suggested following guidelines for interpreting these stress values (See Table 5.1).

Plotting stress values as a scree plot allows the "elbow" point to be identified where increasing the number of dimensions no longer offers a large improvement in fit. The scree plot for this pilot test data is shown in figure 5.8 and it can be seen that modelling the data in two dimensions already gives a good fit to the data, and plotting in 3-D only offers slight improvement.

An additional method of assessing the fit of the MDS model is provided by a Shepard plot [296]. Here the original data dissimilarities are plotted against both a) a nonlinear monotonic transformation of the original dissimilarities (disparities) and b) the inter-point Euclidean distances of the modelled disparities (distances). Figure 5.9 shows the Shepard plot for this listening test data modelled in two dimensions.

The distances between points produced by the transformation are close to the red line showing that the non-metric modelled distances recreate the original disparities well (they are not too scattered about the red line). The plot also shows an almost linear relationship between original disparities and the modelled distances (red line), meaning that small dissimilarities in the modelled visualisation correspond well to small dissimilarities in the original data.

Modelling the data in three dimensions offers only a slight improvement in the relationship between distances and disparities (see Shepard plot in Figure 5.10) with less

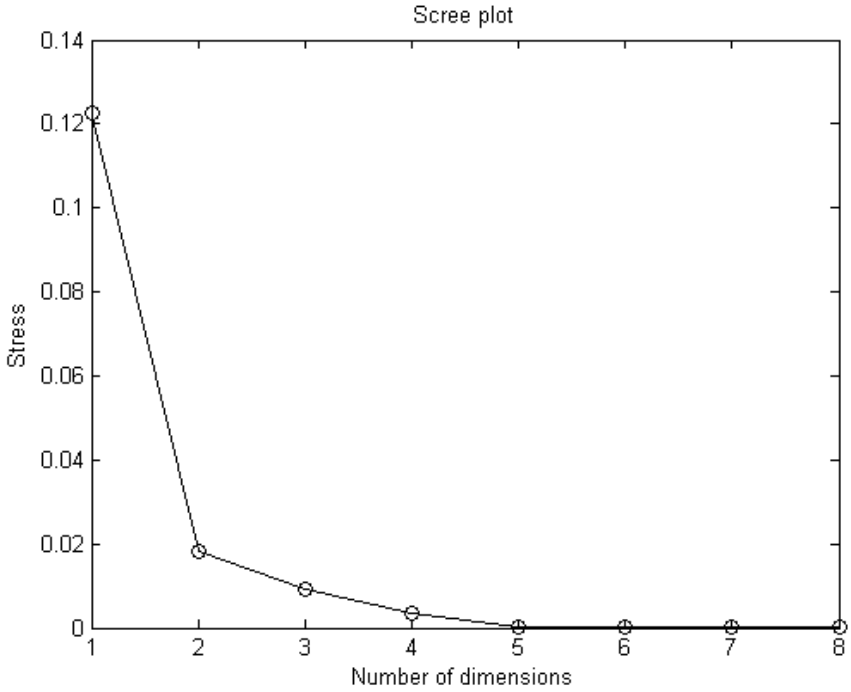


Figure 5.8: *Scree plot of stress measure with increasing dimensions used to model pilot test II dissimilarity data*

scatter seen in the blue circles about the red line. Again the monotonic transformation is almost linear, suggesting a good fit to the original dissimilarities, with most dissimilarities being slightly under-represented by the modelled disparities, but some of the larger dissimilarities (above 0.4) being represented by an almost one-to-one relationship.

MDS solution

The two-dimensional visualisation of the MDS solution is illustrated in Figure 5.11 and the three-dimensional solution is illustrated in Figure 5.12. Red points represent fragments recorded in the anechoic chamber, black are those made in the dry configuration of the *real performance space* (SP), and blue and green points are those made in the medium (MR) and long reverberation (LC) settings respectively.

5.5.3 Discussion

All audio fragments of recorded singing were convolved with an SRIR of the medium acoustic setting of the *real performance space* (listener position MR) and listeners were asked to attend to the performance itself rather than to any room acoustic characteristics they could hear.

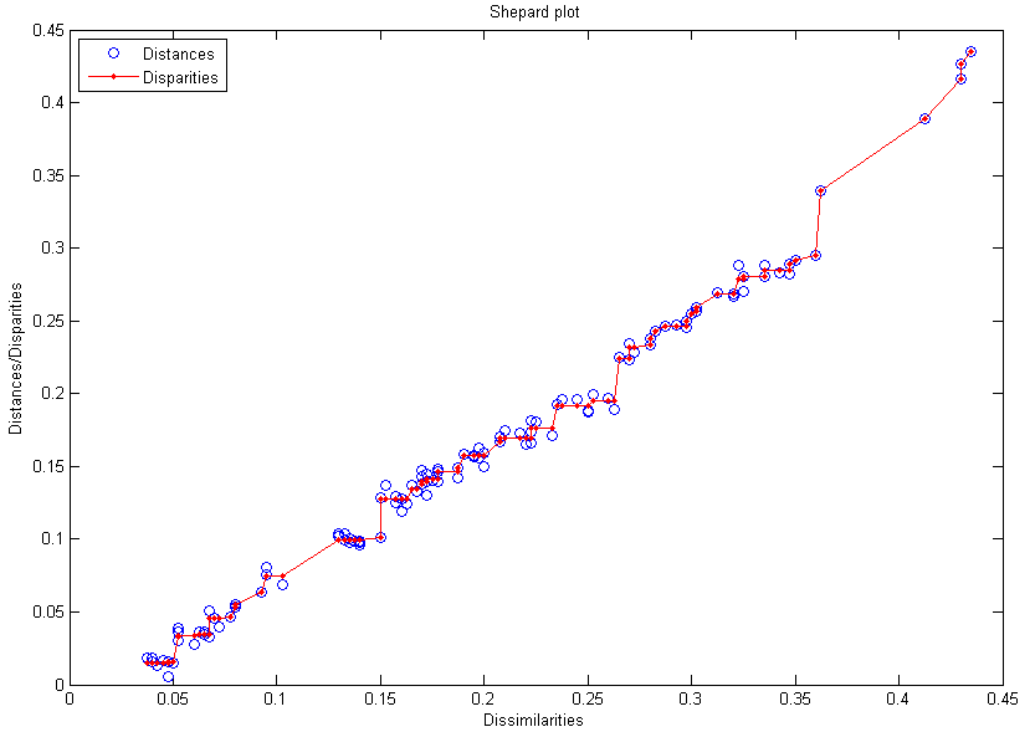


Figure 5.9: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of pilot test II point configuration modelled in 2 dimensions*

The pairwise-comparison listening test was too long and risked listener fatigue whereas the comparison test was easily achieved by all participants. This finding reflects those of Bonebright [17] who conducted a comparison between two different data collection methods for auditory stimuli: a paired comparison task and a computer sorting task, finding that the computer sorting task was a viable alternative to the traditional paired comparison task.

Two-dimensional and three-dimensional MDS solutions to the "sorting" task data are presented in figures 5.11 and 5.12 respectively.

It can easily be seen in both solutions that listeners grouped the anechoic fragments tightly together as being very similar, but placed apart from the next nearest group - the SP fragments - showing that listeners judged similarities within groups of fragments and their distinction to other groups.

Three out of four of the "dry" source (SP) versions (black markers) are grouped closely together and rated as similar. One of the SP versions (*spfbv1*) seems to be an outlier and is positioned in the top left hand corner of the grid in Figure 5.11. In the performance recorded on this fragment it can be heard that the bass singer is out of time with the other

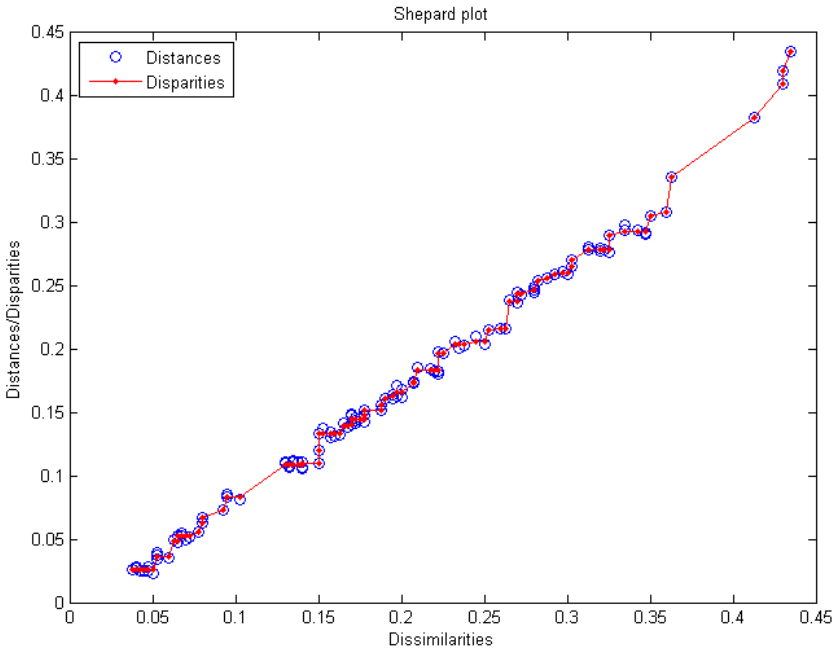


Figure 5.10: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of pilot test II point configuration modelled in 3 dimensions*

singers at the beginning of the excerpt, only synchronising again properly at the final note of the phrase. Therefore it seems that all listeners judged this as highly dissimilar to the other SP fragments.

Similarly one of the LC fragments (lcfbv1) is positioned away from the main groupings of fragments towards the base of the grid slightly to the right of centre. Again, listening to this fragment it can be heard that the balance between the voice parts in this recording is not good, with the soprano singer being much louder than the other voice throughout most of the excerpt, before finally balancing in the final notes of the phrase.

The LC ("reverberant" source) and MR ("medium" source) fragments are less well distinguished as the other groups of fragments, for example LC fragment version 2 (lcfbv2) is positioned close to two of the MR fragments. Ratings of similarity between MR and LC fragments are not surprising as singers reported that the difference between these two acoustic settings in the *real performance venue* was not as striking as the difference between the dry setting (SP) and the others.

The identical versions (versions 3 and 4) of each set of fragments are placed close together e.g. *lcfbv4* and *lcfbv3* (green markers) are very tightly positioned towards the top centre of the grid.

The three-dimensional solution (Figure 5.12) offers similar insights into the perceived similarity between the fragments. The outliers (spfbv1 and lcfbv1) are still easily seen

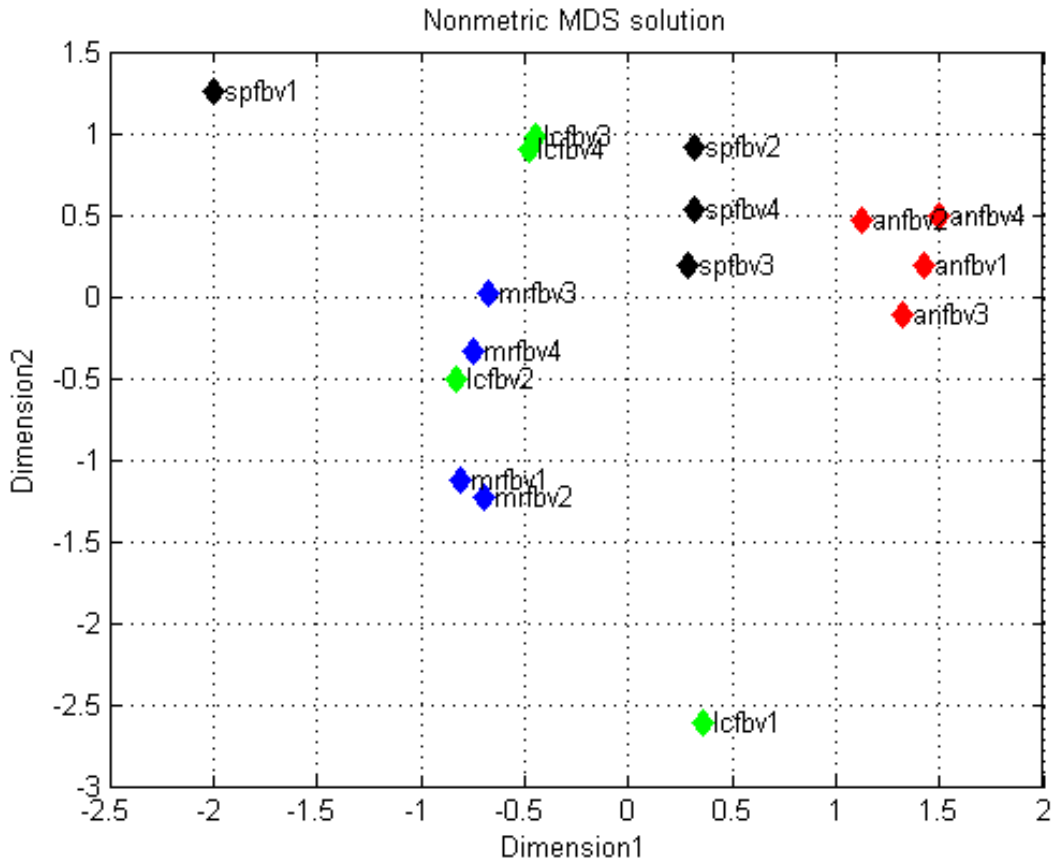


Figure 5.11: *Nonmetric MDS solution to pilot test II similarities, modelled in 2 dimensions*

but an added dimension of dissimilarity between the groups of fragments might be distinguished. Nevertheless, modelling the perceptual space in three dimensions in this case does not improve the fit to the data by a great amount (see 5.5.2) and plotting the map in 3D does not greatly improve the ability to see perceived differences between the audio fragments.

5.6 Summary

Section 5.2 outlined some of the methods which have been used in the perceptual evaluation of audio and music, noting that the evaluation of musical performances provides a multivariate data set, some techniques for dimensionality reduction were discussed. Section 5.3 looked at recent attempts to correlate objective parameters to listeners' judgements of preference or similarity between audio objects or musical performances.

Section 5.4 examined whether producing stimuli, which would then be used for the main listening test in this thesis, could be achieved by taking close-microphone recordings of

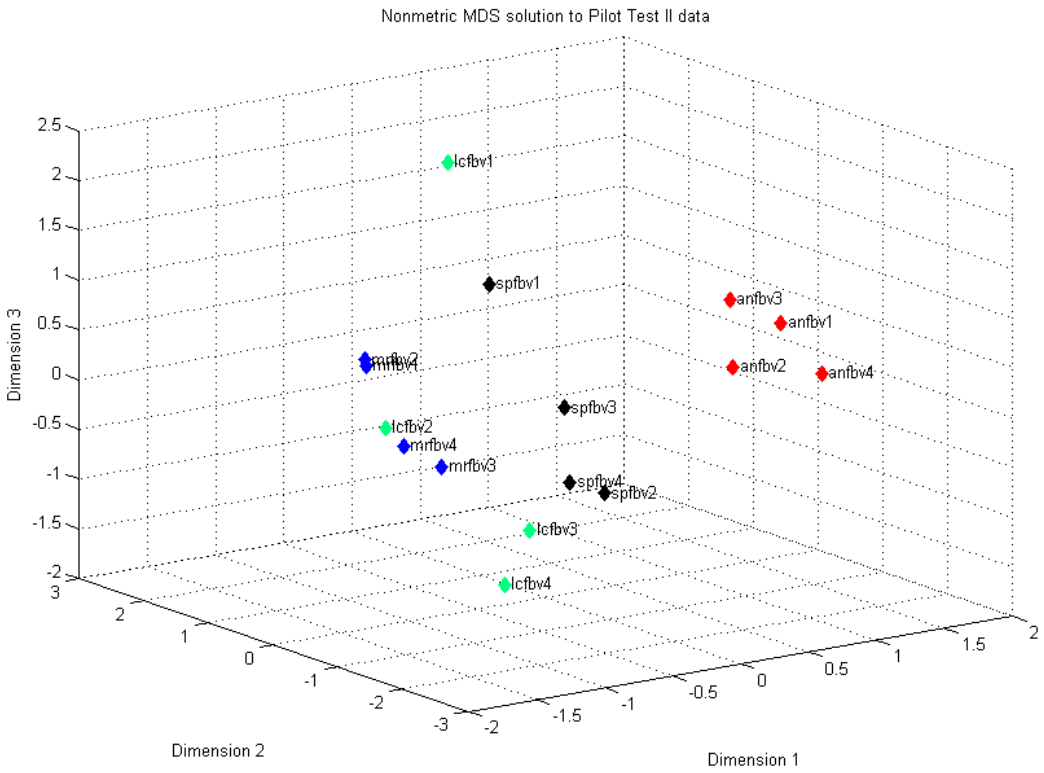


Figure 5.12: *Nonmetric MDS solution to pilot test II data, modelled in 3 dimensions*

singing performances and convolving them with the SRIR measured in the real performance venue.

A pilot listening test 5.5 showed that a sorting task was a reasonable method to gather similarity ratings between the musical fragments, avoiding over long listening test times which might otherwise lead to fatigue of the participants. The non-metric multi-dimensional scaling solutions, representing a perceptual map of listeners' responses, show that listeners were able to rate the similarity between fragments of vocal quartet singing recorded in the different acoustic surroundings. Results suggest that listeners judge singing performance produced in the same acoustic configurations to be similar, but can hear differences between those recorded in different acoustic configurations.

Methodologies for producing stimuli and conducting a listening test to gather participants' judgement on the similarity between singing performances have been tested. MDS analysis has been shown to provide a suitable method for visualising the multi-variate similarity data collected in the listening test. Chapter 6 goes on to describe recordings made in the *real performance space* (NCEM) and the *virtual performance space* provided by the VSS and subsequent listening test, MDS visualisation and acoustic analysis of the singing performances.

Chapter 6

Singing in Real and Virtual Acoustic Environments

6.1 Introduction

This chapter first describes how recordings of solo singers were made in the *real performance space* and the *virtual performance space* (Section 6.2) provided by the Virtual Singing Studio (described in Chapter 3) and present some of the singers' own evaluations of their singing in real space and the simulation.

Section 6.3 describes the listening tests which were undertaken to assess the similarity between fragments of the recorded solo-singing. It also presents the results of the MDS analysis undertaken on the similarity ratings gained from the listening test via a sorting task, the methodology for which was tested in a pilot test as outlined in Section 5.5.

Section 6.4 presents an analysis of the recorded fragments in terms of a number of singing performance attributes, namely fundamental frequency vibrato rate and extent, tempo, and measures of intonation accuracy and precision.

Section 6.5 attempts to infer the dimensions of the MDS derived perceptual maps, by fitting the objective performance attribute data to the perceptual space in order to identify which attributes listeners might use to assess similarity between the solo singing performances.

Naming conventions

In this section and for the remainder of this chapter, fragments of recorded singing are labelled and referred to for brevity using the following convention

- LC/MR/SP - the acoustic configuration in which the recording was made i.e. LC - Large Choral (highly reverberant), MR - Music Recital (medium reveberant), SP -

Speech dry acoustic followed by...

- Real/Virtual - whether recorded in the *real performance space* or the *virtual performance space* followed by ....
- V1/2/3 - version 1, 2 or 3 to identify from which recording the fragment was taken.

In this way *LCvirtualV1* refers to a fragment taken from recording number one, of those recorded in the Large Choral acoustic configuration of the virtual performance space.

In addition fragments in test 212 are also denoted with a letter corresponding to the verse from which the fragment was taken e.g. *LCVirtualC7* is the 7th fragment in the Large Choral acoustic configuration in the virtual performance space, taken from the third vers (verse C) of the song. The lyrics of the verses in this song are found in G

6.2 Recording in real and virtual performance spaces

6.2.1 Method

Recordings made by the professional solo singers who participated in the subjective evaluation of the virtual and real performance spaces (Section 3.6) were used to provide source stimuli for the listening test described in Section 6.3(further details of the singers are given in section 3.6).

Singers were asked to sing familiar pieces of their own choosing and to sing these a number of times in each of the three acoustic configurations. The singers sang in the virtual performance space first, and gave their subjective responses about the acoustic characteristics of the venue, the plausibility of the simulation, and thoughts on how their vocal performance changed in the different acoustic environments.

As far as was possible the singer was then taken to the real performance venue on the same day, for a similar recording session performing the same pieces. In the real performance venue panels and drapes were manipulated to provide the three acoustic configurations which had been simulated in the VSS. The singer sang in these acoustic configurations in a random order, without initially knowing which configuration related to those they had experienced in the virtual performance space. After recordings had been made in each of them the singers were asked to state which of the acoustic configurations corresponded to the simulated acoustics of the virtual performance space.

Laryngographic data was not collected (see section 4.3.5) as it was hoped to keep the performance experience as natural as possible, and it was decided that the close-mounted microphone signal would be sufficient for fundamental frequency estimation using the

Figure 6.1: *Photograph to illustrate position of head-mounted microphone and masking tape used to replicate position between recordings*

AMPACT toolbox (Section 6.4) as there was no cross-talk contamination of the signal by other singers, as was the case in the choral recordings reported in Section 4.7.

Microphone placement In order to be able to compare across singers, the microphone was placed at a fixed distance from the mouth (in this case 5cm). Since even a small discrepancy in the placement of the head-mounted microphone will mean a large change in the relative level of the vocal signal captured by the microphone the microphone placement needs to be as accurate as possible. In practice this was sometimes difficult to achieve and some time was spent on adjusting the microphone and headband accordingly. To ensure consistency of microphone placement for each singer a piece of masking tape was placed on the side of the singer's face and the position of the corner of the mouth and the base of the ear marked on the tape in pencil. This masking tape was used to aid microphone placement in both the the virtual and real performance spaces (See Figure 6.1).

6.2.2 Singer's Evaluation of Solo Singing Performance

After the recordings in the real and virtual spaces had been made, singers were asked via informal interview and via questionnaire (Appendix F)about their own performances in the spaces, specifically to identify which performance attributes the singers themselves felt were altered in the different acoustic settings.

Singer Number	Voice type	Performance Changes Identified
201	Soprano	
211	Alto	Tempo, Length of Notes
212	Mezzo-soprano	Tempo, Length of Notes, Tone, Articulation,
213	Alto	
221	Tenor	
231	Bass	Tempo, Intonation, Length of Notes, Articulation
232	Bass	Tempo, Intonation, Tone

Table 6.1: *Summary of performance changes identified by the singers in the different acoustic configurations of the real and virtual performance spaces*

6.3 Listeners' Evaluation of Solo Singing Performances

The aim of the main listening test is to evaluate the degree of similarity between sung performances rather than to ascertain if listeners can distinguish between the samples, which could be due to idiosyncratic differences and the fact that the sung performances will differ in small ways, because they have been produced on different occasions.

In many ways the listening test undertaken here has opposite aims to many perceptual listening tests which are used for example to test audio codecs, audio processing techniques or headphone/loudspeaker reproduction capabilities. In those tests the objective is often to quantify pereived differences which listeners perceive stemming from the results of processing or reproduction techniques and the source material is the same across the samples [290].

In the present study the objective is to gauge the perceived similarity of the singing performances recorded in the *real* and *virtual spaces*, whilst making sure that listeners are not being unduly influenced by the reproduction method or the characteristics of the room acoustics in which the recordings were made.

Participants

Twenty experienced listeners participated (6F/14M)in the listening test. Participants had between 1 and 40 years of musical training (Mean 15.4, SD 11.24) and over half (11 participants) had specifically vocal training (ranging between 1 and 33 years, mean 8.94 SD 11.74). Ages ranged between 27 and 50 (Mean 30.35 SD 8.84 years). All participants performed music regularly (most singing in a choir or teaching music) and half worked or researched in the area of audio/music technology. 11 out of 20 reported regular singing activity predominantly in Western Art Music tradition including "early music", choral music or opera.

Materials

A number of productions (fragments) of one phrase of a piece by each singer were chosen for evaluation. The fragments were chosen from those excerpts which allowed a good number of suitable versions to be identified. The fragments were taken from the close-mic recorded vocal signals as detailed in Section 6.2.

Recordings of all the fragments used as stimuli in the listening tests can be found on the data CD accompanying this thesis. Details of the fragments used are outlined below.

Test 232a

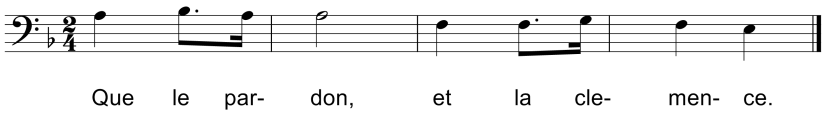
Voice Bass

Lyrics "Que le pardon, et la clemence"

Piece "Si la riguer" Aria from the opera La Juive by Halévy

Number of fragments recorded 5

Names of fragments:



SPRealV1
SPVirtualV1
MRVirtualV1
LCRealV1
LCvirtualV1

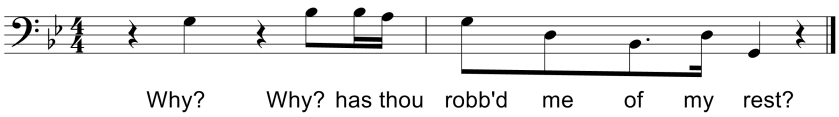
Table 6.2: *Names of Fragments in test 232a*

Test 232b

Voice Bass

Lyrics "Why, Why has thou robbed me of my rest"

Piece from "Saul" by Henry Purcell



Number of fragments recorded 8

Names of fragments:

SPRealV1	SPRealV2	
SPVirtualV1		
MRVirtualV1	MRvirtualV2	MRvirtualV3
LCRealV1		
LCvirtualV1		

Table 6.3: *Names of Fragments in test 232b*

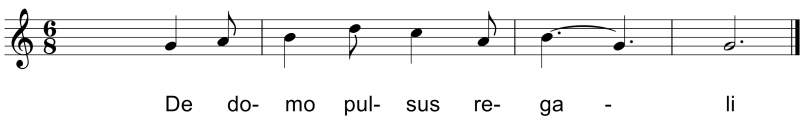
Test 221

Voice Tenor

Lyrics "de domo pulsus regali" , French medieval

Number of fragments recorded 11

Names of fragments:



SPRealV1	SPRealV2	SPRealV3
SPVirtualV1		
MRVirtualV1		
LCRealV1	LCRealV2	LCRealV3
LCvirtualV2	LCvirtualV3	

Table 6.4: *Names of Fragments in test 221*

Test 212

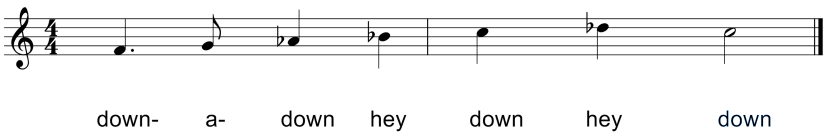
Voice Mezzo-soprano

Lyrics "Down a down a down hey down" from "The Three Ravens" an English folk ballad

Number of fragments recorded 34

Names of fragments:

Lyrics of the Verses of this song are found in Appendix G



SPrealB2	SPrealC3				
SPvirtualA1	SPvirtualA2	SPvirtualB3	SPvirtualC4	SPvirtualD5	
MRvirtualA1	MRvirtualA2	MRvirtualA3	MRvirtualA4	MRvirtualB5	
MRvirtualB6	MRvirtualB7	MRvirtualC8	MRvirtualD9	MRvirtualE10	MRvirtualE11
LCrealA4	LCrealC6	LCrealD7	LCrealD8	LCrealE9	
LCvirtualA1	LCvirtualA2	LCvirtualA5	LCvirtualB3	LCvirtualB4	LCvirtualB6
LCvirtualC7	LCvirtualD8	LCvirtualE10	LCvirtualE9		
MRrealA1					

Table 6.5: *Names of Fragments in test 212*

6.3.1 Producing Stimuli

All recorded source fragments were convolved with the same SRIR measured in the Musical Recital setting (MR) of the real performance venue - full details of these SRIR recordings are found in 3.4.1 and the convolution process outlined in section 5.5.1. Stimuli were between 4 and 10 seconds long.

The results of Pilot Test I (see section 5.4) suggested that some listeners were able to distinguish between convolved (virtual acoustic) recordings produced with "trimmed" and "un-trimmed" source material, meaning that differences in the level and decay of early reverberation might be noticeable and lead the listener to make distinctions between the recorded fragments. In order that the salience of the room acoustic characteristics of the

performance venue were minimized, participants in the listening test were asked to attend to the content of the details of the singing performance itself, rather than listening to the room acoustic characteristics of the recordings.

6.3.2 Procedure

A series of four listening tests was undertaken as follows:

Test 232a Bass voice, "Que le pardon, que la clemence" : 5 fragments

Test 232a Bass voice, "Why, why has thou robbed me of my rest?" : 8 fragments

Test 221 Tenor voice, "De domo pulsus regali" : 11 fragments

Test 212 Mezzo-soprano voice: "Down a down a down, hey, down" : 34 fragments

Details of the fragments used as stimuli in all tests are found in Appendix G.

Dissimilarity ratings between fragments were obtained via a sorting task using Sonic Mapper software [294] as described in Section 5.5.1. Sonic Mapper ran in Windows 7 on a custom-built PC (Intel Core i7 3GHz, 12GB RAM) with a RME Fireface 800 soundcard, and audio presented over Beyerdynamic DT990 closed-back headphones. Average completion times for the tests ranged from 7 minutes for Test 232a to 25 minutes for Test 212.

6.3.3 Data Analysis

It was shown in Pilot Listening Test II (Section5.5) that Multi-Dimensional Scaling (MDS) was a suitable method of visualising the multi-dimensional data of participants' ratings of dissimilarity between the recorded singing fragments.

Aggregated dissimilarity matrices for each listening test were produced in MATLAB using the *mdscale* function included in the Statistics Toolbox^{TM} [305]

Stress values for MDS solutions obtained using a number of different dimensions (1 to 8) were calculated using Kruskal's method of stress measurement [306] (See section 5.5.2) and Shepard plots for the 2 and 3-dimensional MDS solutions were produced (see section 5.5.2).

In order to aid interpretation of the MDS derived perceptual maps for each listening test, agglomerative hierarchical clustering (AHC) was carried out using the "average" linking method, which computes the unweighted average distance from all items in one cluster (group) to all the items in another group, meaning that clusters with small variances

are combined. Dendrogram plots are also produced to illustrate the clusters identified through this procedure, where the arrangements of the linking bars indicate which items are most similar to each other: the length of the joining branch relates to the degree of similarity between the items which form the "leaves", with shorter branches reflecting higher degrees of similarity.

In summary, for each listening test the following plots were produced.

- A scree plot of stress measure for 1 to 8 dimensions to allow the goodness of fit of the MDS solution to be evaluated in (presented in Appendix I)
- Shepard plots of the 2-dimensional and (where relevant) 3-dimensional MDS solutions
- Plots of the MDS solution "common perceptual space" in two and/or three dimensions, where appropriate
- A dendrogram representing the hierarchical clustering of fragments

A full list of fragments and recordings is included in Appendix G and all recorded fragments are available on the data CD accompanying this thesis. Appendix H gives a full list of listening test participants' comments on the fragments.

6.3.4 Results: Test 232a Bass

The screeplot for Test 232a in Appendix I Figure I.1 shows that a 2-dimensional MDS model represents this data excellently, and this is confirmed by the Shepard plot in Figure 6.2 which shows that the modelled disparities recreate the original distances extremely accurately. This is to be expected given the small number of fragments in this test.

The 2-dimensional MDS solution for Test232a is shown in Figure 6.3. The *SP* and *MR* fragments form a cluster distinct from the two *LC* fragments. The two *LC* fragments, real and virtual, differ from each other along dimension 2 but are similar to each other, and distinct from the other (*SP* and *MR*) fragments along dimension 1. The two *SP* fragments, real and virtual, are similar to each other in dimension 1.

Figure 6.4 shows one group clearly, with *SPrealV1*, *SPvirtualV1* and *MRvirtualV1* closely grouped together with *LCrealV1* and *LCvirtualV1* forming the only other cluster.

Discussion

Listeners remark that *LCrealV1* and *LCrealV2* have similar phrasing and vibrato and that these fragments seem more theatrical and powerful than the others. A number of listeners commented on the timbre of vowels, level of loudness and articulation of the text.

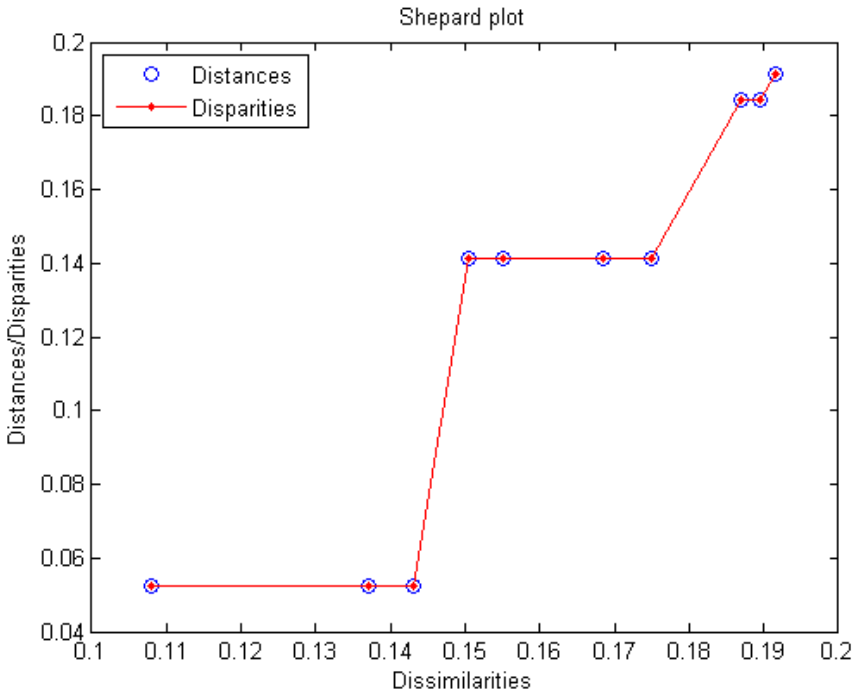


Figure 6.2: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232a modelled in 2 dimensions*

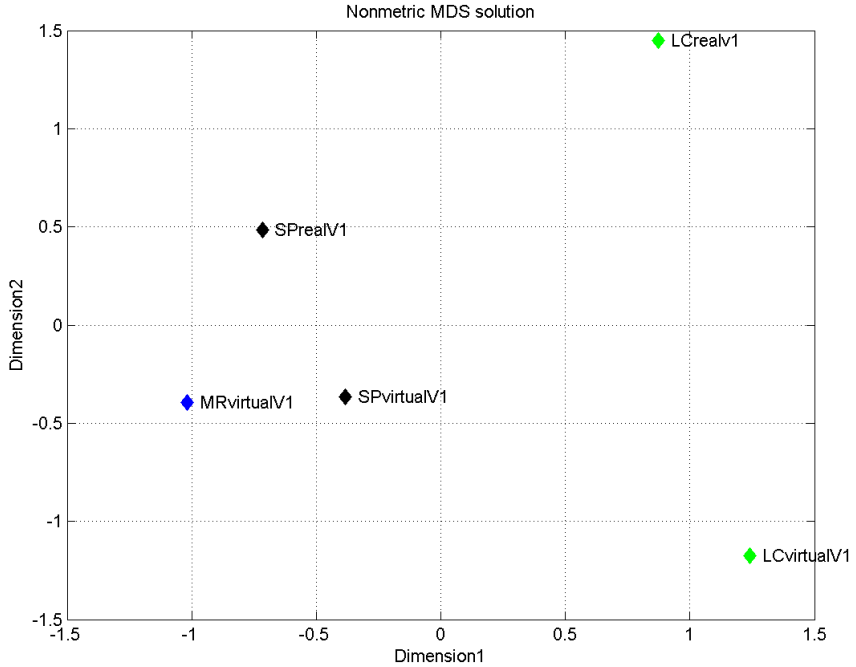


Figure 6.3: *Nonmetric MDS solution for listening test 232a, modelled in 2 dimensions*

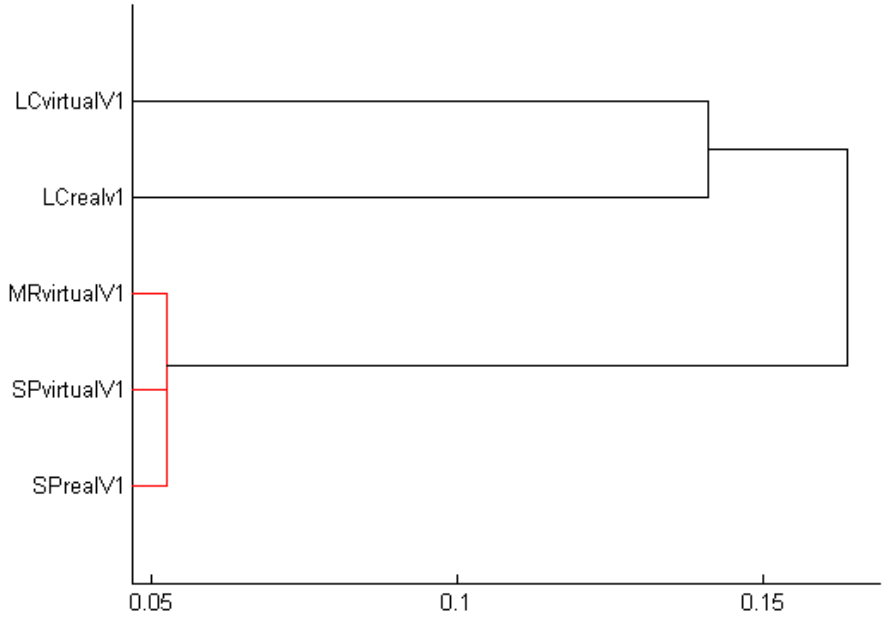


Figure 6.4: *Dendrogram of hierarchical clustering of fragments from a 2-dimensional MDS solution for test 232a*

Some listeners suggest that the group of *SPRealV1, MRVIrtualV1* and *SPVirtualV1* includes performances that are more uniform in delivery where both halves of the phrase are similar to each other, as opposed to the dramatic LC fragments where there is a greater contrast between the two phrases in each fragment.

6.3.5 Results: Test 232b Bass

The scree plot in Appendix I Figure I.2 shows that a 2-dimensional solution gives a fair representation of the dissimilarities between fragments, but that adding a third dimensional means that the model fits exactly. The Shepard plot in Figure 6.5 of the 2-dimensional MDS solution shows that although the transformation of dissimilarities to disparities is monotonic, the mid-range distances are not as well represented by the modelled disparities as in the solution for Test 232a, which can be seen in the scattering of blue circles (representing distances) around the red line (representing disparities).

A 2-dimensional MDS solution for this data is plotted in Figure 6.6. There is an obvious clustering of *MRvirtualV1*, *LCvirtualV1* and *SPvirtualV1* in the central upper portion of the plot, which is also seen in the close grouping of these fragments in Figure 6.7.

There are no other obvious clusters or tightly similar groups shown in the dendrogram.

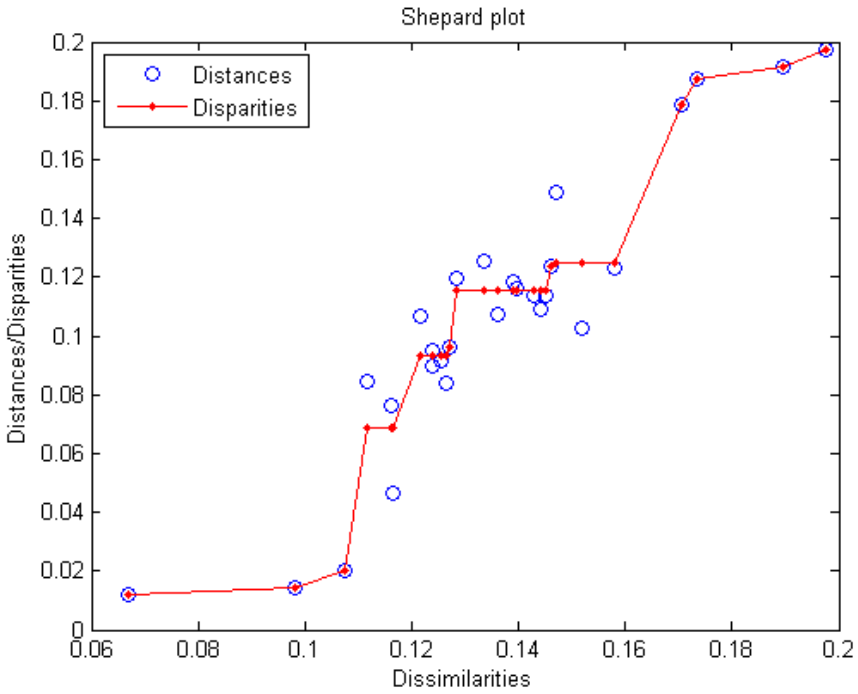


Figure 6.5: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232b modelled in 2 dimensions*

However two main groupings of fragments do seem to arise, one of which (denoted in red on figure 6.7) only includes fragments from recordings made in the virtual space. *LCrealV1* is highly dissimilar to all the other fragments although most listeners do not distinguish this fragment in their comments, with only one listener suggesting that the articulation of the final consonant cluster "st" of the final word was a distinguishing feature.

The Shepard plot in 6.8 for the 3-dimensional MDS solution for the same data, shows an exact matching of original distances to modelled disparities; the blue circles are all placed exactly on the blue line.

The MDS solution modelled in three dimensions for this data is plotted in Figure 6.9 and a plot of dimension 1 vs dimension 3 is found in 6.10 to aid visualisation.

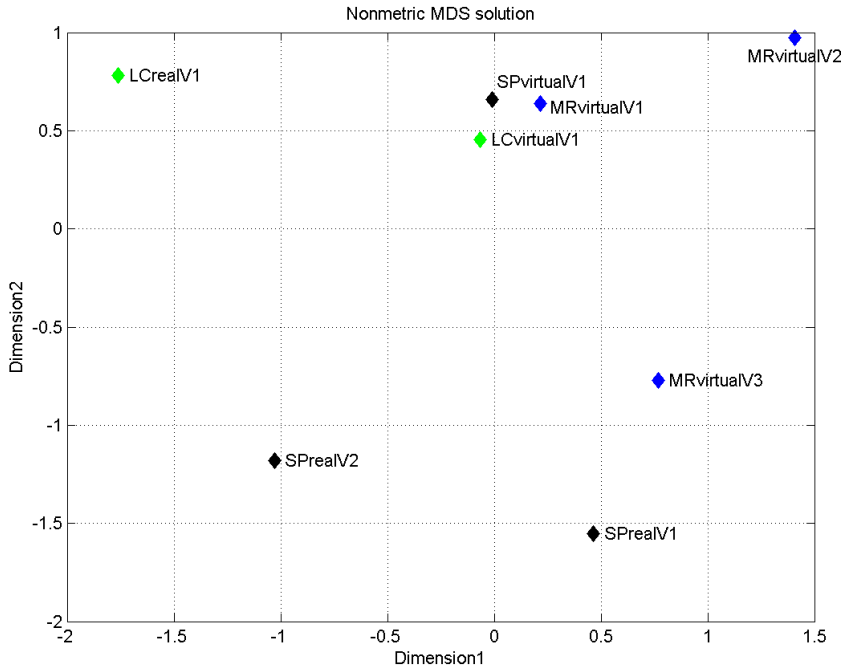


Figure 6.6: *Nonmetric MDS solution for listening test 232b, modelled in 2 dimensions: dimension 1 and dimension 2 plotted*

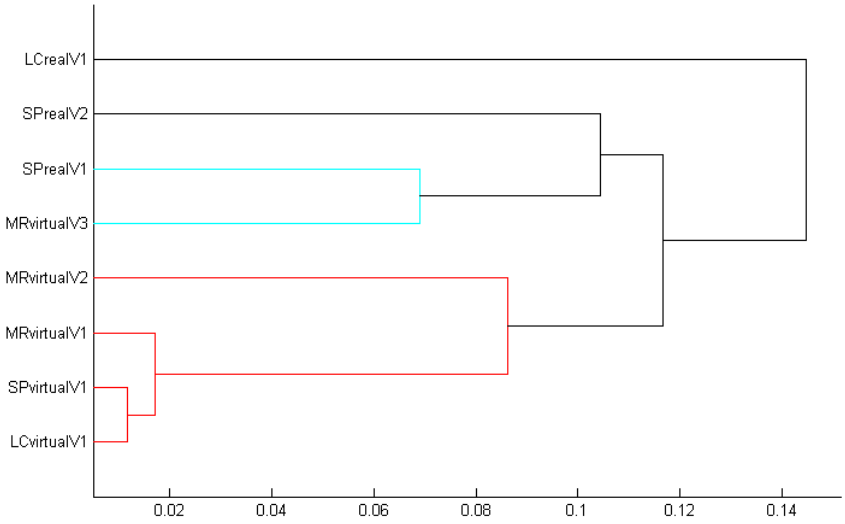


Figure 6.7: *Dendrogram of hierarchical clustering of fragments of a two-dimensional MDS solution for test 232b*

The dendrogram shows two main groupings; the one at the top of the figure containing only fragments from the real performance space, whereas the other main group includes

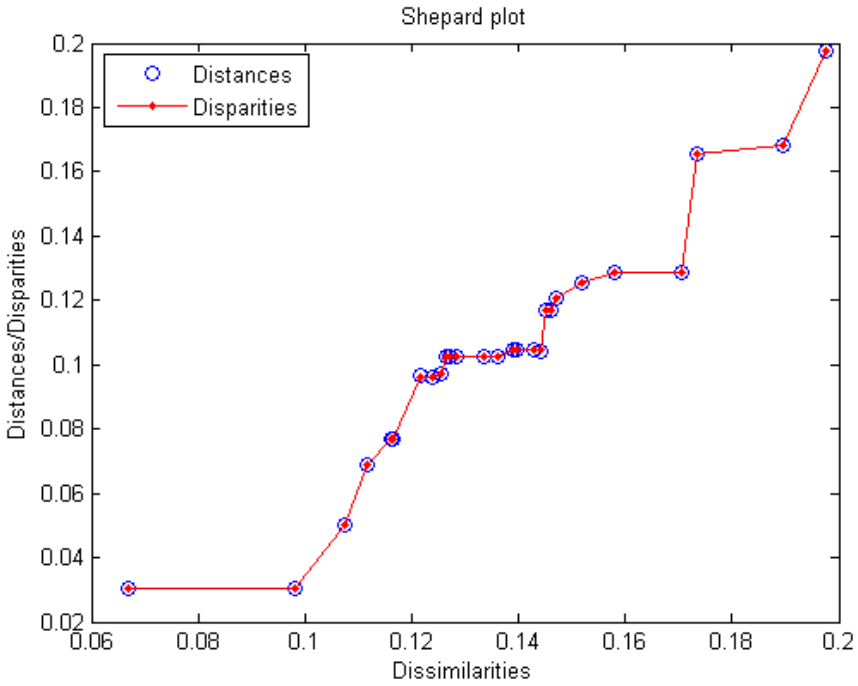


Figure 6.8: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 232b modelled in 3 dimensions*

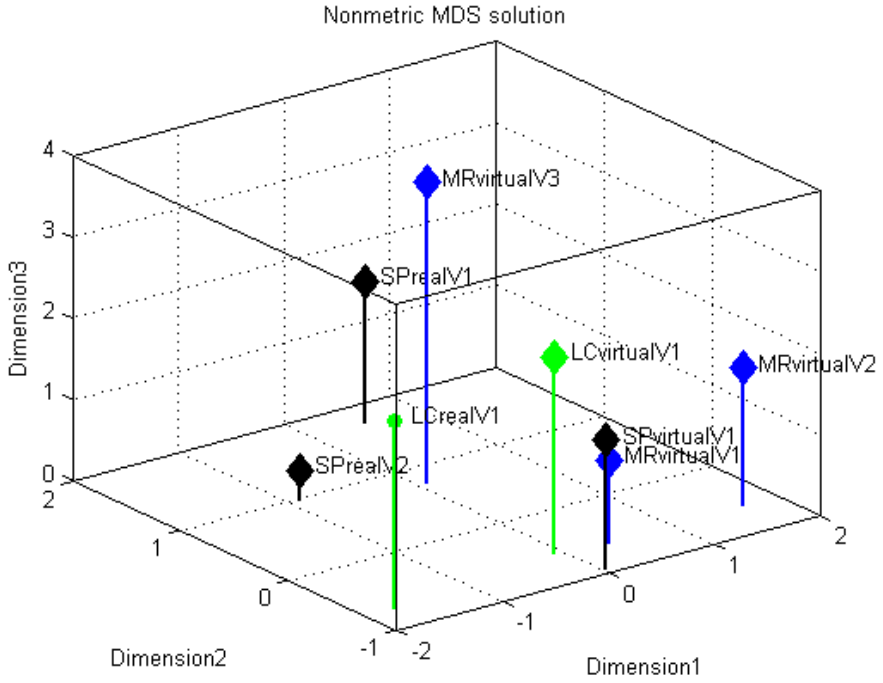


Figure 6.9: *Nonmetric MDS solution modelled in three dimensions for listening Test 232b*

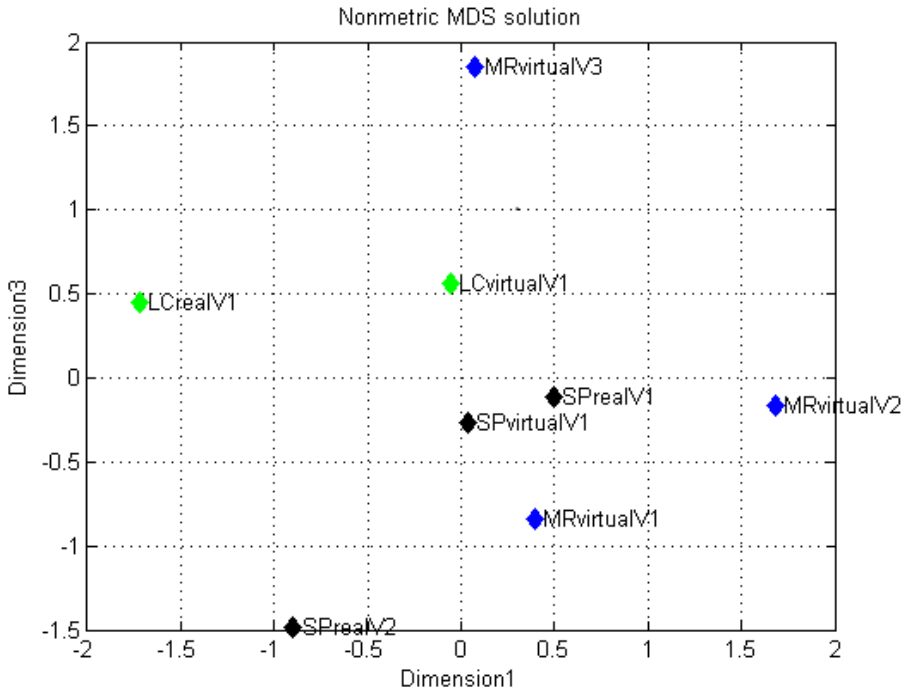


Figure 6.10: *Nonmetric MDS solution for listening test 232b, modelled in 3 dimensions: dimension 1 vs. dimension 3*

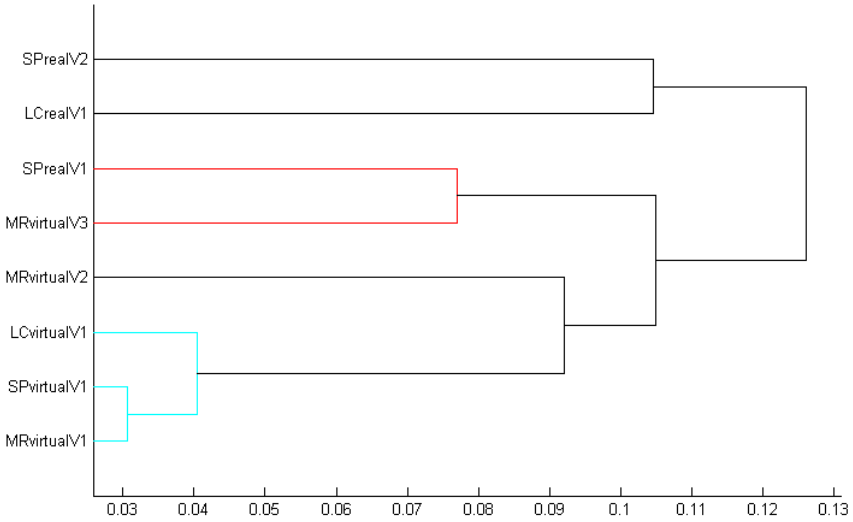


Figure 6.11: *Dendrogram of hierarchical clustering of fragments from 3-dimensional MDS solution for test 232b*

the majority of fragments from the virtual space.

Discussion

The LC and MR fragments are judged as similar to each other along dimension 2 where they are distinguished from the SP fragments. Within the SP fragments (*SPrealV1* and *SPvirtualV1*) are highly similar to each in both dimension 1 and 3, and the LC fragments are also similar in dimension 3, but distinct from each other along dimension 1.

There seems to be a distinction along dimensions 1 and 2 between most of the *virtual* fragments which are clustered towards the top right hand corner of the MDS plot (Figure 6.6). However, real and virtual fragments are clustered more tightly together along dimension 3 (see Figure 6.10) with *SPRvirtualV1* and *SPRealV1* judged as highly similar.

Some listeners comment on individual differences between the fragments such as the pronunciation of the word "rest" at the end of the phrase but most comment on more global aspects of the fragments such as the use of vibrato, tuning and length of notes. The group of *MRvirtualV1*, *LCvirtualV1* and *SPvirutalV1* have the pronunciation of "rest", brighter timbre and faster tempo in common.

6.3.6 Results: Test 221 Tenor

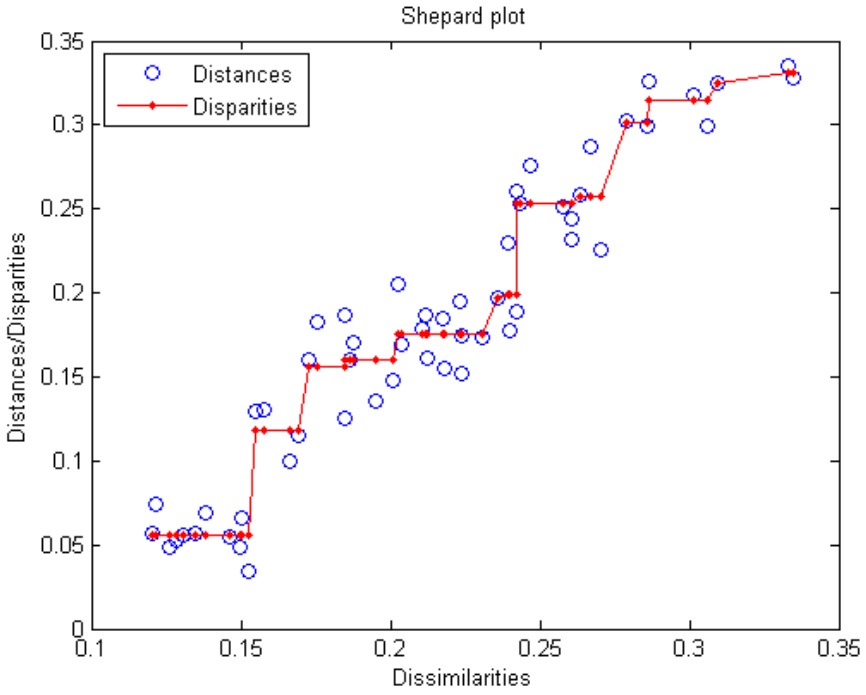


Figure 6.12: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 2 dimensions*

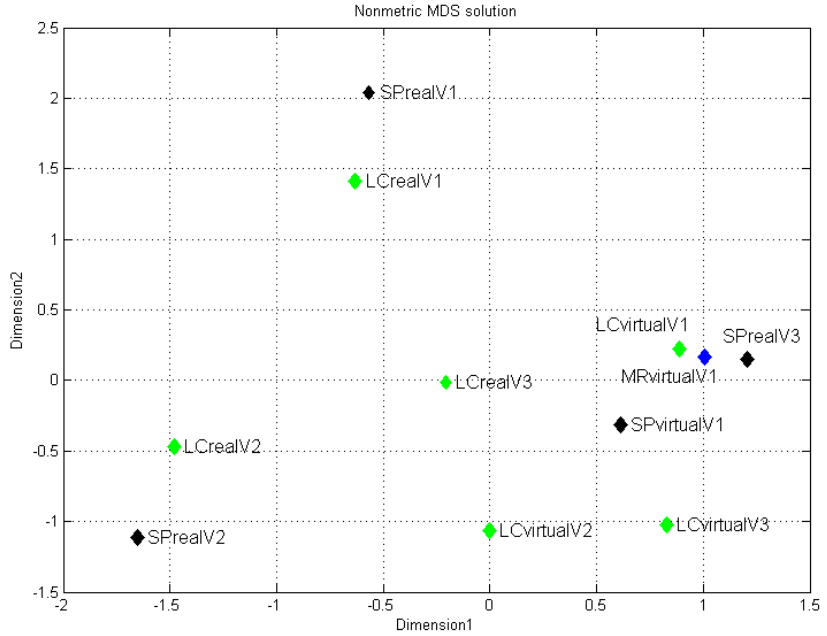


Figure 6.13: *Nonmetric MDS solution for listening test 221, modelled in 2 dimensions*

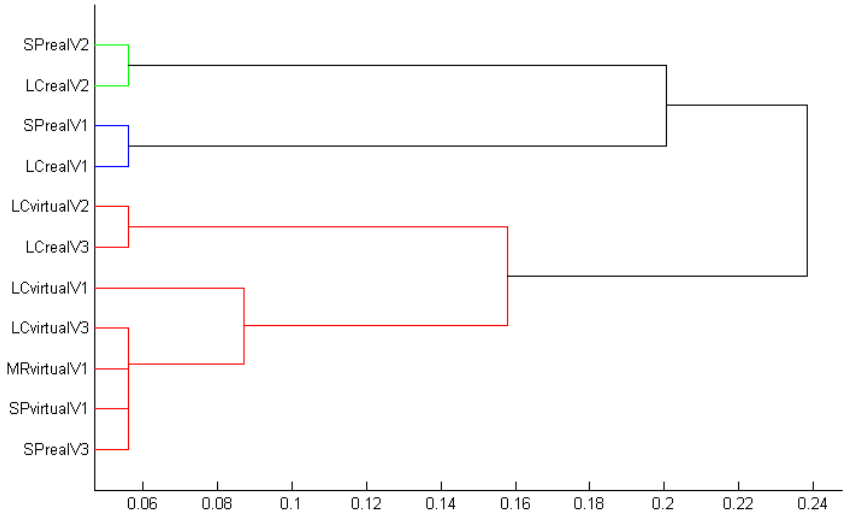


Figure 6.14: *Dendrogram of hierarchical clustering of fragments from 2-dimensional MDS solution for test 221*

The screeplot in Appendix I Figure I.3 shows that a two-dimensional MDS solution gives only a fair representation of the dissimilarity data for this test, whereas a three-dimensional solution gives a better representation. This is reflected in the Shepard plots in Figures 6.12 and 6.15 for the two and 3-dimensional solutions respectively.

Figure 6.12 shows that modelled distances are somewhat scattered about the red line reflecting a certain amount of variability in the match between the original dissimilarities and the modelled solution.

In the 2-dimensional MDS solution in Figure 6.13 there is one very clear clustering of *LCvirtualV1*, *MRvirtualV1* and *SPrealV3* which can be seen towards the centre right-hand side of the plot. This tight cluster forms part of a larger group including mostly both real and virtual versions recorded in the LC acoustic setting which is also seen as the red grouping in the dendrogram in Figure 6.14.

LCrealV2 and *SPrealV2* are grouped tightly together in the bottom left-hand corner of the perceptual map, whereas *SPrealV1* and *LCrealV1* are close together in the centre top of the plot - again these groupings are also shown as closely similar fragments at the top of the hierarchical clustering dendrogram.

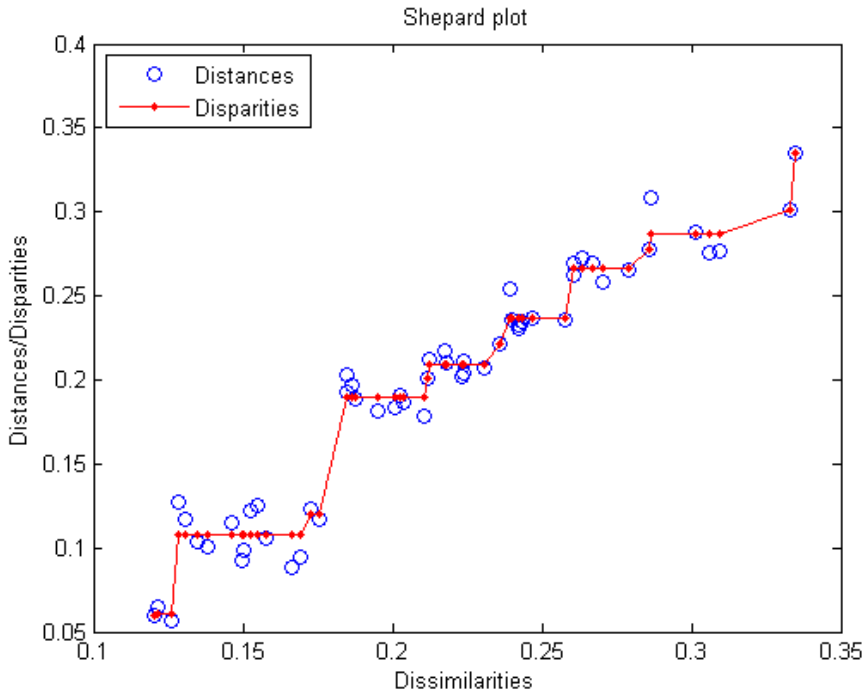


Figure 6.15: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 3 dimensions*

The 3-dimensional solution illustrated in Figure 6.16 shows that most of the LC fragments are similar to each other along dimension 3 and differ from three out of four of the SP fragments in this dimension. *SPrealV3* and *SPrealV1* are grouped closely together and are similar to *SPrealV1* along dimension 3. The single MR fragment sits between the SP fragments and the LC fragments in dimensions 3.

Although *SPrealV1* and *LCrealV1* are close together in the 2-dimensional MDS solution

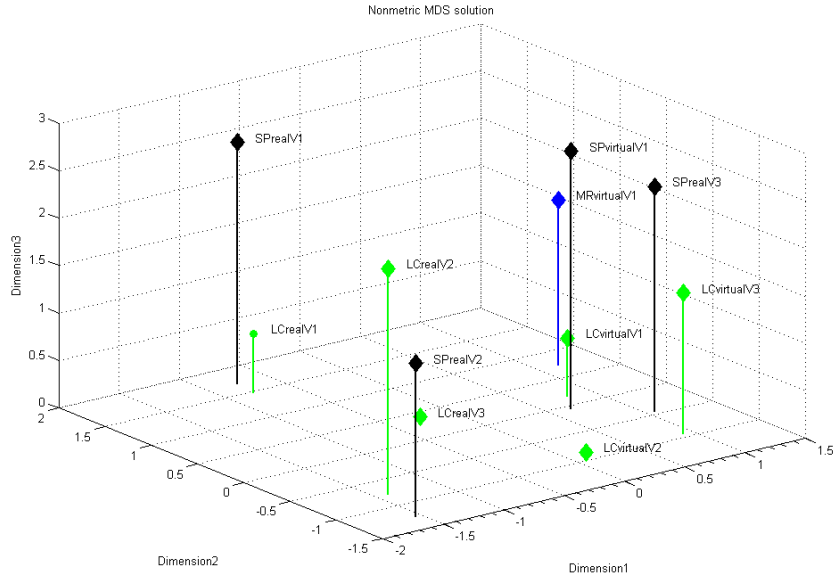


Figure 6.16: *Nonmetric MDS solution for listening test 221, modelled in 3 dimensions*

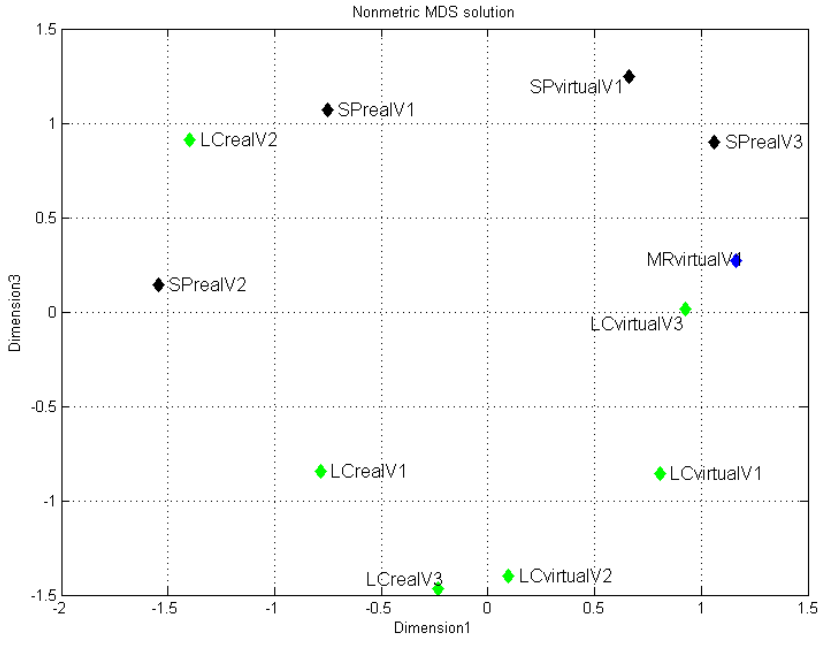


Figure 6.17: *Nonmetric 3-dimensional MDS solution for listening Test 221, Dimension 1 vs. Dimension 3*

of this data (see Figure 6.13) in the three-dimensional solution they are seem to differ along dimension 3. The majority of the virtual fragments are distinguished from the recordings in the real space along dimension 1.

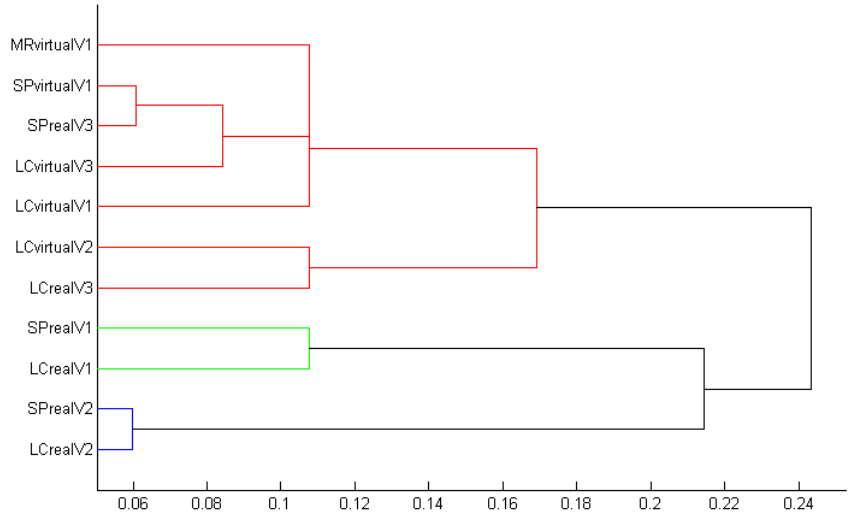


Figure 6.18: *Dendrogram of hierarchical clustering of fragments from the 3-dimensional MDS solution for test 221*

Figure 6.18 confirms the two tight clusters each containing a real and virtual fragment together: *SPvirtualV1* with *SPrealV3* and *SprealV2* with *LCrealV2*. The dendrogram also shows two main groups, with *SPrealV1*, *LCrealV1*, *SPrealV2* and *LCrealV2* forming one group contrasting with the other fragments.

Discussion

The singer reported that he had sung two very different interpretations of the piece for this test (see Appendix F for the score). One interpretation was from the "unmeasured score", that is, where the rhythm of the notes is not specified (the way that this medieval piece was originally written); the other interpretation is from a modern rhythmic transcription. "Rhythmic" transcription interpretations (namely *SPrealV1*, *LCrealV1*, *SPrealV2* and *LCrealV2*) are slower in tempo with clear "dotted" rhythms for pairs of notes.

Figure 6.18 does indeed show two main groupings for these fragments, one of which includes *SPrealV1*, *LCrealV1*, *SPrealV2* and *LCrealV2*, namely all the "rhythmic" interpretation versions. Listener comments also reflect these differences in tempo and 'legato" singing.

6.3.7 Results: Test 212 Mezzo-Soprano

Fragments for this listening test were taken from the repeated "Down-a-down hey down, hey down" stanza which appears during each verse of the song. Fragments are labelled

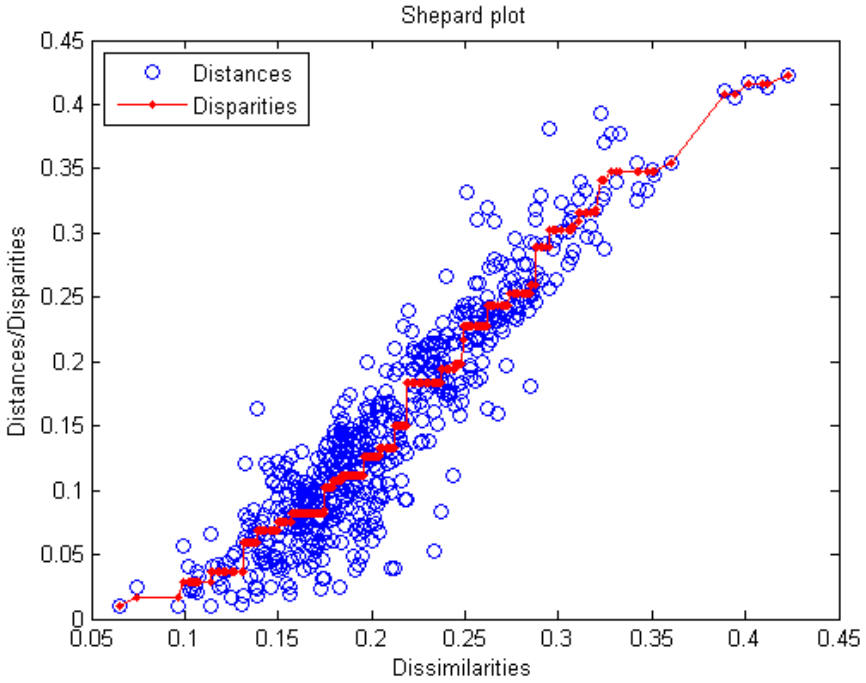


Figure 6.19: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 212 modelled in 2 dimensions*

according to acoustic setting (LC/MR/SP and Real/Virtual) but also include an indication from which verse of the song they were taken e.g. *LCvirtualB3* is a fragment taken from Verse B of recording 3 made in the LC setting of the virtual performance space.

The screeplot in Appendix I Figure I.4 shows that a three-dimensional MDS solution allows a much better fit of the data; the elbow of this screeplot is at 3 dimensions. The better fit of the three-dimensional MDS solution is confirmed by the Shepard plots in Figures 6.19 and 6.21 relating to the two-dimensional and three-dimensional solution respectively. The three-dimensional solution Shepard plot shows a more compact clustering of the modelled distances about the original disparities, and the transformation of the dissimilarities into disparities/distances is almost linear.

The two-dimensional MDS solution for this data is shown in Figure 6.20. There is one clear grouping towards the right hand side top of the plot which are all fragments taken from Verse E of the song. The LC versions in this group differ from the MR versions along dimension 1.

There is a central grouping of a mix of MR, SP and LC fragments taken from the different verses and recorded in both real and virtual environments. There does not seem to be any clear patterning of fragments recorded in the different acoustic configurations.

Two fragments (*LCrealC6* and *MRvirtC8*) in the top left-hand corner of the plot seem

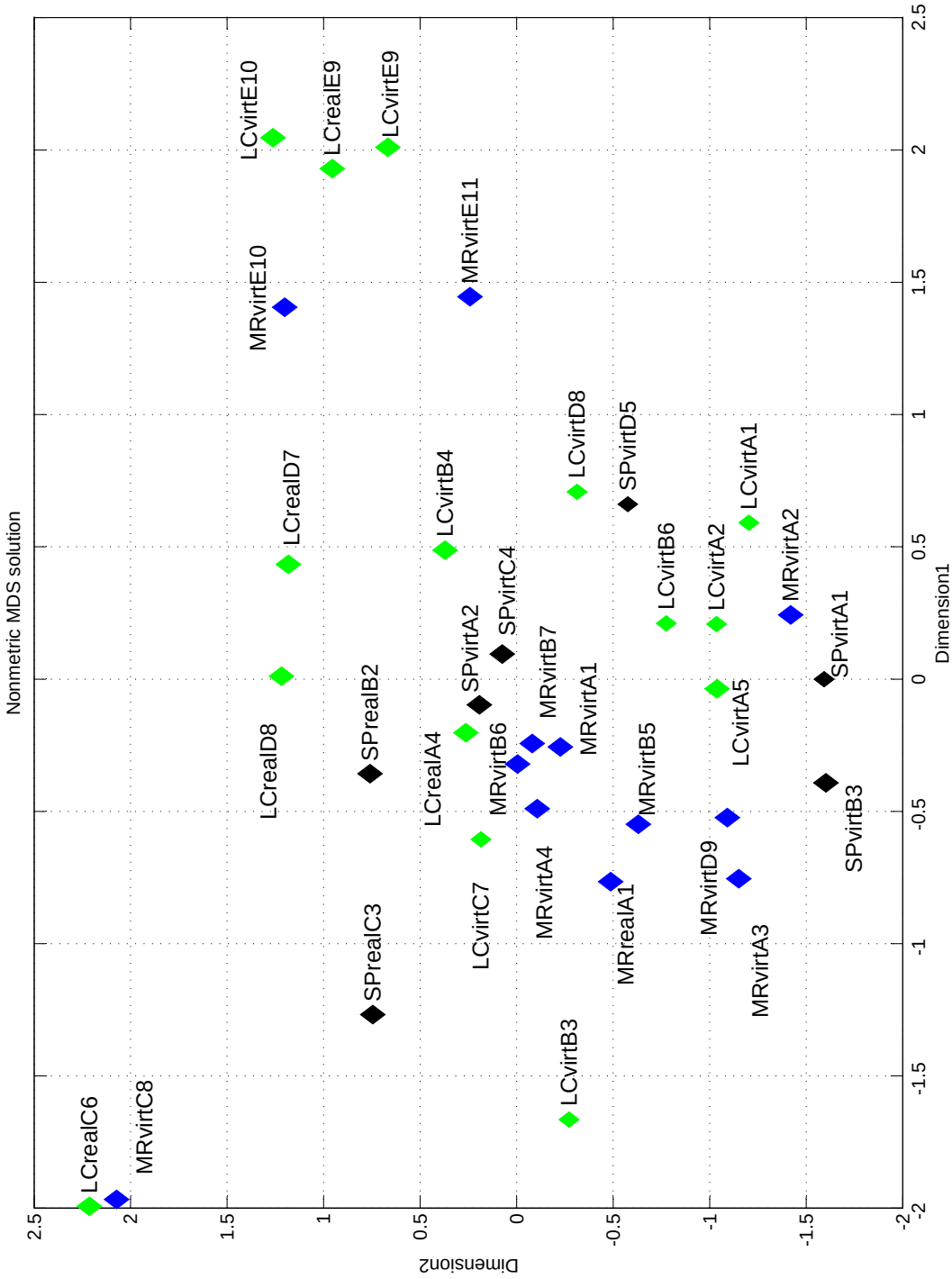


Figure 6.20: *Nonmetric MDS solution for listening test 212, modelled in 2 dimensions*

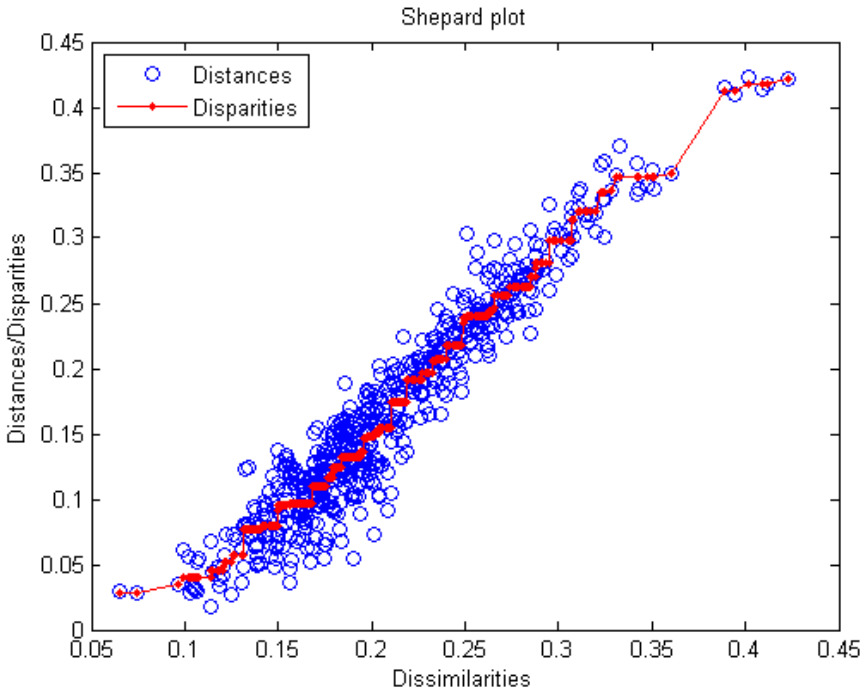


Figure 6.21: *Shepard plot displaying the relationship between original dissimilarities, distances and disparities of the point configuration for test 221 modelled in 3 dimensions*

to be highly similar but distinct from all the other fragments and might be classed as outliers. Indeed most listeners comment on these two fragments in terms of heavy vibrato, use of rubato, and emotional intent such as "power", "drama" and "anger".

The dendrogram in 6.24 gives a slightly clearer picture of the groups within this MDS solution. The verse E fragments are grouped together in blue at the top of the dendrogram. Next comes a large group (in green) which contains a mix of verse fragments A to D with no obvious patterning between verses, acoustic settings or environments. The third group (red) contains mainly fragments from verse C.

LCvirtB3 is not closely grouped with other fragments and indeed listeners remark on the "wiffy tuning", rich tone, greater vibrato, pronounced articulation of "d's" and husky breathy tone which seem to distinguish this fragment from the others from verse B. (see Appendix H)

Discussion

The MDS solutions for this test are complex and difficult to interpret since there is a mixture of fragments taken from different verses of the song. If MDS solutions for verse D and Verse E are plotted separately in Figures 6.25 and 6.26 then a slightly clearer picture emerges.

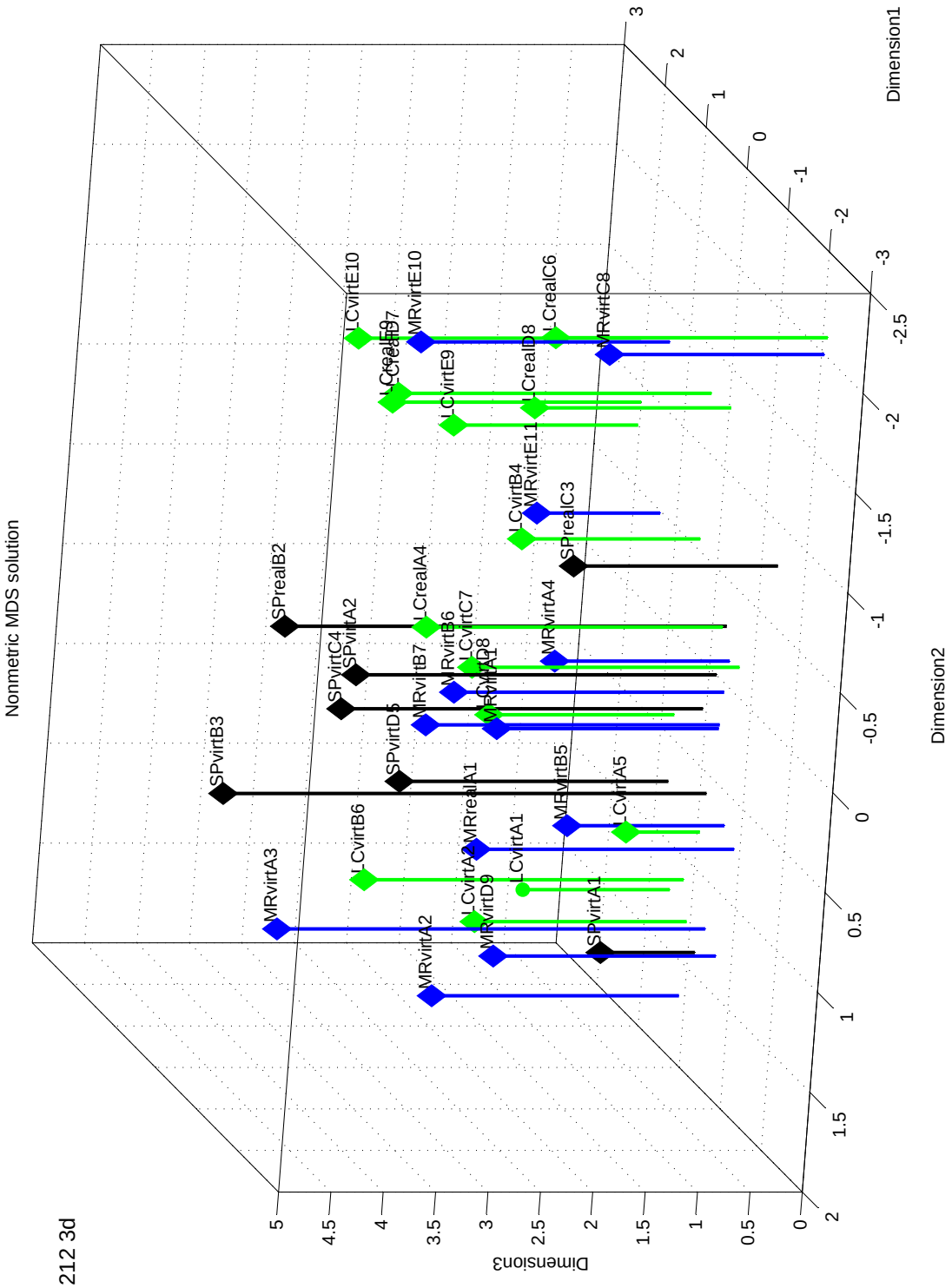


Figure 6.22: *Nonmetric MDS solution for listening test 212, modelled in 3 dimensions*

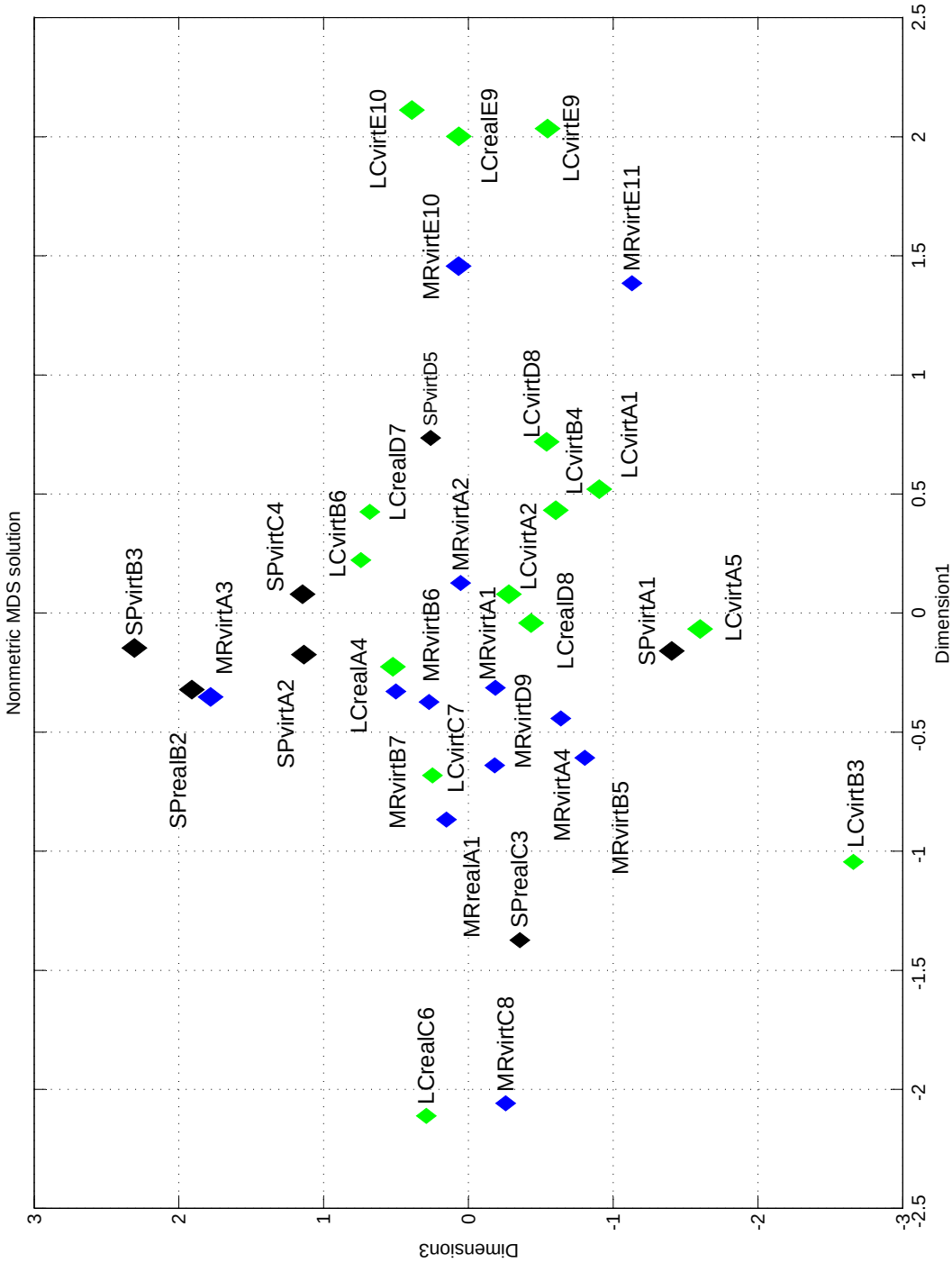


Figure 6.23: *Nonmetric MDS solution for listening Test 212, Dimension 1 and Dimension 3*

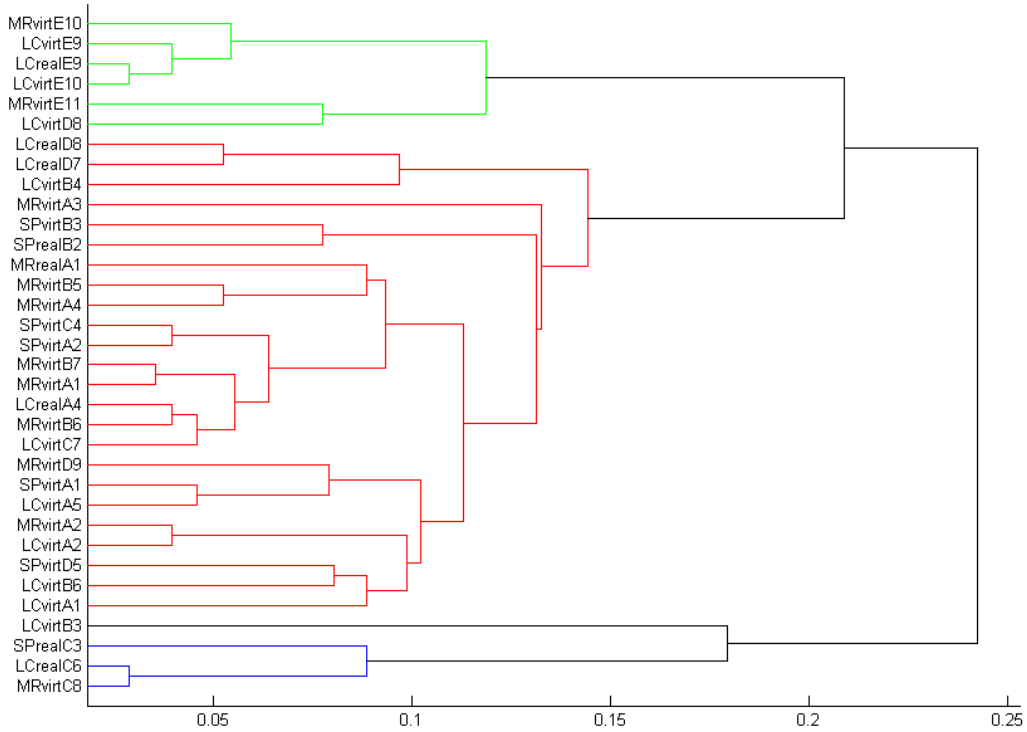


Figure 6.24: *Dendrogram of hierarchical clustering of fragments from the 3-dimensional MDS solution*

The screeplot of the goodness of fit for Verse D fragments (Appendix I, Figure I.5)shows that the MDS 2-dimensional solution represents the similarity data excellently.

There is no obvious grouping of fragments, but the SP fragment seems to differ from the others along dimension 2. Listener comments suggest that this distinction could lie with the "slow attack" of the SP fragment and the "powerful" middle portion of the phrase (see Appendix H).

Within the Verse E fragments LC productions seem to be distinct from MR productions; listeners remark that LC fragments of this verse are slow, long, continuous, soft, slow, gentle, with less rich tone and less vibrato whereas the MR productions have a richer tone, more vibrato, with a slow attack and powerful middle section.

6.3.8 Limitation of Sorting Task

A recognised problem with using a sorting task (implemented in Sonic Mapper software) to collect similarity ratings is that it only allows listeners to arrange the sound items only on a 2-dimensional plane [294]. This may unnaturally force a listener to regard the

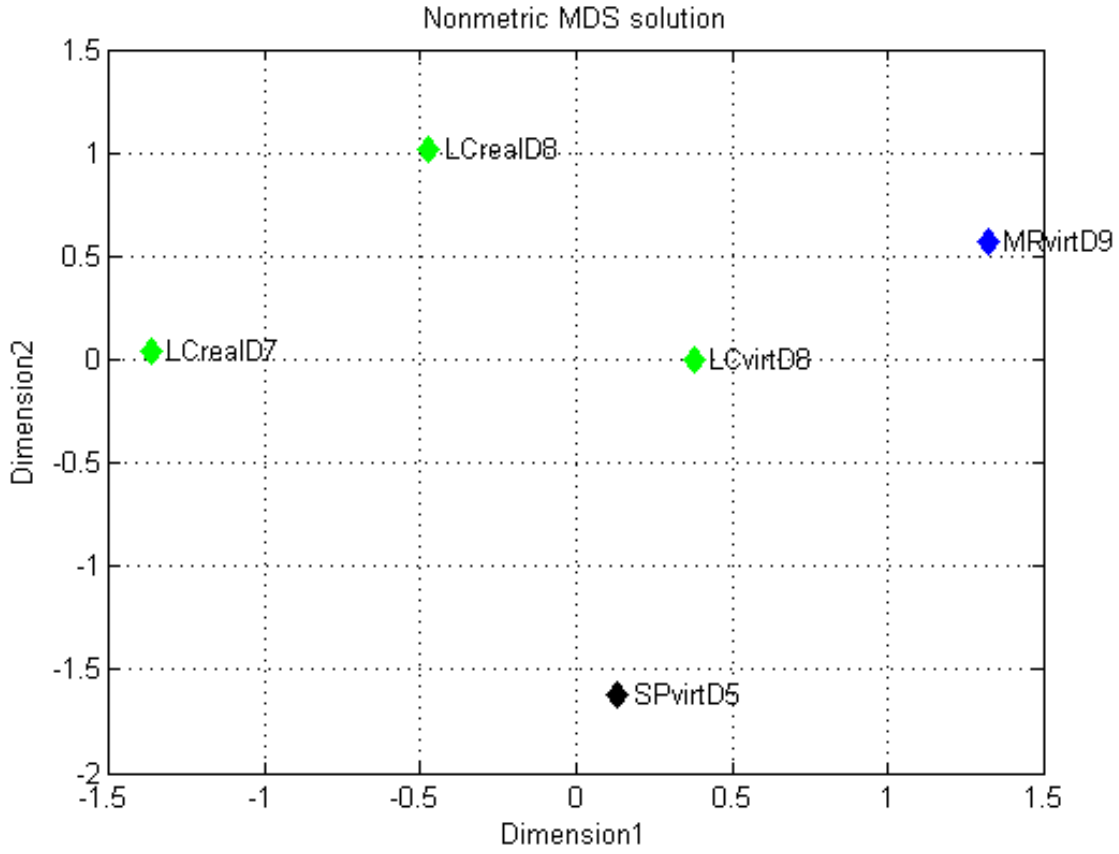


Figure 6.25: *Nonmetric MDS solution for listening test 212 Verse D fragments, modelled in 2 dimensions*

sound tokens in a 2-dimensional way e.g. only attending to two attributes at any one time, such as "tuning" and "tone", whereas a 3-dimensional representation might be more appropriate. Some participants did report that they felt this hampered the sorting task they undertook, but most accommodated for this limitation by effective labelling of any grouping that they produced.

A small number of participants wanted to overlap some groups which they were unable to do with the software used, but again they worked around this limitation by labelling the groups they had made, and using descriptions for the individual sung fragments. The descriptions of individual fragments given by listening test participants are laid out in the table found in Appendix H

6.3.9 Summary of listeners' evaluation

In general MDS models of the dissimilarity ratings obtained from the perceptual sorting task seems to provide a reasonable solution in either two or three dimensions. In test

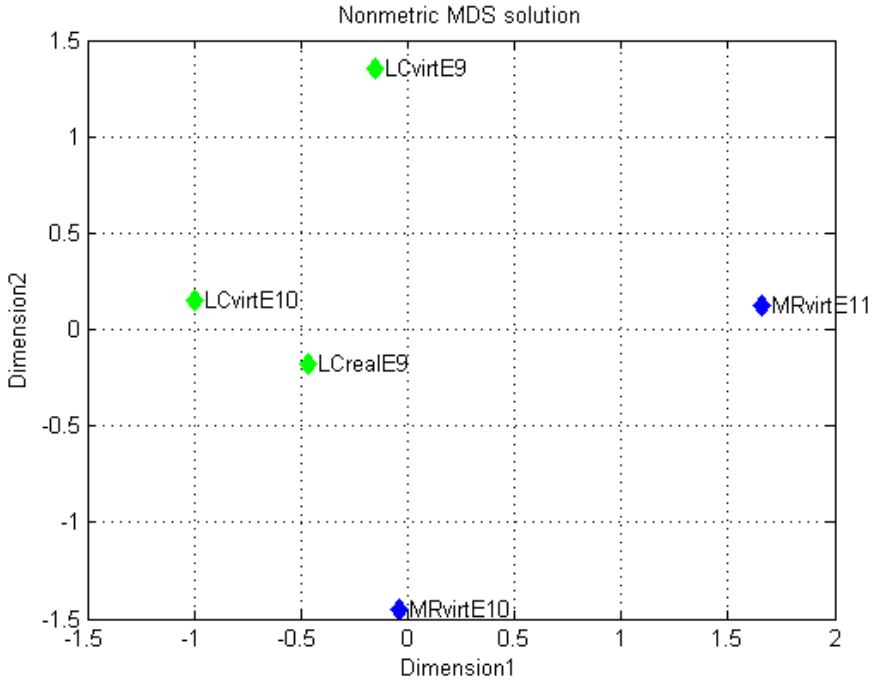


Figure 6.26: *Nonmetric MDS solution for listening test 212 Verse E fragments, modelled in 2 dimensions*

232a and 232b obvious clusterings of the fragments are apparent, whereas in test 221 and test 212 the groupings are not quite as clear cut.

Those fragments which are rated as highly dissimilar by listeners often have comments associated with them which pick out one or another particular aspect of the performance and examples have been cited above. Similarly those grouped closely together in the MDS solution are seen to share some common aspects identified by listeners in their comments.

On the whole production in the different acoustic configurations (LC vs MR vs SP) are seen to differ from each other. In all of the tests, SP and MR fragments pattern together in their distinction from LC fragments in some respect. In test 212 there are perceived dissimilarities between SP and MR performances within the parameters of each verse, but the greatest distinctions are judged between verses, which can be seen in the obvious grouping of all fragments from verse E.

The distinctions between productions recorded in the *real* and in the *virtual space* are less well defined. For example, for the bass singer in Test 232a and 232b fragments recorded in the real and virtual are judged as similar along dimension 1 but differ along dimension 2. In tests 232b and 221 there does seem to be a distinction between real and virtual fragments, at least in one of the three dimensions modelled. On the other hand real and virtual fragments are judged as similar in test 221 in dimension 1, and in test

212 real and virtual fragments are often grouped closely together.

It is also the case that *real* and *virtual* versions are often judged as similar between and across the acoustic configurations e.g. an *SPreal* fragment is sometimes judged similar to *LCvirtual*. However, on the whole, there is no clear distinction or patterning of the fragments recorded in the *real* vs. the *virtual* space.

MDS solutions offer an insight into the perceived similarities of the sung fragments performed in the different acoustic settings and virtual and real spaces. It should be noted that, since auditory attributes of the stimuli were not specified and a compositional approach to MDS was employed, the dimensions modelled by the MDS analysis are not necessarily the same between the different tests.

In order to try to infer what the dimensions of the modelled perceptual common space might represent, objective parameters of the sung performances must be analysed (Section 6.4). Quantitative measures of the singing performances can then be combined with the MDS perceptual maps as is described in Section 6.5.

6.4 Acoustic Analysis of Solo Singing Performances

In section 4.6.4 a number of musical performance attributes were highlighted as those which have been shown to alter according to the room acoustic characteristics of the performance venue.

Previous research in this area has studied changes in *tempo*, *vibrato*, *intonation*, *pitch glides* (portamento), as well as *spectral characteristics*. Vocal quartet singers who took part in the case study described in Section 4.7 identified changes in their own singing performance in the different acoustic as changes in *intonation* and *timing*. The solo singers who were recorded in the real and virtual performance spaces identified most often *tempo*, *length of notes*, *intonation* as changes in their own singing (see Section 6.2.2)

Participants who made comments in the listening test (Section 6.3) commented on *dynamics*, *tempo*, *vibrato*, *articulation of the text* (See appendix H).

It was decided that two of the most frequently mentioned attributes by singers, namely intonation and tempo, would be investigated as possible objective measures to explain the groupings as obtained through the MDS analyses (Section 6.3.4). In addition, due to the good amount of previous research that has established vibrato as a performance attribute which is altered in different acoustic environments, vibrato rate and extent were also evaluated for the fragments.

The Automatic Music Performance Analysis and Comparison Toolkit (AMPACT) developed by Devaney and Mandel [151] is used for initial segmentation of the audio waveform to identify note positions. AMPACT is a MATLAB toolbox which was specifically

developed for singing voice recordings where a score of the performance is available. It is optimised to provide note onset and offset estimates for tones with non-percussive onsets, such as singing voice and incorporates the Hidden Markov Model MATLAB toolkit by Kevin Murphy [307] and the Dynamic Time Warp MATLAB toolkit by Dan Ellis [308]. Additional MATLAB scripts were developed for this project to aid analysis of tempo, vibrato and intonation as described below.

6.4.1 Method

The original source material taken from the head-mounted microphone recordings of the fragments were used in the following acoustic analyses. A MIDI file for each of the 4 phrases used in the four listening tests was produced. This is used by the AMPACT toolkit to aid identification of note positions and guide the fundamental frequency estimation. Full details on the signal processing techniques used in AMPACT can be found in [150] and [145].

The lyrics of the sung phrase are coded, indicating the relative timings of musical rests (silence), voiceless consonants (transient) at the beginning or ending of a note, and vowels (steady state portions). This encoding is used to form a Hidden Markov Model (HMM) to guide the detection of note onsets and offsets, which improves the initial Dynamic Time Warp (DTW) matching of the MIDI file to the audio waveform. An illustration of the states used in the HMM and the audio waveform of a sung phrase is given in Figure 6.27.

AMPACT makes an initial alignment of the MIDI file to the recorded audio waveform as illustrated in Figure 6.28.

If this initial alignment is not fully correct, alterations can be made by hand - in this case by inspecting the waveform in Sonic Visualiser [149] and outputting new values for note onsets and offsets which are then incorporated into the alignment in AMPACT. The improved alignment can then be visualised as before to check that timings are correct - for example see Figure 6.29.

A vector of states (steady-state, silence, transient) for each note of the phrase is produced for each sung fragment. The timing data contained in this vector facilitates the analysis of musical performance parameters which can be further analysed as described below.

Timings for each note are extracted from the matched MIDI file, and a MIDI file mirroring each performed fragment is produced. An illustration of the original (score-based) MIDI file and the performance-matched MIDI file is given in Figures 6.30 and 6.31.

Once the researcher is happy that the time points for note onset and offsets, vowels and consonant portions are correct, the YIN [138] fundamental frequency estimation algorithm

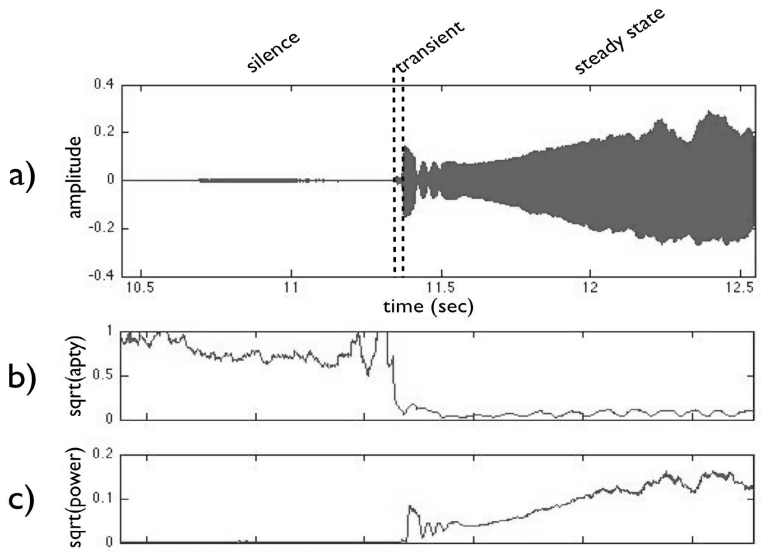


Figure 6.27: *Illustration of HMM states defined in the AMPACT alignment algorithm a) time domain representation of the sung note with HMM states labeled, b) aperiodicity measure from the YIN algorithm c) power measure from the YIN algorithm (figure used, with permission, from [150])*

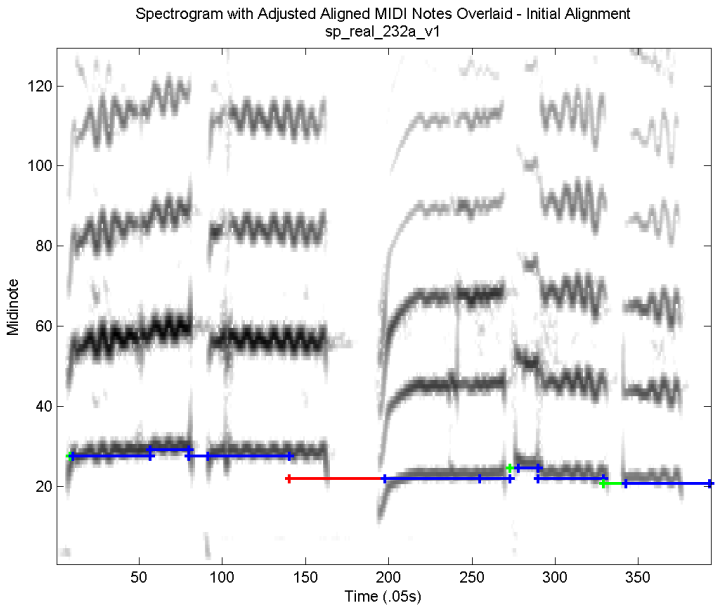


Figure 6.28: *Initial alignment in AMPACT of MIDI file timings (coloured bars)overlaid on a spectrogram of the: blue bars depict vowels, green bars depict voiceless consonants and red bars depict silences*

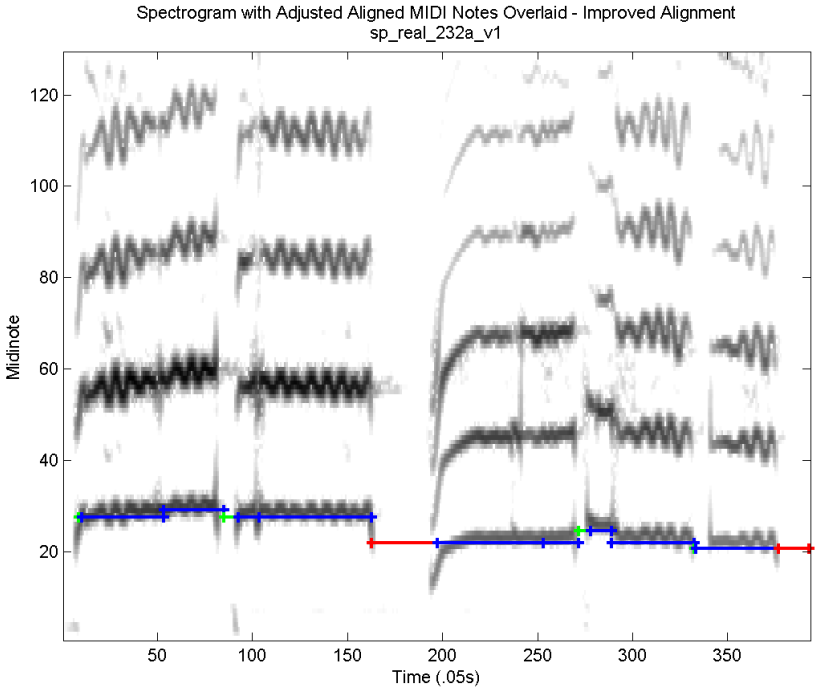


Figure 6.29: *Improved alignment after visual inspection and adjustments made in Sonic Visualiser. MIDI file timings (coloured bars) are overlaid on a spectrogram of the audio file : blue bars depict vowels, green bars depict voiceless consonants and red bars depict silences*

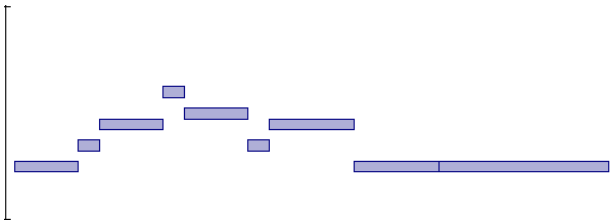


Figure 6.30: *MIDI representation of Test 221 phrase (De Domo) with timings taken from the score*

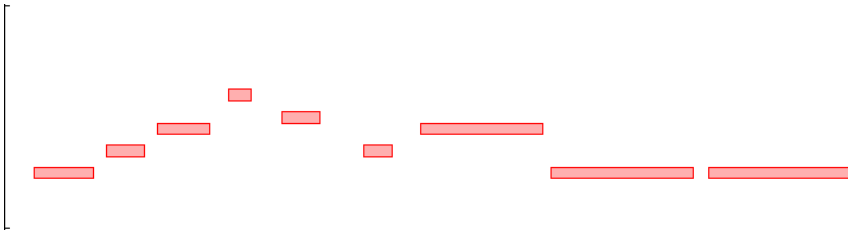


Figure 6.31: *MIDI representation of Test 221 phrase (De Domo) with timings as performed by the singer*

is applied. Values for estimated fundamental frequency are extracted for all steady-state (vowel and vowel-like) portions of each note and reported in octaves relative to A440Hz.

Onset timings and fundamental frequency values allow further analysis of tempo, vibrato and intonation to be undertaken as described below.

6.4.2 Analysis of Tempo

Method

Beats were identified by hand for each phrase according to the score and inter-onset-intervals between the first notes of each beat were calculated. Local tempo is the tempo value (expressed in beats per minute) for each beat identified in the phrase, the mean of these local beats is then calculated as global tempo.

The locations of beats for each phrase are indicated in Figures 6.32, 6.34, 6.36 and 6.38. Note that for Test 232a (Figure 6.32) bar durations were extracted from the note timings and a tempo value for each bar was calculated assuming two beats per bar.

Results

Local (beat-level) and global tempo (mean across phrase) for all fragments in each test are found in Appendix J.

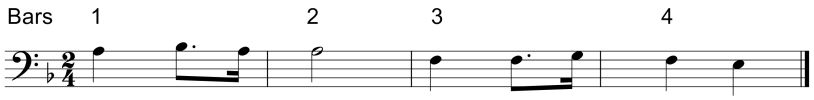


Figure 6.32: *Location of bars in phrase for Test 232a*

232a Global tempo over the five fragments ranged from 48.7 bpm to 53.5 bpm (Mean:52.4, SD 2.4). Local tempo for individual notes ranged from 42 bpm (beat 1) to 67.6 (beat 2).

232b Global tempo over the eight fragments ranged from 46 bpm to 58.4 bpm (Mean:51.0, SD 4.3). Local tempo for individual notes ranged from 30.9 bpm (beat 1) to 102.7 (beat 2).

221 Global tempo over the eleven fragments ranged from 59 bpm to 77.4 bpm (Mean:67.0, SD 7.2 . Local tempo for individual notes ranged from 33 bpm (beat 5) to 136 (beat 6).

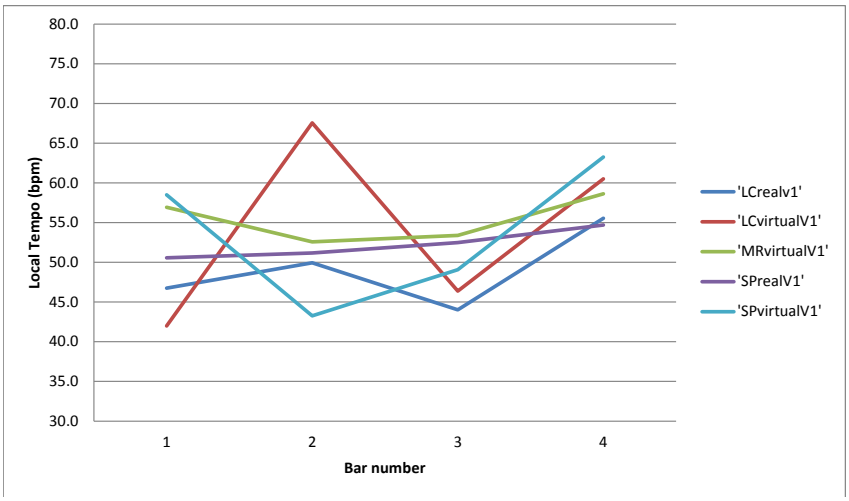


Figure 6.33: *Plot of local tempo (beats per minute) for bars 1-4 over 5 fragments in Test 232a*

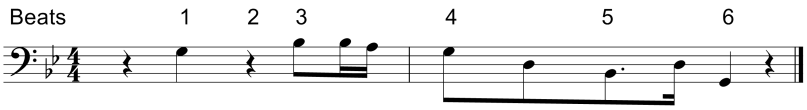


Figure 6.34: *Location of beats in phrase for Test 232b*

212 Global tempo over the 34 fragments ranged from 32.8 bpm to 62.3 bpm (Mean:51.3, SD 6.0. Local tempo for individual notes ranged from 26 bpm (beat 1) to 88.5 (beat 4).

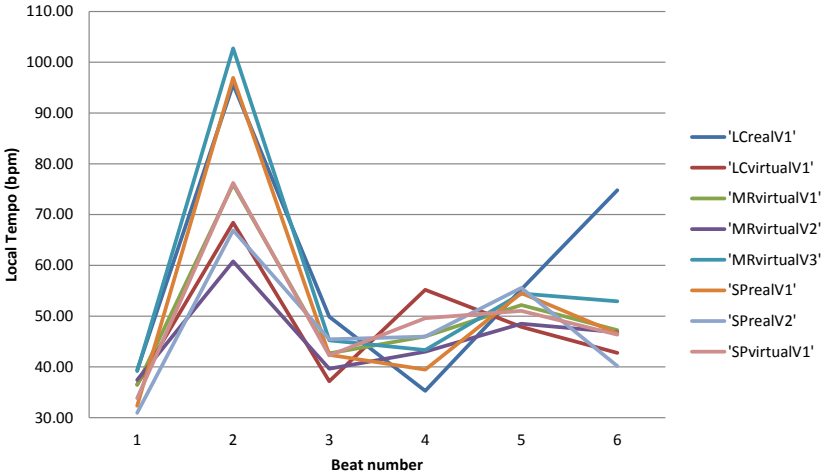


Figure 6.35: *Plot of local tempo (beats per minute) for beats 1-6 over eight fragments in Test 232b*

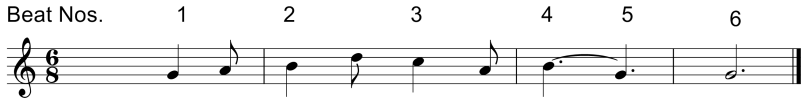


Figure 6.36: *Location of beats in phrase for Test 221*

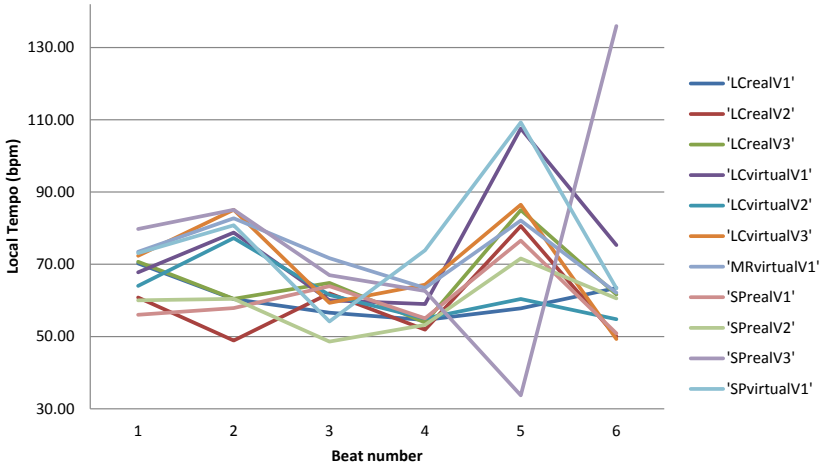


Figure 6.37: *Plot of local tempo (beats per minute) for beats 1-6 over eleven fragments in Test 221*

Discussion

The bass singer in tests 232a and 232b was most consistent in tempo with small standard deviation of global tempo of 2.4 bpm and 4.3 bpm respectively. The second and last beats of many of the phrases seem to be most variable in tempo with minimum and maximum local beat values appearing most often in these positions.

6.4.3 Analysis of Intonation

Method

Framewise fundamental frequency values output from the YIN toolbox [138] are retrieved for each steady-state portion of each note, using timing data obtained from the audio/midi alignment.

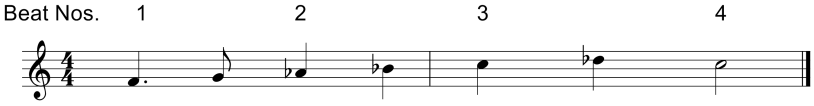


Figure 6.38: *Location of beats in phrase for Test 212*

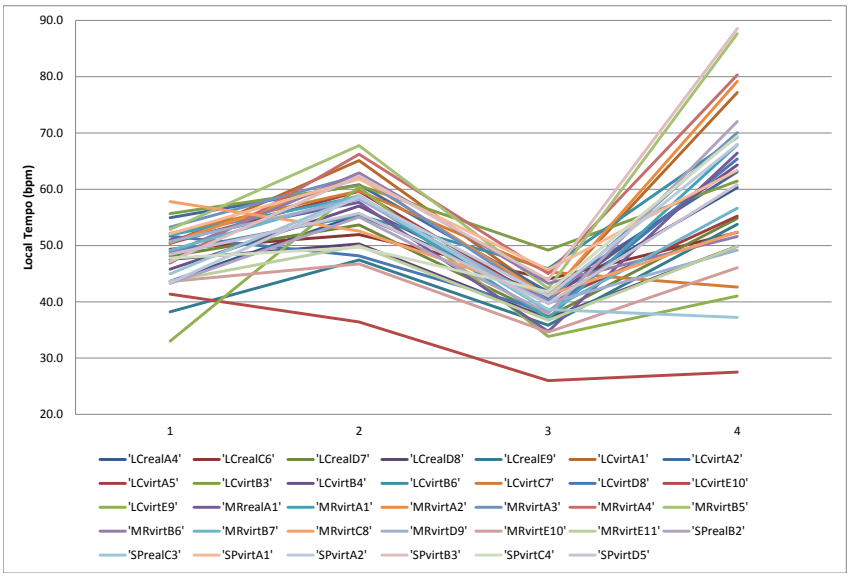


Figure 6.39: *Plot of local tempo (beats per minute) for beats 1-4 over 34 fragments in Test 212*

The pitch of each note of the phrase is estimated in AMPACT using a model of pitch perception for vibrato tones, based on work by Gockel, Morre and Carolyn [266], which calculates a weighted mean of the rate of change of the fundamental frequency. Frames of the audio waveform where the $F0$ changes more rapidly are given higher weighting than those where the $F0$ changes more slowly.

It should be noted that these pitch estimates are not fully comparable with the measure of "median pitch" used by Mauch et al. [251] and Dalla Bella et al. [234].

Three intonation metrics were computed for each fragment, one measure of accuracy and two measures of precision (see Section 4.3.11).

Mean Absolute Interval Error (MAIE) Mean Absolute Interval Error (MAIE), is calculated with respect to equal tempered intervals (ET) in the same way as in [251].

$$MAIE = \frac{1}{M-1}\sum_{i=2}^{M}|e_i^{int}| \tag{6.1}$$

where i is index of the note of the phrase in question, M is the total number of notes in the phrase and e_i^{int} is a measure of interval error, the difference between the produced interval and the target interval (see section 4.3.11 equation 4.5).

Mean Absolute Pitch Precision (MAPP) is computed by first determining a measure of pitch stability for each note in each fragment, by comparing the produced pitch with the average pitch for the corresponding note across all productions of the phrase (fragments).

$$s_i = p_i - mp_i \tag{6.2}$$

where s_i is pitch stability, p_i is the pitch of the i^{th} note of the fragment (phrase) and mp_i is the mean produced pitch for that note of the phrase. The mean of the absolute values of pitch stability are then calculated for each note of each fragment.

$$MAPP = \frac{1}{M}\sum_{i=1}^{M}|s_i| \tag{6.3}$$

MAPP used here is equivalent to *pitch stability* (measured by Dalla Bella et al. [234]) and similar in some respects to *note precision* (Pfordesher et al. [233]) except that *note precision* is calculated within pitch classes, whereas MAPP is calculated for each note of the phrase.

Mean Absolute Interval Precision (MAIP) is evaluated by first taking a measure of interval stability

$$s_i^{int} = \triangle p_i - \triangle p_i^m \tag{6.4}$$

where (s_i^{int}) is the difference between $\triangle p_i$, the size of the interval leading to the i^{th} pitch and $\triangle p_i^m$, the average interval size computed across all fragments for each interval in the phrase.

The mean of the absolute values of interval stability are then calculated for each note of each fragment

$$MAIP = \frac{1}{M-1}\sum_{i=2}^{M}|s_i^{int}| \tag{6.5}$$

MAIP is a measure similar to that of *Interval Precision* used by Pfordresher et al. [233] except that it is calculated for each individual interval of each phrase, rather than within an interval class.

It was decided not to calculate measures of MAPE (Mean Absolute Pitch Error) for two reasons. Firstly the singers in the study did not all use a pitch reference (tuning fork or electric tuner) when singing in the real and virtual spaces, so it cannot be guaranteed that they were always on "pitch" (relative to A=440Hz). Secondly, most of the recorded fragments were not taken from the beginning of a piece, and as such, it is possible that the singer may have already drifted in pitch by the time of the chosen fragment. Both of these issues, would mean that MAPE values would be unreliable, e.g. a sung fragment might achieve a high MAPE value (leading to the conclusion that the singer was not

accurate in producing the "correct " pitches), whereas this high inaccuracy score might in fact stem from the singer consistently "inaccurate" due to pitch drift, or starting the passage/song under- or over-pitch, (that is with an incorrect reference pitch).

It is reasonable on the other hand to measure MAIE against expected target pitches in equal temperament as others have shown that in solo-singing, where melodic intonation is most appropriate, singers tend toward equal tempered tuning [151, 145] Also, the differences between equal tempered intervals and those of a justly tuned scale differ in by less than 16 cents which is smaller than the average interval precision scores which range from 19 cents to over 40 cents in this study.

Results

Average values for the intonation metrics calculated are presented in Table 6.6, full results for MAIE, MAPP and MAIP for all fragments in each test are found in Appendix K

Test	**MAIE**	**MAPP**	**MAIP**
232a	44.7	23.4	32.5
232b	57.7	29.7	45.9
221	24.2	17.0	19.0
212	35.4	28.5	37.1
Mean (StDev)	**40.5 (12.3)**	**24.6 (5)**	**33.6 (9.7)**

Table 6.6: *Mean (St Dev) values of Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) for each test (cents)*

Overall there was a high degree of variability between the singers found in the measures of intonation.

Discussion

The tenor singer in this study (test 221) was the most precise and accurate, whereas the bass singer (test 232a and 232b) was the most inaccurate and imprecise. The highest values of average MAIE were produced by the bass singer (tests 232a and 232b), who did in fact verbally report difficulties in maintaining stable intonation both in the real and virtual performance spaces.

MAIE Average MAIE across the productions of the fragments within one test ranged from 24.2 cents (Test 221) to 57.7 cents (Test 232b). These values are smaller than those found by Berkowska [235] in a study of occasional singers but are in line with results found by Mauch et al. [251] (mean MAIE of 26 cents) and Dalla Bella [234] (mean MAIE 20 - 40 cents) whose studies involved more experienced singers.

MAPP Average MAPP across the productions of the fragments within one test ranged from 17 cents (test 221) to 29.7 cents (test 232b). These findings are similar to pitch precision values ranging from 3 to 24 cents (mean 13 cents) found amongst choral singers by TernstrÖm and Sundberg 1982 [245]. An a study of occasional singers Berkowska and Dalla Bella found mean pitch precision values of 41.2 cents. As expected the professional singers in this study are more precise in pitching than occasional singers.

MAIP Average MAIP scores across the productions of the fragments within a test ranged from 19 cents (Test 221) to 45.9 cents (Test 232b) which are in line with those values found amongst singers by Berkowska and Dalla Bella [235] of average interval precision socres of 49.4 cents.

Singers in this test are more precise than accurate, showing on the whole lower values of MAIP and MAPP rather than MAIE, which might suggest that they are not in fact tending towards equal temperament as a tuning system even in solo singing. This is an area which should be investigated further in future research.

6.4.4 Analysis of Vibrato

Method

The framewise fundamental frequency estimates for each note were obtained in AMPACT as described in section 5.5.1. These fundamental frequency traces are considered as a signal and, following the method used by Gleiser, Friberg and Granqvist [309], bandpass filtered in MATLAB using a 2nd order Butterworth filter with a lower cut-off frequency of 4 Hz and an upper cut-off frequency of 15 Hz.

Vibrato rate and extent can then be evaluated from the frequency domain representation of the filtered fundamental frequency traces (vibrato contours). The amplitude of the strongest component is equivalent to the vibrato extent and the frequency of the vibrato trace corresponds to the vibrato rate.

Filtering the $F0$ trace in this way also serves to eliminate onset pitch glides, and any offset/onset transients which may have been erroneously included by the vowel-portion by the state selection implemented in AMPACT (see section 5.5.1) as is illustrated in Figures 6.40 and 6.41

In evaluating the vibrato rate and extent of sung tones, most authors do not include notes which do not have sufficient duration to allow vibrato to be established. For example Prame [243] includes only notes which were long enough to provide reliable data, choosing these by hand from the musical score, which also excludes those notes where vibrato extent might be mis-calculated due to the inclusion of an onset pitch glide. In the present

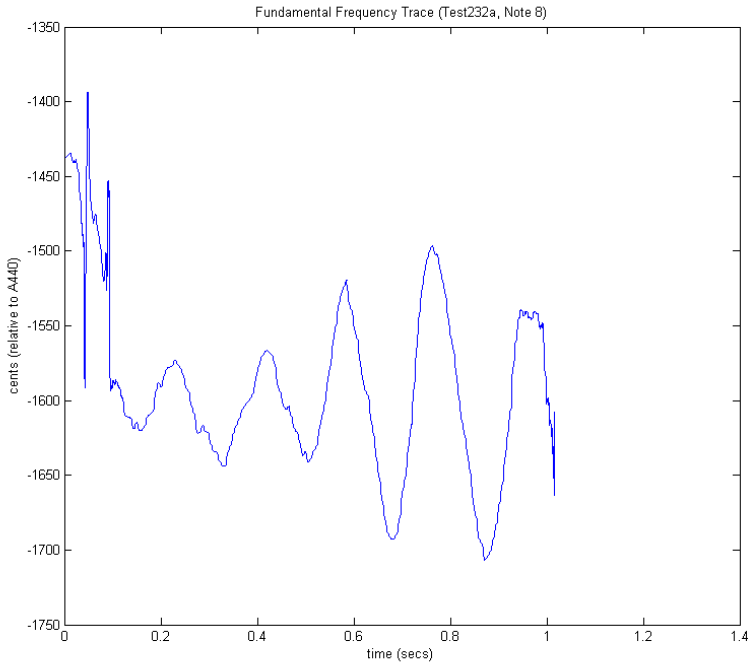


Figure 6.40: *Unfiltered fundamental frequency trace of values estimated by the YIN algorithm implemented in AMPACT, for note number 8 of MRVirtV1 fragment in Test 232a*

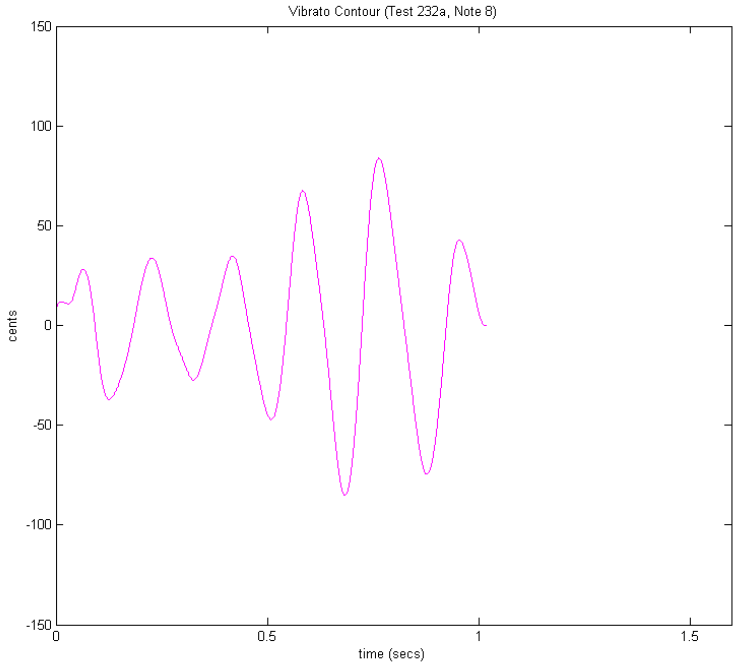


Figure 6.41: *Filtered fundamental frequency trace to provide vibrato contour of note number 8 of MRVirtV1 fragment in test 232a*

analysis only tones with 4 complete cycles or more were included in the mean vibrato rate and extent calculations for each production.

Results

In order to gain an impression of the vibrato characteristics of the different productions of each fragment in the tests figures 6.42 to 6.53 show the mean vibrato rate and mean vibrato extent, (with standard deviation) which give an indication of the amount of variation within the fragment.

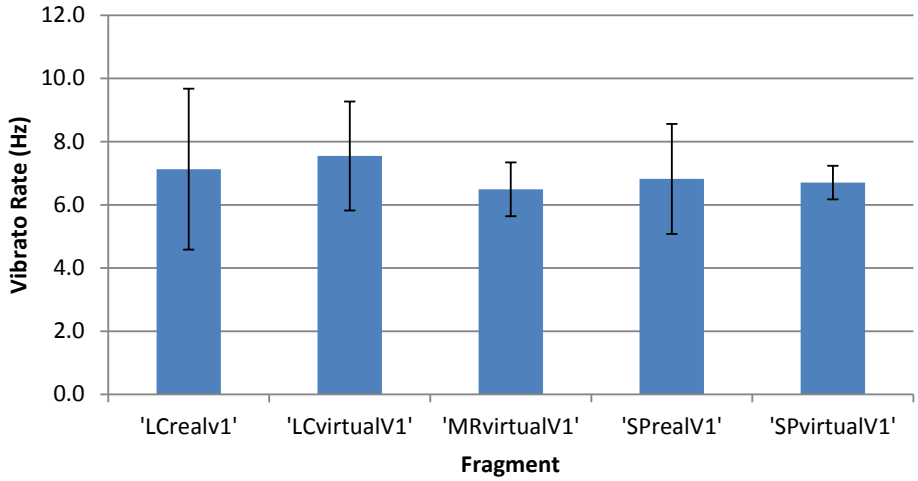


Figure 6.42: *Mean vibrato rate and standard deviation for each fragment in test 232a*

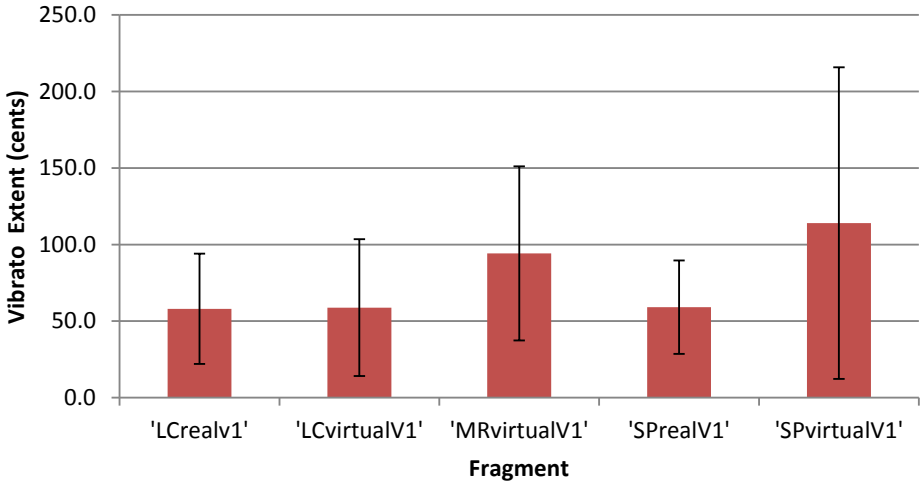


Figure 6.43: *Mean vibrato extent and standard deviation for each fragment in test 232a*

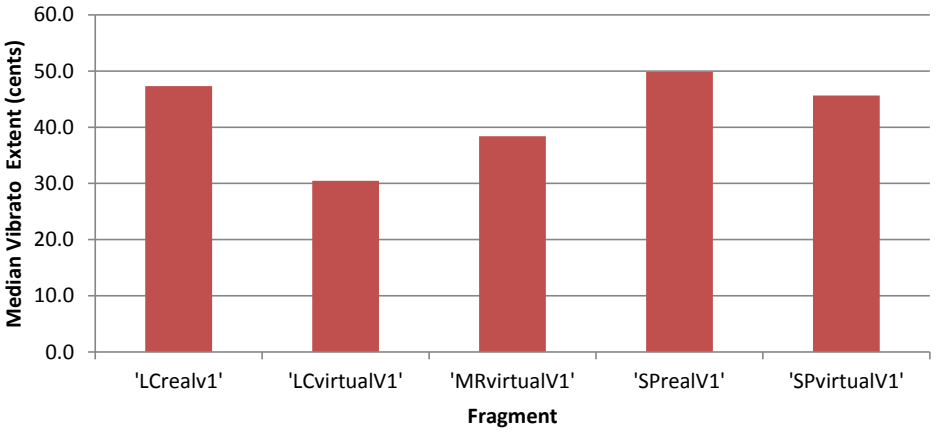


Figure 6.44: *Median vibrato extent for each fragment in test 232a*

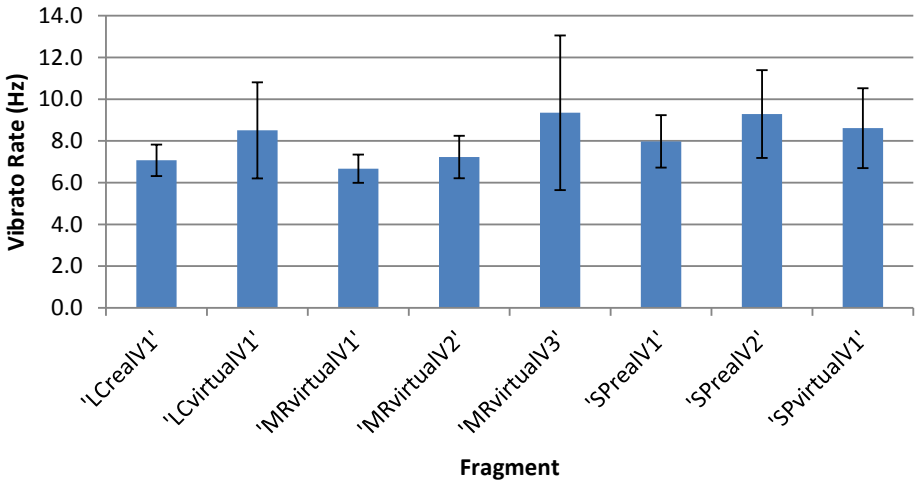


Figure 6.45: *Mean vibrato rate and standard deviation for each fragment in test 232b*

Table 6.7 shows the mean (and standard deviation) vibrato rate and extent across the fragments in each test, the median vibrato extent, as well as minimum and maximum values of the vibrato parameters.

	Vibrato Rate (Hz)			***Vibrato Extent (cents)***			
Test	Mean (SD)	Max	Min	Mean (SD)	Median	Max	Min
232a	6.9 (0.4)	7.5	6.5	76.8 (25.9)	42.3	114.0	58.0
232b	8.1 (1.0)	9.3	6.7	79.1 (35.8)	41.3	134.9	39.9
221.0	7.1 (1.1)	8.7	5.7	88.3 (41.7)	40.6	168.0	35.9
212.0	7.6 (0.8)	10.36	6.41	75.7 (22.6)	35.3	122.8	36.9

Table 6.7: *Average Mean, standard deviation, median, maximum and minimum values of vibrato rate (Hz) and vibrato extent (cents) across all fragments in each test*

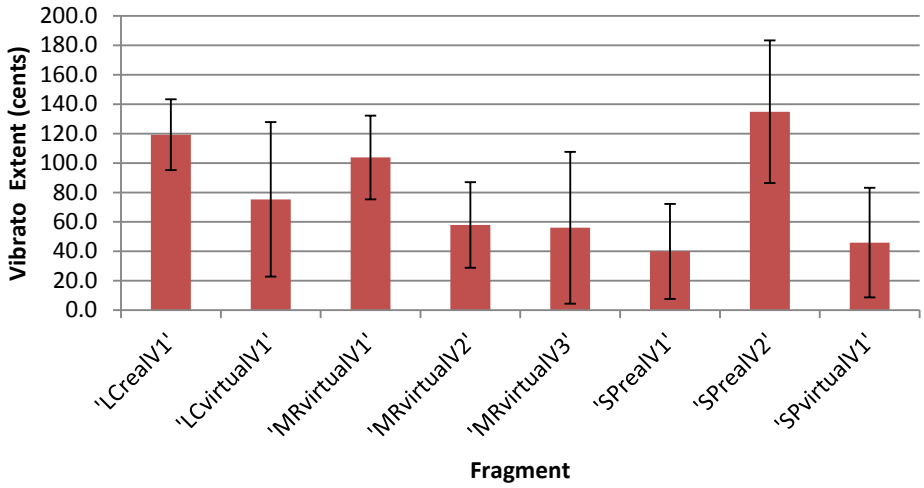


Figure 6.46: *Mean vibrato extent and standard deviation for each fragment in test 232b*

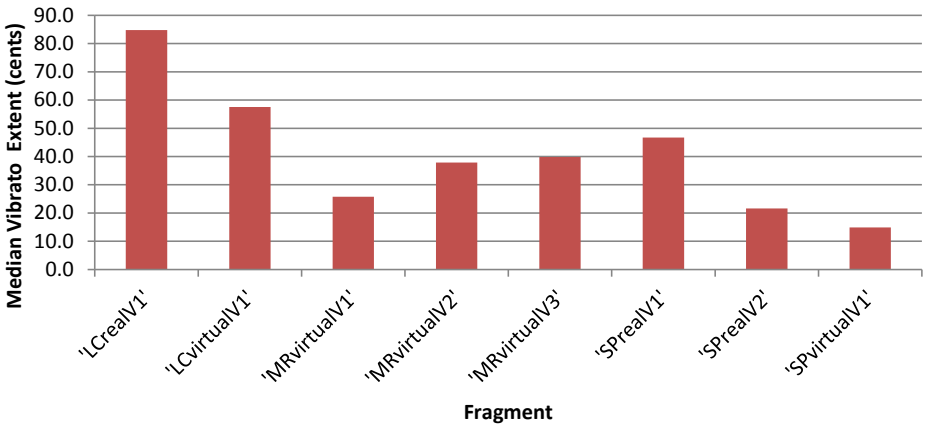


Figure 6.47: *Median vibrato extent for each fragment in test 232b*

Discussion

Singers in this study generally exhibit expected mean vibrato rates of 6 - 8 Hz with small standard deviations meaning that singers vibrato rate is generally quite consistent across the different productions. Mean vibrato extent, on the other hand, varies more widely between the fragments within each test, ranging from 35 - 168 cents (standard deviations between 22.6 - 41.7 cents). However, mean vibrato extent values between singers are fairly similar lying between 75 - 88 cents.

Standard deviations of vibrato extent for each performance, which can be taken as a measure of the variability within a performance, are larger on the whole than those found by Timmers [166] in a study of vocal expression in Schubert songs. Median vibrato extent values are more consistent across the fragments of each test as can be seen in Figures 6.44, 6.47, 6.50 and 6.53. Average median vibrato extent ranges from 35.3 cents to 42.3 cents

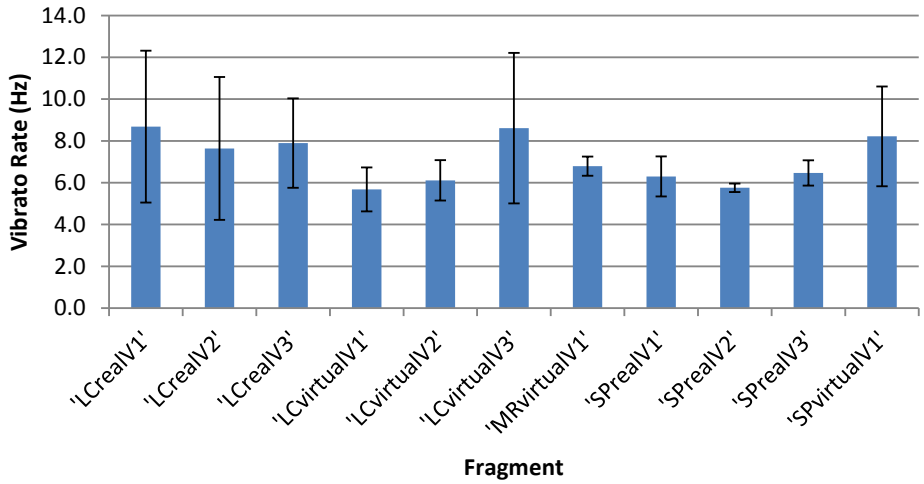


Figure 6.48: *Mean vibrato rate and standard deviation for each fragment in test 221*

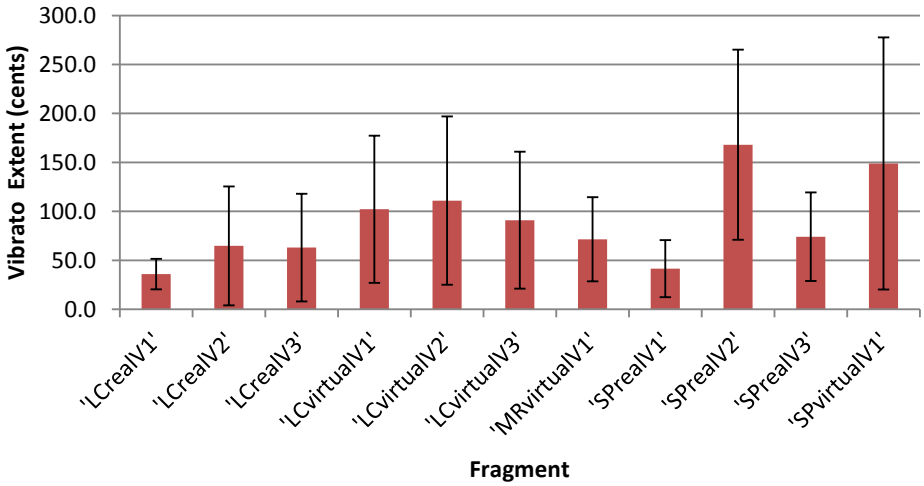


Figure 6.49: *Mean vibrato extent and standard deviation for each fragment in test 221*

across the singers (Table 6.7).

The large variability in mean vibrato extent may be explained by the fact that singers in this study were all early music specialists. In the historically informed singing style prevalent amongst early music specialists, fundamental frequency vibrato is used as an effect, or musical ornament, to decorate or add shape and colour to particular notes, rather than as a constant characteristic of the vocal sound. Other studies of vibrato in singers have involved opera singers, where vibrato is almost invariably present in each note and is an important part of the operatic technique.

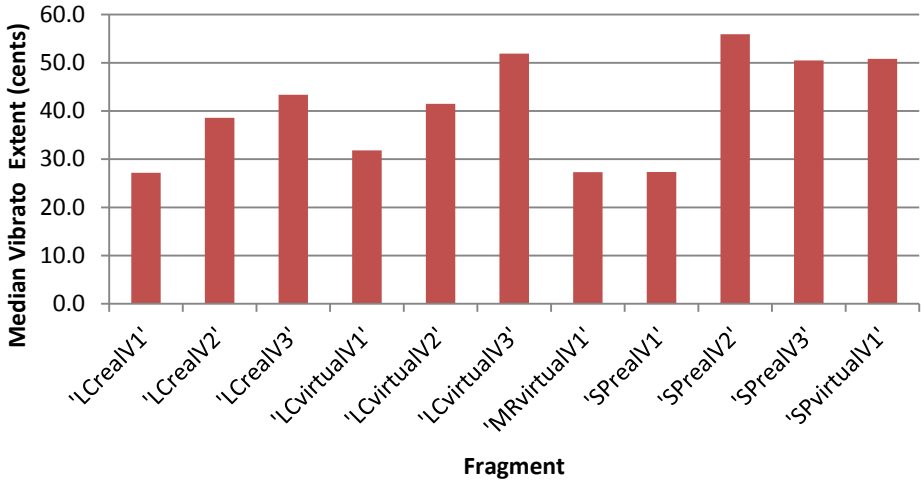


Figure 6.50: *Median vibrato extent for each fragment in test 221*

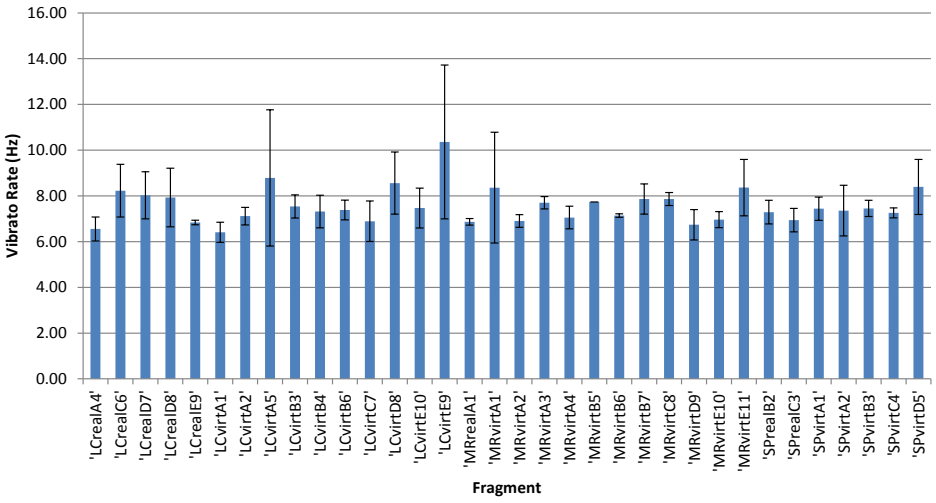


Figure 6.51: *Mean vibrato rate and standard deviation for each fragment in test 212*

6.5 Correlation of Performance Attributes and Perceptual Evaluation

6.5.1 Method

Multiple linear regression can be used infer the relationships between the position of the stimuli (musical fragments) in the perceptual space, as visualised via the MDS analysis, and the quantitative properties of the stimuli; this method is sometimes referred to as "vector property fitting" (PROFIT) [310].

Multiple linear regression was performed in the present project using the MDS solution positions for each listening test as the independent variables, and the measured performance

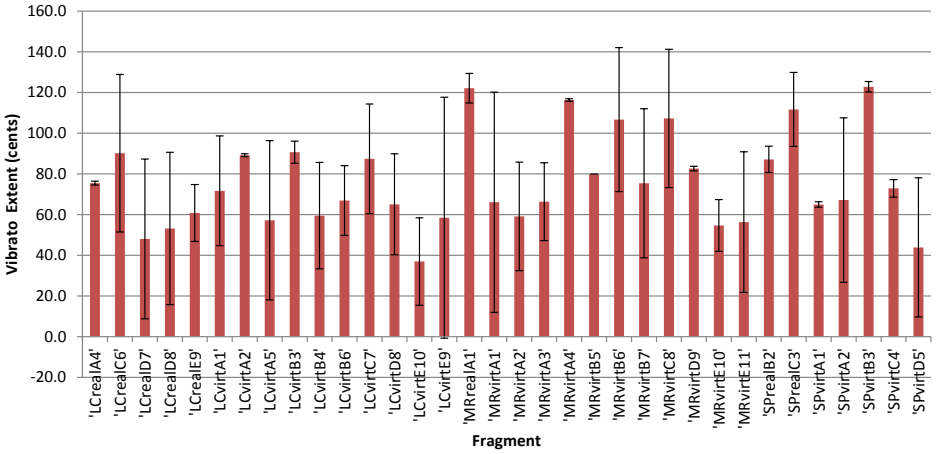


Figure 6.52: *Mean vibrato extent and standard deviation for each fragment in test 212*

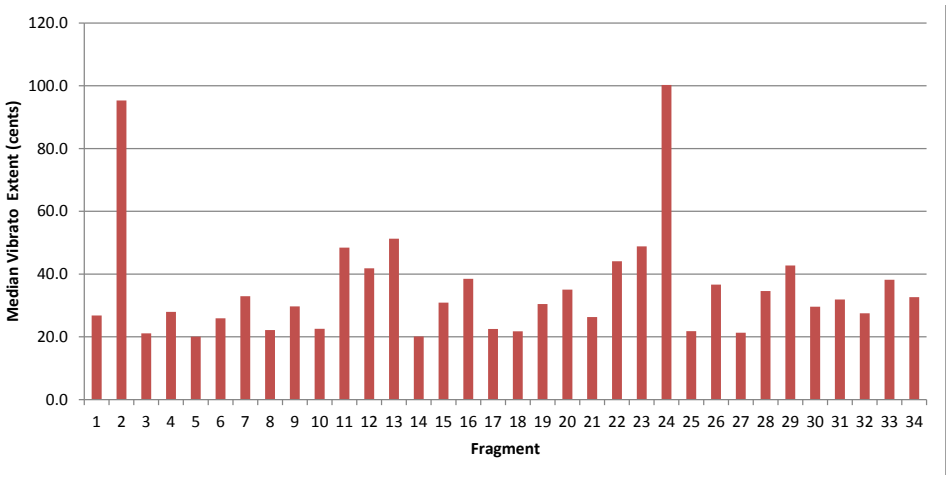


Figure 6.53: *Median vibrato extent for each fragment in test 212*

attributes as the dependent variables. Residual case order plots displaying the confidence intervals of the residuals from the regression analysis were produced in order to identify any possible outliers in the data. Outliers were removed and the regression analysis repeated with the new data set.

Following the procedure as outlined by Bonebright [17, p.136], where the ratio of beta weights of the two predictor variables define the slope of the "best fit vector", a line is drawn passing through the origin with the arrowhead indicating the direction of increasing property value. Linear regression was performed for the six performance attributes, as reported in section 6.4 for each test. Only property vectors with R^2 values greater than 0.40 are fitted to the MDS solutions here, with the length of each property arrow indicating the "goodness of fit" of the linear regression model.

6.5.2 Results

Figures 6.54 to 6.58 below illustrate the vector fitting of objective performance attribute data to the MDS solutions presented in section 6.3. Only performance attributes which account for a reasonable proportion of variance in the regression model, with R^2 values greater than 0.4, and with confidence ratings of 10 % or less (p value < 0.1) are plotted here.

6.5.3 Test 232a Bass

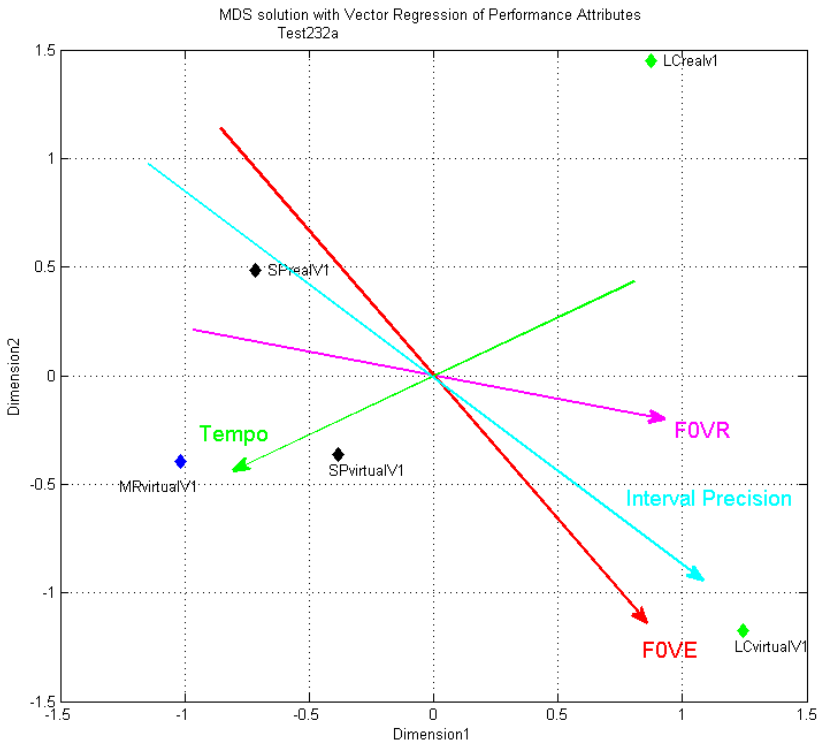


Figure 6.54: *Vector property fitting of performance attributes onto MDS perceptual map for Test 232a*

Discussion 232a

The two vibrato measures (extent and rate) and Interval Precision (MAIP) account for a very large proportion of the variance in the similarity data and tempo also accounts for 70 % of the variance in the MDS solution in this test.

The vector property fitting does seem to confirm listeners' comments that *LCrealV1* and *LCrealV2* differ from the other fragments in this test in terms of vibrato. Listeners also comment on the power and drama of these performances compared with the fragments

Performance Attribute	R^2
Vibrato Rate (F0VR)	**0.99***
Vibrato Extent (F0VE)	**0.99***
Global Tempo (GT)	**0.70***
Pitch Precision (MAPP)	0.33
Interval Precision (MAIP)	**0.99**
Interval Error (MAIE)	0.41

Table 6.8: *Regression analysis between performance attributes and positions in MDS perceptual map for Test 232a (boldface p<0.1, * p<.05)*

from MR and SP, which may relate to the tempo of the fragments, with the LC fragments enjoying a slower delivery than the others.

It seems that intonation and vibrato explain the distinction between the *virtual* and *real* fragments which reflects the singer's own report that he found maintaining good intonation in the virtual space more difficult than in the real.

6.5.4 Test 232b Bass

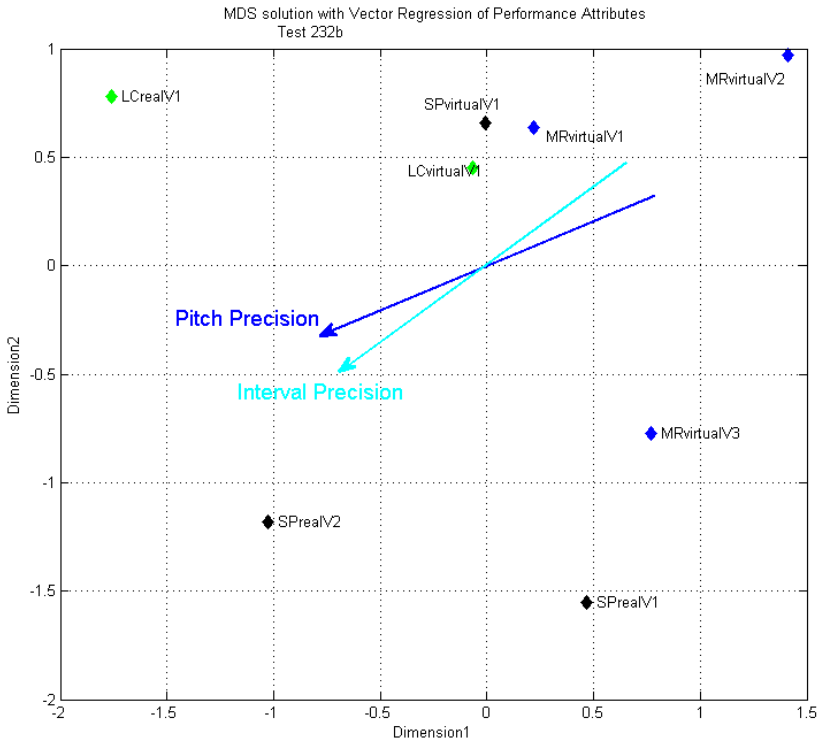


Figure 6.55: *Vector property fitting of performance parameters onto MDS perceptual map for Test 232b*

Performance Attribute	$\boldsymbol{R^2}$
Vibrato Rate (F0VR)	0.34
Vibrato Extent (F0VE)	0.56
Global Tempo (GT)	0.16
Pitch Precision (MAPP)	**0.72***
Interval Precision (MAIP)	**0.68***
Interval Error (MAIE)	0.57

Table 6.9: *Regression analysis between performance attributes and positions in MDS perceptual map for Test 232b (boldface p<0.1, * p<.05)*

Discussion

Although *LCRealV1* appears to be an outlier there were in fact no outliers identified in the regression analyses. The intonation measures explain most of the variance in the MDS dissimilarity data plotted. Again this reflects the singer's own feeling that he struggled to maintain good intonation in the SP acoustic configuration for the piece in this test. However, it seems in this test that intonation is less stable in the *real performance space* in contrast to the intonation problems reported in the *virtual space* for test 232a.

It was suggested in Section 6.7.1 that a difference in the spectral locus of reflected sound in a venue may alter a singers' perception of the pitch being produced leading to unstable intonation. This was commented upon by the bass singer who experienced this phenomenon in both the real and virtual spaces to a greater or lesser extent. The tessitura (pitch range) of the song in this test is low, reaching down to D2. The acoustic characteristics of the real space (see Section 3.5.2) would mean that for these lower notes the upper harmonics fall in the more reverberant mid-frequency range, perhaps leading to a perceived rise in pitch due to a shift in spectral locus. This effect and its subsequent perceptual implications for singers is ripe for further investigation in a future study.

The differences in tempo as reported by listeners between the group of fragments in the top right-hand corner of the MDS plot (*MRvirtualV1*, *LCvirtualV1* and *SPvirutalV1*) are not reflected in the regression analysis here as global tempo, although it accounts for 16% of the variance in the data, it cannot be used to significantly predict these differences (p value over 0.1). There may be other distinguishing features amongst these fragments picked up by listeners, but not captured by the performance parameter analysis such as pronunciation of words, length of notes and timbre.

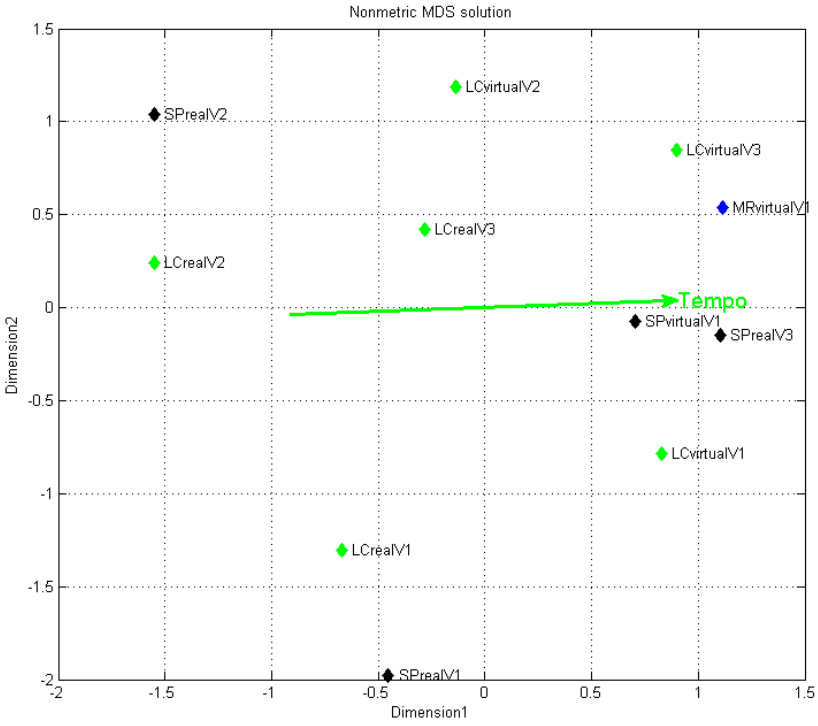


Figure 6.56: *Vector property fitting of performance parameters onto MDS perceptual map for Test 221*

Performance Attribute	R^2
Vibrato Rate (F0VR)	0.02
Vibrato Extent (F0VE)	0.36
Global Tempo (GT)	**0.82*****
Pitch Precision (MAPP)	0.45
Interval Precision (MAIP)	0.25
Interval Error (MAIE)	0.35

Table 6.10: *Regression analysis between performance attributes and positions in MDS perceptual map for Test 221 (boldface p<0.1, * p<.05, ** p<.01, ***p<.001)*

6.5.5 Test 221 Tenor

Discussion

SPrealV2 is an outlier in the regression analysis of vibrato extent, and indeed listeners comment on this fragment's excessive and prolonged vibrato, which is out of keeping with the other fragments.

The two different types of performance here - the slow measured performance and the faster rhythmical performance - are accounted for by the tempo values found in

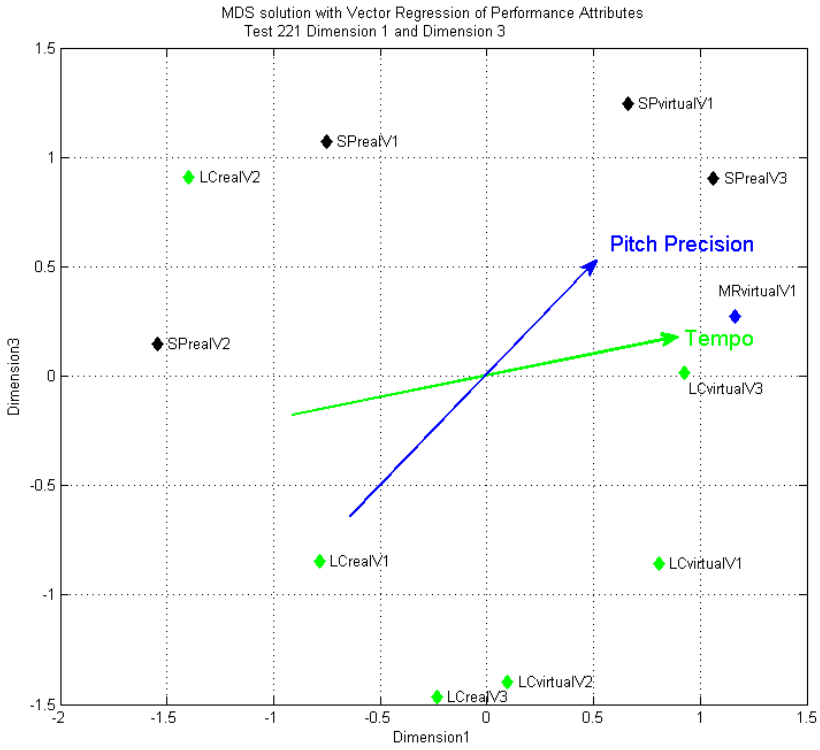


Figure 6.57: *Vector property fitting of performance parameters onto MDS perceptual map for Test 221 Dim 1 vs Dim 3*

Performance Attribute	R^2
Vibrato Rate (F0VR)	0.01
Vibrato Extent (F0VE)	0.29
Global Tempo (GT)	**0.47**
Pitch Precision (MAPP)	**0.70****
Interval Precision (MAIP)	0.39
Interval Error (MAIE)	0.43

Table 6.11: *Regression analysis between performance attributes and positions in 3-D MDS solution for Test 221: dimensions 1 and 3 (boldface p<0.1, * p<.05, ** p<.01)*

the acoustic analysis of the fragments. Tempo increases from left to right of the MDS plot in Figure 6.56. No other perfromance parameter is significantly correlated with the differences between fragments in dimensions 1 and 2.

Nevertheless when performance attributes are regressed onto the point configuration taken from the 3-dimensional MDS solution, (Figure 6.57) pitch precision (MAPP) accounts for 70% of the positioning of fragments in this map. Global tempo is also reasonable predictor of the fragment locations in this MDS solution. In general, fragments taken from the *virtual space* lie to the right of these two vectors stemming from their quicker

tempo and less stable intonation.

6.5.6 Test 212 Mezzo-Soprano

Performance Attribute	$\boldsymbol{R^2}$
Vibrato Rate (F0VR)	0.01
Vibrato Extent (F0VE)	**0.51*****
Global Tempo (GT)	**0.80*****
Pitch Precision (MAPP)	0.02
Interval Precision (MAIP)	0.05
Interval Error (MAIE)	**0.14**

Table 6.12: *Regression analysis between performance attributes and positions in MDS perceptual map for Test 212: dimensions 1 and 2 (boldface p<0.1, * p<.05, ** p<.01, ***p<.001)*

Performance Attribute	$\boldsymbol{R^2}$
Vibrato Rate (F0VR)	0.02
Vibrato Extent (F0VE)	**0.52*****
Global Tempo (GT)	**0.29****
Pitch Precision (MAPP)	0.02
Interval Precision (MAIP)	0.05
Interval Error (MAIE)	0.09

Table 6.13: *Regression analysis between performance attributes and positions in MDS perceptual map for Test 212: dimensions 1 and 3 (boldface p<0.1, * p<.05, ** p<.01, ***p<.001)*

Discussion

Tempo and vibrato extent are the most effective at predicting the dissimilarities between the performed fragments in plots of 2d and 3d MDS solutions (Figures 6.58 and 6.59) The additional variation between fragments expressed along dimension 3 is not well accounted for by any of the performance attributes analysed here.

Two fragments (*LCrealC6* and *MRvirtC8*) in the top left-hand corner of the plot were indeed outliers for some of the regression analyses. *LCRealC6* is an outlier in terms of the intonation whereas *MRvirtC8* is an outlier in the regression of tempo data. The group of fragments taken from verse E are all at the right-hand side of the tempo vector, reflecting their slow tempo.

Most of the fragments taken from verse D appear to the right-hand side of the Tempo, Fundamental Frequency Vibrato Extent (F0VE) vectors. The majority of fragments from

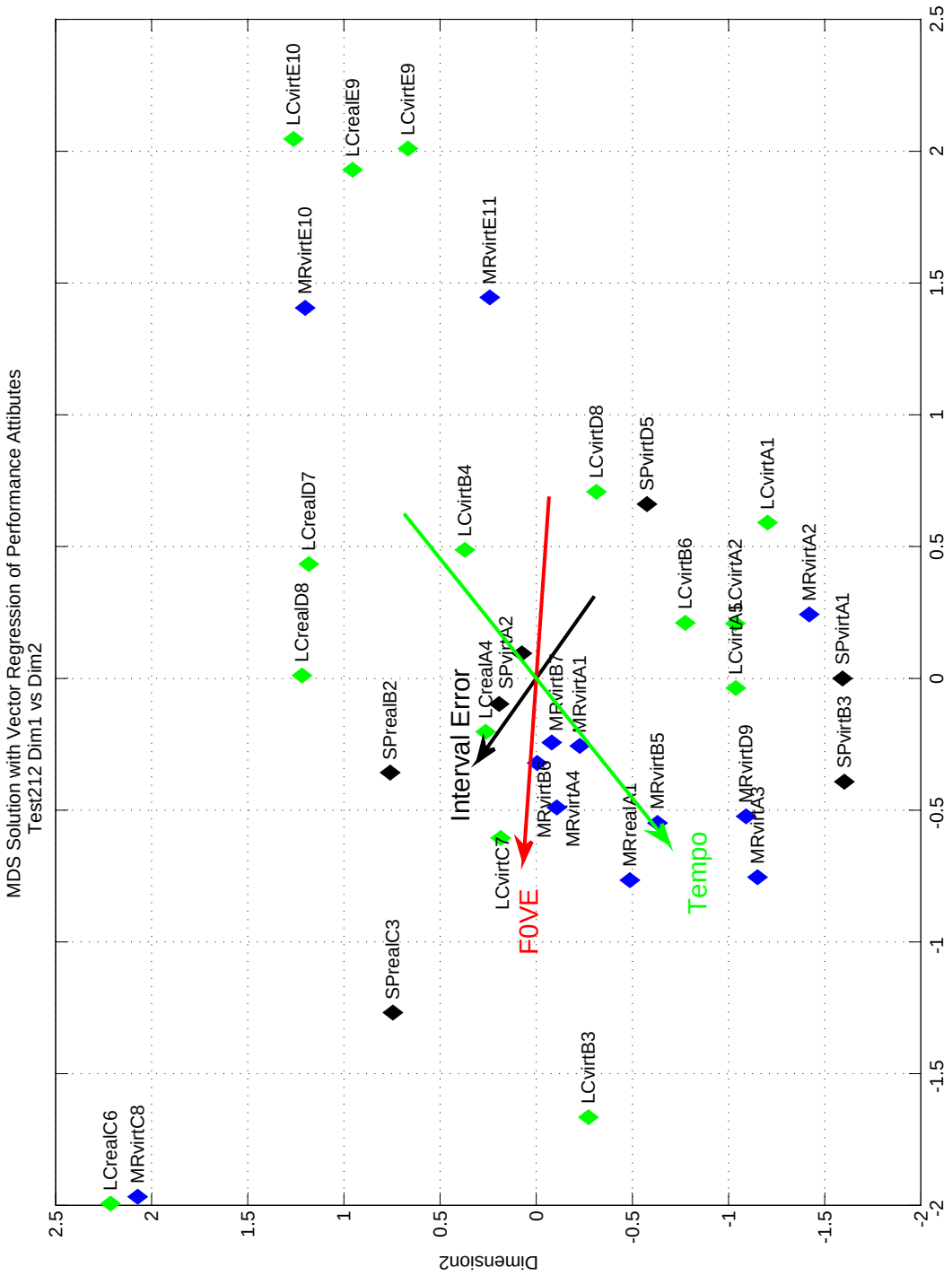


Figure 6.58: *Vector property fitting of performance parameters onto MDS perceptual map for Test 212 dimension 1 and dimension 2*

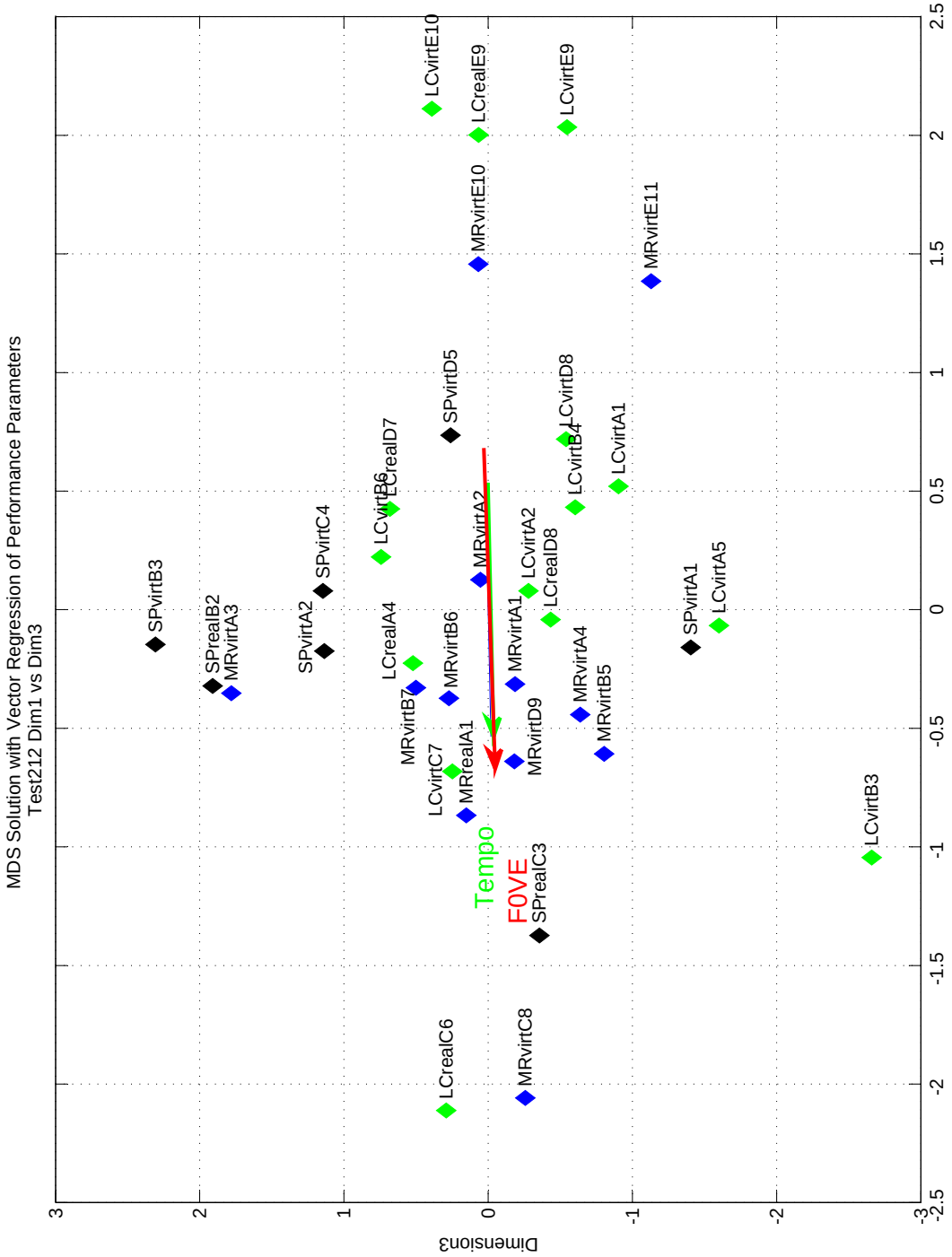


Figure 6.59: *Vector property fitting of performance parameters onto MDS perceptual map for Test 212 Dim 1 vs Dim 3*

the real performance space appear to be less stable in terms of pitching (higher pitch precision values).

Simple linear correlation analyses were carried out to assess the relationship between performance parameters, room acoustic configurations, song verses and real/virtual performance space for fragments in this test.

Performance Parameter	***Acoustic Configuration (LC/MR/SP)***	***Space (Virtual/ Real)***	***Verse***
FOVE	**0.28**	0.18	**-0.55*** **
FOVR	-0.13	-0.1	0.32
GT	0.2	0.19	**-0.76**** **
MAIE	-0.02	**0.42****	-0.15
MAIP	0.04	-0.19	0.27
MAPP	-0.02	-0.07	0.33

Table 6.14: *Results of linear correlation between performance parameters and acoustic configuration, performance space, and song verse (boldface p<0.1, * p<.05, ** p<.01, ***p<.001, ****p<.0001))*

For the correlation analyses presented in Table 6.14 "acoustic configuration" was coded as LC = 2, MR = 3 and SP = 4, verses were coded as verse A, B & C = 1, D= 2 and E=3 and "performance space" was coded as Real = 1, Virtual = 0.

There is a small correlation between vibrato extent F0VE and acoustic configuration, with vibrato extent (F0VE) increasing as reverberation time decreases, but this is probably on the limit of significance with a p-value just less than 0.1. More significant is the correlation between interval accuracy Mean Absolute Interval Error (MAIE) and performance space, with a positive correlation in the real performance space. The most significant correlations however are the negative correlations between the song verse and vibrato extent and global tempo - global tempo is highly correlated with verse, with the later verses exhibiting slower tempo and smaller vibrato extent, probably stemming from the performer's desire to express the emotional content of the latter verses.

6.6 Testing effect of Acoustic Setting and Simulation

In order to investigate any effect of acoustic setting or simulation on the measured musical performance attributes, and the listeners' similarity ratings, a data set of forty fragments was established, which included fragments from across the four listening tests. The data set was chosen to include similar numbers of fragments recorded in the real and virtual spaces (19 from the real and 21 from the virtual simulation) and more balanced numbers

of fragments from the three acoustic settings (13 fragments from the SP setting, 13 from the MR setting, 14 from the LC setting).

6.6.1 ANOVA on performance attributes

A series of two way ANOVAs were performed, with *acoustic* (LC/MR/SP) and *simulation* (real vs virtual) as the independent variables and each of the performance attributes as the dependent variable. Standardised measures of global tempo were used in order to allow comparison across fragments from the different tests, which were taken from different songs. Results are presented in Table 6.15.

Dependent Variable	**Source**	**Type III sum of squares**	***df***	**Mean square**	***F***	**Sig.**
Mean Vibrato Extent	Acoustic	3.0487	2	1.52435	1.5	0.2376
	Simulation	2.7571	1	2.75714	2.71	0.1088
	Acoustic*Simulation	1.7838	2	0.8919	0.88	0.425
Median Vibrato Extent	Acoustic	1209.2	2	604.615	2.78	0.076
	Simulation	988.6	1	988.63	4.55	* ***0.0402***
	Acoustic*Simulation	18.5	2	9.226	0.04	0.9585
Vibrato Rate	Acoustic	0.0643	2	0.03216	0.03	0.9684
	Simulation	0.0043	1	0.0043	0	0.9482
	Acoustic*Simulation	4.4563	2	2.22817	2.22	0.1236
Global Tempo	Acoustic	5.3696	2	2.68479	2.78	0.0762
	Simulation	0.224	1	0.22405	0.23	0.6332
	Acoustic*Simulation	1.3878	2	0.69391	0.72	0.4948
MAIE	Acoustic	65.67	2	32.837	0.3	0.7426
	Simulation	102.92	1	102.916	0.94	0.3389
	Acoustic*Simulation	257.8	2	128.9	1.18	0.3201
MAIP	Acoustic	1568.2	2	784.1	1.02	0.3725
	Simulation	1463.8	1	1463.84	1.9	0.1773
	Acoustic*Simulation	592.5	2	296.27	0.38	0.6839
MAPP	Acoustic	228.62	2	114.309	0.6	0.5536
	Simulation	127.57	1	127.572	0.67	0.4183
	Acoustic*Simulation	394.84	2	197.422	1.04	0.3647

Table 6.15: *Results of ANOVA on effect of acoustic (LC/MR/SP) and simulation (virtual vs real) on performance parameters (boldface p<0.1, * p<.05)*

There was only one significant effect of *simulation* on median vibrato extent values ($F_{(1,39)}$=4.55, p <0.05). No other significant effects of either *acoustic* or *simulation* were found on the performance attributes

6.6.2 ANOVA on similarity ratings

The same data set as described above was used to test the effect of *acoustic* and *simulation* on the similarity ratings of pairs of fragments. Each pair of fragments was coded according to whether they were from a congruent pair of *acoustic* or *simulation* or not e.g. two fragments both recorded in the virtual simulation were coded as "congruent" whereas a pair of fragments including one from the *real performance space* and one from the *virtual* simulation were coded as "non-congruent".

A two-way ANOVA was performed with *acoustic* (congruent vs non-congruent) and *simulation* (congruent vs non-congruent) as independent variables, and similarity ratings as the dependent variable for each separate listening test, and for the full data set described above.

Test	**Source**	**Type III sum of squares**	***df***	**Mean sqaure**	**F**	**Sig.**
232a	Acoustic	0.04401	1	0.04401	2.98	0.0858
	Simulation	0.00306	1	0.00306	0.21	0.6493
232b	Acoustic	0.00078	1	0.00078	0.07	0.7875
	Simulation	0.03662	1	0.03662	3.41	0.0654
221	Acoustic	0.0044	1	0.00443	0.16	0.6854
	Simulation	0.0208	1	0.02078	0.77	0.3801
212	Acoustic	0.007	1	0.00713	0.25	0.6198
	Simulation	0.365	1	0.36509	12.6	**0.0004 *****
Full set	Acoustic	0.002	1	0.00247	0.08	0.7742
	Simulation	0.127	1	0.12693	4.23	**0.0397***

Table 6.16: *Results of ANOVA on effect of acoustic (congruent vs non-congruent) and simulation (congruent vs non-congruent) on similarity ratings of pairs of fragments (boldface p<0.1, * p<.05, ** p<.01, ***p<.001)*

There was a significant effect of *simulation* on the similarity ratings ($F_{(1,39)}$=4.23 $p <$ 0.05). Since the similarity rating data was found to be highly skewed (skewness value of 0.95) a Wilcoxon rank-sum test was also performed. This is a non-parametric test which has fewer assumptions about normality of distribution of the data. The result of this test also showed a significant effect of *simulation* on the similarity ratings. ($p < 0.001$)

6.6.3 Discussion

It is surprising that the acoustic setting did not have a significant effect on more of the performance attributes measured, as singers reported changes in their own performances

between the different acoustic settings. Although *simulation* (real vs virtual) did not have a significant effect on most of the performance attributes, it did have a significant effect on median vibrato extent ($F_{(1,39)}$=4.55, p <0.05).

Simulation (congruent vs non-congruent) also had a significant effect on the similarity ratings between pairs of fragments, which means that listeners did indeed rate pairs of fragments from matching simulations (e.g. real/real or virtual/virtual) differently to those from non-matching simulations (e.g. real/virtual). This result suggests that the *real* and *virtual spaces* might not match closely enough to result in singing performances in the two spaces being rated as similar overall.

Nevertheless, both of these results should be viewed in the context of the data set used, which contained a greater number of fragments by the soprano singer (test 212) in the *virtual space*, than of the other singers in the *virtual space*, which might have an undue influence on the results. Future work will include establishing a more systematic set of recordings from which it will be easier to draw statistically significant results. This will be undertaken once improvements, which are outlined in 7.1.1 to the VSS have been completed.

6.7 Summary

This section has reiterated the MDS perceptual maps presented in Section 6.3 and has used the technique of vector property fitting to attempt to infer the objective performance attributes which explain the dimensions of the perceptual space.

Global Tempo

The majority of singers who sang in the VSS advised that one of the main alterations of their performance according to the room acoustic conditions was in tempo and timing.

It was shown in Sections 6.5.3, 6.5.5 and 6.5.6 that global tempo is generally highly correlated with the perceived differences between the sung fragments, with R^2 values above 0.70 for three out of four of the listening tests.

Vibrato

Although none of the singers here mentioned vibrato as an aspect of singing which would change in different acoustics, vibrato rate and extent were analysed since other authors in similar studies had found alterations in this attribute.

In fact, vibrato rate was only correlated with the perceived differences for the bass singer in Test 232a alone and not for the other singers. It was noted previously that

average vibrato rates did not vary greatly between the fragments in each test, as the professional singers in this research maintained steady consistency of vibrato rate across performances.

The effect of *simulation* on median vibrato was shown to be significant in Test 212 and when compared across all tests. It was also shown to be a good indicator of the perceived differences between fragments for both the bass in Test 232a and the mezzo-soprano in Test 212. For the mezzo-soprano F0VE was also significantly correlated with both the acoustic configuration of the space and the verse from which the fragment was taken. There was moderate negative correlation for this singer between vibrato extent F0VE and reverberation time, (see section 6.5.4), which reflects with a finding by Ueno et al. [3] that F0VE increased as reverberation times in the performance venues studied decreased.

Intonation

Four of the seven singers in the study thought that their intonation or tuning was affected by the different room acoustic conditions. Intonation measures (Mean Absolute Pitch Precision (MAPP), Mean Absolute Interval Precision (MAIP) and MAIE) were indeed strongly correlated with performance differences for the bass singer in test 232b, and MAIP (interval precision) was very highly correlated for the same singer in test 232a. In contrast, intonation measures were only moderately correlated for the other singers. MAPP was correlated with perceived differences for the tenor (Test 221) and seemed to relate to dimension 3 of the perceptual map.

Singers in this study on the whole were more precise than accurate. Whereas [233] found that imprecise singing was widespread amongst the general population, precise singing is to be expected of professional singers. One might also expect professional singers to be highly accurate but the higher interval error values for these singers might in fact hint that they are not singing equally tempered intervals (against which errors here are calculated) but could be using a temperament closer to just tuning (as was discussed in Section 4.3.11).

6.7.1 Acoustic Configurations

For most of the tests, groupings of similar fragments do seem to reflect the acoustic surroundings in which they were made. For example, in Test 232b there is a similarity between all the MR fragments along dimension 1 of the perceptual map, and in Test 212 within those fragments taken from Verse E, there is a clear grouping of LC fragments from MR fragments along dimension 1. In addition, for this test, it was shown that Vibrato Extent was moderately correlated with acoustic configuration, with vibrato extent

increasing with shorter reverberation times.

In Test 232a Global Tempo is shown to explain the differences between fragments recorded in the SP/MR configurations and those in the more reverberant (LC) condition. This echoes findings by Kato et al. [14] who found slower tempos in more reverberant concert halls and Ueno et al. [7] who measured faster tempos for less reverberant acoustics.

In Test 221 (tenor) pitch precision and tempo vector both point towards most of the SP and MR versions and away from the LC versions. Tempo and intonation have been shown to be correlated in other studies, such as that by Jers and Ternström [311], who found that deviation in mean fundamental frequency exhibited less scatter with slower tempos, suggesting that accurate pitching is easier for the singer at slower tempos.

Control of sung pitch is influenced by the auditory feedback the singer receives [283, 288, 162] and which can of course be altered by the room acoustic characteristics of the performance space. Two main factors at play in this feedback loop have been identified which relate to room acoustics.

Firstly reverberation time, which means that pitches of previous notes "linger" in the air, has been shown to affect the singer's ability to tune subsequent notes, and can have either a detrimental or beneficial effect depending upon the performer and the style of music being performed.

Secondly, any spectral colouration of aural feedback arising from room acoustic conditions in the space can lead to perception of the singer's own voice being "coloured" by the reflected sound. Ternström and Sundberg [250] showed that the presence of higher frequency partials in a complex tone increased the ease of tuning an interval for a singer. It is likely that the frequency response of a performance venue, and its boosting or dampening effect on the balance of partials across the spectrum could also affect the ease of intonation for the singer. Indeed the bass singer in this study complained that intonation was hard to maintain in the LC setting (most reverberant) of both the real and virtual performance spaces, especially in the piece for test 232b which had a wide pitch range and low tessitura (musical range of pitches). This is an area which should be further investigated, and could be facilitated by adjustments to the VSS.

6.7.2 Real vs. Virtual

In Test 221 the fitted property vector of Global Tempo points towards most of the virtual fragments reflecting faster tempos in the virtual performance space, although similar differences do not appear for the other singers in this research.

Ueno et al. found that when singer support was lower, singers became more careful in production and hence tempo was slower, whereas Schärer Kalkandjiev and Wienzierl [10] found increased tempo (for a cellist) was predicted by low levels of G_e stemming from fewer

or lower level early reflections on the concert hall stage. Tempo was also highly influenced by reverberation times, whereby the musician might lengthen notes to compensate for a lack of room reverberation and hence a slower tempo in drier acoustics, but on the other hand a desire to stop notes blurring into each other might also lead to a slower tempo in highly reverberant spaces. The interaction between the room acoustic parameters of Strength, Stage Support and Reverberation Time is not clear-cut, may differ for different musicians and styles of music, and is an area ready for further parametric investigation.

For the mezzo-soprano (Test 212) intonation accuracy, as measured by interval error, correlated significantly with differences between the real and virtual performance spaces with higher values of interval error - less accurate pitching - correlating with the real space.

For the tenor (Test 221) pitch precision points towards mainly virtual fragments and away from the real performance space fragments, again suggesting that intonation is more difficult in the virtual performance space. However, tempo is also correlated in this way, so it might in fact be increased tempo which leads this singer to less precise pitching.

For the bass singer too, although no correlation analysis was undertaken, in Test 232a the interval precision vector points toward the virtual fragments and away from the real fragments, indicating that intonation was less stable in the virtual performance space. This is reflected by this particular singer's comments that he found intonation difficult to maintain in the virtual performance space.

However, for the same singer, in Test 232b, intonation precision values seem to increase (less precise singing) in the real performance space and point away from the virtual fragments, which were all recorded in the MR configuration. Again this reflects the singer's comments that intonation was more difficult in this test in the real space, but easier in the medium reverberant setting.

Both differences in tempo and differences in intonation accuracy in the virtual performance space are likely explained by the differences in relative levels of energy in the early and late parts of the room impulse response in different regions of the spectrum (see Section 3.5.4). These differences lead to a mismatch as evaluated in terms of Support, a stage acoustic parameter which has been shown to be important for a musician's sense of support for his/her own sound.

Errors in Early and Late Support values were shown to exist in the VSS at both low and mid frequencies. It was suggested in Section 3.5.4 that these errors may arise from a lower level of floor reflection in the VSS compared to the real space, and reduced reverberation time in the 2000Hz octave band. Both of these aspects might affect singers' intonation in the VSS.

The positive errors in Early and Late Support in the simulation in the lower octave

bands (250Hz and 500Hz) might suggest that the singer has more difficulty "hearing oneself" in the virtual space and hence intonation will be impaired. Nevertheless not all singers exhibited less precise or less accurate intonation in the virtual space. The bass singer was less precise at producing intervals in the real space and the mezzo-soprano was less accurate in producing intervals in the real space.

Overall the effect of congruent (real/real or virtual/virtual) or non-congruent (real/virtual) pairs was significant on the similarity ratings by listeners. This suggests, together with the aspects outlined above, that the virtual simulation is not yet a sufficiently accurate match to the real performance space, and that further improvements should be made: these are outlined in Section 7.1.1

6.7.3 Emotional Content

The most salient variation in the mezzo-soprano performances seems to be the emotional intent of the performance which was able to be investigated here as different verses of the same piece with different emotional content were investigated. Both F0VE and tempo were highly and significantly correlated with the verse number, whereby the earlier verses (see Appendix G), which set the scene of the story and could be described as having neutral emotion, show higher values of F0VE and tempo whereas the later verses, which are more sorrowful, are characterised by smaller vibrato extent and slower tempo.

6.7.4 Unexplained differences

In some of the tests, not all of the posited perceptual dimensions as captured by the MDS analysis are explained by the performance attributes analysed. For example, perceptual dimension 2 is not well explained in tests 232b (bass) and 221 (tenor), and similarly dimension 3 does not correlate well to any attribute analysed in test 212 (mezzo-soprano). Fragments grouped closely along these dimensions could be similar in ways not captured by the performance parameters analysed in this thesis.

Listener comments often include details about idiosyncratic differences between the fragments such as variation in breathing between phrases, contrasts between the first and second half of the fragment, changes in tempo (rubato) during the fragment. These shorter-term features, such as articulation of consonants, different breathing patterns, or lengthening of certain notes in comparison to others are not captured by average values used in this analysis. Average values, such as those analysed here, lose some of the detailing in the phrases which singers use to communicate emotional intent. Future work should seek some way of capturing the "shape" and "direction" of phrases as well as differences between individual notes in the phrase.

6.7.5 Limitations of the study - future work

There was a desire to provide realistic rather than fully controlled experimental conditions whereby singers were asked to sing in the virtual and real spaces as if rehearsing prior to a performance. For this reason it was not always possible to extract a parametric set of recordings of a set number of productions of each fragment in each acoustic configuration in the real and virtual settings.

6.8 Conclusions

This chapter outlined the analysis, comparison and evaluation of recordings made by solo singers made in the *real performance space* and the *virtual performance space*.

Reports by singers about their own singing (Section 6.2) were combined with listener comments and the results of a perceptual listening test and subsequent MDS analysis, as described in Section 6.3. In addition, objective musical performance attributes of the solo singing were analysed and used to describe some of the perceived similarities between sung fragments. Tempo, vibrato extent and intonation precision accounted for most of the variability in listeners' perception.

Singing fragments recorded in the different acoustic configurations (LC/MR/SP - most reverberant to less reverberant) were perceived on the whole as being different along at least one perceptual dimension. More often than not this dimension was explained by one of the performance attributes examined here, but not all perceptual dimensions have been correlated to performance attributes.

The main impetus for the listening tests is to judge similarity between singing produced in the real and virtual performance spaces and hence to gather evidence that the room acoustic simulation is indeed plausible. Listeners did not systematically perceive a difference between fragments recorded in the real space in contrast to those in the virtual performance space.

Indeed, in all tests the fragments recorded in the real and virtual spaces, across the configurations, share perceived similarities explained by at least one attribute. However, some of the dimensions of the perceptual space from the MDS analysis are not explained by the performance attributes analysed here. Other performance parameters, such as dynamics (loudness), articulation of the text and vocal timbre could explain some of the perceived similarities and such be studied in future work in this area.

On the whole, global tempo explained the perceived similarities between fragments for the majority of the singers, with differences in vibrato extent and intonation precision also playing a role for listeners' judgements of similarity.

Chapter 7

Conclusion

The main aim of this research was to investigate the perception of singing performances in real acoustic environments and room acoustic simulations, focussing in particular on perceived similarities as judged by listeners.

In order to investigate these singing performances, an interactive room acoustic simulation of a performance space was implemented and evaluated by professional singers as being sufficiently realistic to elicit alterations in their singing performance, as would be expected in different room acoustic conditions of real performance spaces.

7.0.1 Summary of Thesis

Chapter 1 - Introduction

Chapter 1 introduced the idea that singing performance alters according to the acoustic characteristics of stage and auditorium of the performance venue. The need to use anechoic recordings of musical sources in auralisation was described, and it was suggested that anechoic recordings of musicians in general, and singers in particular, would not fully reflect recordings made in real venues, because of the highly unnatural nature of recording in a room devoid of sound reflections. This introductory chapter set out the need for the implementation of a real-time room acoustic simulation, which would allow a singer to sing in an acoustically treated room and hear his/herself as if in a concert. If such a simulation were plausible it would not only allow the study of how singing performance changes in different acoustic conditions, but also provide a means for obtaining more natural "dry enough" (but not anechoic) recordings for subsequent use in auralisations of concert halls and other performance spaces.

Chapter 2 - Simulating Room Acoustics

Chapter 2 began by introducing the study of room acoustics and describing Spatial Room Impulse Response measurement techniques, which were used in the implementation of the VSS. Recent work on developing techniques for auralization, virtual acoustic environments and interactive real-time room acoustic simulations were outlined. The importance and relevance for musicians of particular room acoustic parameters such as the presence of early reflections, levels of Stage Support, Running Reverberation (RR160) and reverberation time T30 and it was these parameters which were used in the verification of the VSS in the next chapter.

Chapter 3 - Virtual Singing Studio: Implementation and Verification

Chapter 3 described the implementation of the Virtual Singing Studio, which provided the singer with the ability to sing and hear his/herself as if in a real performance venue. Three different room acoustic configurations were simulated, mirroring those available in the real performance venue which differed most obviously in terms of reverberation time. Subjective responses of the singers were collected through questionnaire and interview regarding the naturalness and plausibility of the virtual space, ease of use and the effect the different simulated room acoustic on their singing performance.

Errors introduced by the signal processing implemented in the VSS were assessed by comparing room acoustic parameters of the real and virtual performance space. It was shown however that a plausible interactive room acoustic simulation was implemented in the VSS, with most errors lying within the subjective limen chosen.

All singers who used the VSS were able to state a preference for one or more of the three room acoustic configurations which were simulated, and were also able to describe the similarities between the simulated space and the room acoustics of the real venue. Furthermore, all singers rated the VSS highly, reported that they enjoyed singing in the virtual space and agreed that it replicated the real performance venue. The comparison of singers' experience in the real venue and the simulation is one of the novel aspects of this thesis.

Chapter 4 - Singing in Space(s)

Chapter 4 outlined some of the previous work which has investigated changes in musical performance in general, and singing performance in particular according to the acoustic conditions of the performance venue. After a brief history of music performance analysis research a number of available techniques for extracting music performance parameters were outlined. Previous findings relating to the analysis of vibrato, tempo and intonation

were presented and the results of more recent investigations of alterations in musical performance in different room acoustics were examined. A case study which involved a vocal quartet singing in three room acoustic configurations in the real performance space and also in the anechoic chamber showed that singers recognise the differences between acoustic settings, but are able to describe their own impressions of how their singing might change.

Chapter 5 - Singing Performance Analysis and Evaluation

This chapter examined some of the methods available for the perceptual evaluation of similarity of audio and music. It also laid out some of the methods of correlating objective and subjective data used recently by other researchers in similar areas such as PCA, correlation and regression analyses. It tested the method proposed for producing material for the main listening test described in Chapter 6. A second pilot listening test showed that a sorting task was a suitable method of obtaining similarity ratings for a number of audio objects (fragments of recorded singing). A MDS analysis carried out showed that listeners could determine differences between singing performances recorded in different acoustic settings, and groups of fragments made in the same settings were grouped together in the MDS perceptual maps.

Chapter 6 - Singing in Real and Virtual Acoustic Environments

In chapter 6 Multi-dimensional Scaling analysis was used to reduce the high-dimensionality of listeners' similarity ratings of singing recorded in the real and virtual performance spaces. The resulting "common perceptual maps" provided a means of visualising the similarity data.

A number of objective singing performance attributes, namely vibrato rate and extent, tempo and intonation accuracy and precision, were analysed to allow comparison between singing performances by singers in the virtual and real performance spaces. These objective attributes were then correlated to the perceived similarities between sung fragments, as visualised in the MDS analyses, in order to infer the interpretations of the dimensions of the MDS derived perceptual maps.

Vibrato rate did not exhibit any significant variation between singers or within the fragments for each singer and did not appear to be important for listeners' perception of the singing. On the other hand vibrato extent and tempo were shown to be highly significant in explaining a large proportion of the perceived similarity between singing performances. The use of dimensionality reduction analyses is novel in regards to the assessment of singing performances and has been shown to be a useful complement to

traditional objective acoustic analysis and other methods of perceptual evaluation.

Since no significant effect of room acoustic conditions was found for all but one of the performance attributes measured, it has not been conclusively shown that the simulation implemented in the VSS elicits changes in singing performance according to the different room acoustic conditions. Nevertheless, some changes in singing according to the different acoustic configurations were reported by singers who took part in the study, and commented upon by listeners such as vibrato extent and global tempo.

It is interesting that performances in the virtual performance space and real performance space did not match in terms of intonation accuracy and precision. However, intonation was not always less precise in the virtual space - in fact two of the singers exhibited differences in inaccuracy and precision of musical interval production in the real space (but not in the virtual). Nevertheless interview and questionnaire feedback from the singers, and comparison of objective Support parameters, show that levels of Stage Support and the ability of the singer to "hear his/her own voice" is altered in the room acoustic simulation. Future work will seek to overcome this limitation.

7.0.2 Restatement of Hypothesis

The main hypothesis of this work was:

> **A plausible interactive room acoustic simulation will elicit changes in singing performance which replicate those occurring in different real acoustic environments.**

The hypothesis was supported by:

1. Rendering a virtual simulation of a performance space which allows a singer/speaker to hear their vocal performance in real-time as if in the real performance space.
    - The Virtual Singing Studio was implemented and judged to be effective by all singers who took part in the trial and recordings in the room acoustic simulation.
2. Comparing objective room acoustic measurements of the real space and the virtual simulation.
    - Objective room acoustic parameters were measured in the real and virtual spaces and compared. EDT values matched least well whereas T30 values were better replicated and within the double subjective limen. Levels of Stage Support - ST_{early}, ST_{late} and ST_{total} - exhibited errors in the 500Hz octave

band and the 200Hz octave band. Running Reverberation values were well matched across the spectrum.

3. Recording vocal performances in the virtual and real performance spaces.

    - Seven professional singers used the virtual singing studio and were recorded in the virtual space. They were also recorded singing the same pieces in the real performance space. All the singers advised that the virtual space matched their impressions of the real performance venue and reported their enjoyment of singing in the VSS.

4. Collecting subjective responses from singers about their own performances in the real and virtual space.

    - Singers who sang in the real and virtual spaces were interviewed and also completed questionnaires asking them to rate specific perceptual aspects of their experience of singing in both spaces. All singers were able to recognise and articulate the different characteristics of the varied acoustic configurations in the real venue and the simulation. Ratings of the perceptual aspects of the different acoustic settings reflected the variation seen in T30 and ST.

5. Analysing and comparing vocal performance parameters of singing in the real and virtual space.

    - Tempo, fundamental frequency rate and extent, and three measures of intonation, namely MAIP, MAIE and MAPP were analysed in a number of productions of phrases chosen from the singing performances studied.

6. Asking listeners to judge the similarities between vocal performances recorded in the real and virtual space.

    - A listening test incorporating a sorting task allowed listeners to judge the similarities between fragments of the recorded vocal performances. Dimensionality reduction analyses of the perceptual similarity data showed that singing performances do indeed change according to room acoustic conditions. Nevertheless, there was a significant effect of *simulation* on the similarity ratings of pairs of vocal performances, suggesting that although some performance attributes changed according to room acoustic configurations, the simulation does not yet sufficiently replicate the real performance space

Although there were some mismatches in the patterning of intonation accuracy and precision between the real and virtual performance spaces, tempo and vibrato extent changes, which accounted for a largest proportion of the perceived differences between sung fragments, matched across the real space and the simulation. Overall the Virtual Singing Studio was indeed able to provide a "plausible" room acoustic simulation for use by singers to hear themselves sing as if in a real performance venue. However, it is not yet fully "realistic" enough to replicate the real performance venue, and further improvements have been suggested, which include adding a visual element to the simulation.

The majority of studies of musical performance characteristics have involved a small number of participants, for example, one soloist or one duet pair ([137, 114, 12, 11]). No other studies have yet involved the analysis of more than five musicians, (e.g.[14, 15, 277, 3]) since collecting and analysing musical performance data from larger numbers of musicians is still a time consuming and onerous task. Music information retrieval techniques have already shown some benefit in this area, and will continue to provide increasingly effective means of analysing data from larger groups of musicians.

Although only a small number of singers (seven) participated in the study, their reported experiences of singing in the virtual and real performance space showed much similarity. All singers were experienced professional singers with specialism in the performance of early music (pre-1750). All singers involved in the study spent a reasonable time singing in the different acoustic conditions of both the real and virtual spaces and representative performances from three of these singers were used in the musical performance analysis. Although it is accepted that the results of this study may not be fully generalisable to all singers in all genres, this study has nevertheless provided a worthwhile contribution to the implementation of room acoustic simulations for research into musical performance and perception.

7.1 Further work

7.1.1 Improving the VSS

In the room acoustic simulation three out of four of the singers exhibited differences in intonation in comparison to their performances in the real space. It was argued that this might stem from the lack of some early reflections leading to lower levels of *Support* in the simulation, suggesting that control feedback loop for the singer might not fully replicate that which is available in the real space. Alternatively intonation could stem from the difference in the real and simulated reverberation times in the upper regions of the frequency spectrum (with errors particularly apparent in the VSS in the 2000 Hz

octave band) which might lead to a shift in spectral locus of the reverberant soundfield.

Rendering the Soundfield

The VSS described in this thesis was implemented using first order Ambisonics decoded for presentation over sixteen loudspeakers in an octagon plus cube array.

It is possible that the use of other rendering techniques, such as Higher Order Ambisonics or SIRR could improve the simulation, by replicating better the timing and direction of early reflections, and might also provide an increased sweet-spot and better off-centre listening experience.

VSS for multiple singers

Since source-localisation was no of immediate concern in the VSS first order Ambisonics was used quite adequately; However, it is suggested that higher order Ambisonics or SIRR might improve the simulation of early reflections leading to increased levels of singer *Support*, which has been shown to be an important aspect of musicians' impressions of stage acoustics. Such techniques would also increase the sweet spot and might allow small groups of singers to use the VSS.

Woszczyk et al. [12] have shown that it is possible to present a virtual acoustic performance space which can be shared by more than one musician and an audience. This enables the communication between musicians and listeners to be effectively represented. Further development of the VSS might enable more than one singer to perform together for rehearsal.

Listening position for performers

Future work would seek to include the possibility for the performer to hear their own performance from listener position, or indeed a variety of positions in the auditorium or on stage. This was suggested by professional musician Tom Beghinn who took part in investigations of virtual stage acoustics undertaken by Woszczyk et al. [13] that would allow musicians to monitor their performance being and provide a very useful tool for training and rehearsal purposes.

Calibration

An improved method of calibrating the VSS should be developed. In this regard recent research by Brunskog [53] who measured Room Gain (G_{RG}) and by Pelegrín- Garcia [54] who measured Voice Support could be implemented. A head and torso simulator is needed for the measurement of these room acoustic parameters.

Dammerud measured total, early and late parameters of G (Strength) for concert hall stages in order to compare across different concert halls. In all cases the reference level is an impulse response taken with the same measurement equipment in the same set up in an anechoic chamber measured with a source-to-receiver distance of 10m.

Kalkandjiev and Wienzierl [10] balanced the reproduction level of a virtual acoustic environment for performance by recording a single cello tone, both by the instrument close-microphone and dummy head at a distance of 5m. This recorded tone was then played through a binaural simulation of an anechoic chamber over headphones on a dummy head (HATS) with a source-receiver distance of 5m. A subsequent recording was made and the RMS of both dummy head recordings were matched using a scaling factor.

Brunskog [53] measured Room Gain (G_{RG}) which relates the energy contained in the impulse response measured between mouth and ear of a dummy head and torso (HATS) with that of a corresponding measurement in an anechoic chamber where only direct sound is present.

Further Verification

Implementing methods to measure more recently posited parameters such as Voice Support and Room Gain (outlined in Section 2.4.3), both of which involve the use of a Head and Torso Simulator (HATS) would allow additional performer-relevant room acoustic parameters to be compared between the real venue and the simulation.

7.1.2 Music Performance Analysis

Some of the perceived differences between sung fragments were not captured by the singing performance attributes analysed in this research. One salient feature of sung performances is likely the articulation of the text. Some useful way of measuring and quantifying "articulation" as an attribute should be investigated, which would probably include temporal and spectral characteristics.

Loudness was not studied in this thesis, but recently Timmers has shown that attributes which combine measures of dynamics and tempo can account significantly for variations in listeners' perception of [CORRECTION] musical performance. [224]

The length of notes, as measured by note-on-ratio, was not analysed in this thesis, but it was commented on by listeners, and other authors such as Kato et al. [15] have found changes in this parameter according to reverberation time of the space.

Future work would seek to include the analysis of a greater number of performance parameters, some of which might account for some of the unexplained perceptual dimensions of the MDS analyses presented in this thesis.

It was not possible in this study to make comparisons between singers, as they all sang different pieces. Future work might be improved by asking the singers to sing the same piece, for example "Ave Maria" (Schubert) which is used often by researchers in this area, and would allow comparison between singers and to other studies. A balance of fragments recorded, ensuring an equal number of real vs virtual performance space recordings and in each of the three different acoustic configurations (LC, MR, SP) would also facilitate comparison and statistical analysis.

Future work might seek to include a measure of intonation which does not rely on absolute values, but rather maintains the signed value. This might be informative, as it is possible that singers might differ in the direction of error in different acoustic environments e.g. singing"sharp" in reverberant environments. The intonation metrics used in this study lose this aspect of intonation practice.

7.2 Application of research

VSS for rehearsal and performance

If the simulation of room acoustics for the performer can be plausibly rendered, then there is potential for singers to use the VSS as a rehearsal tool with subsequent investigations as to whether the VSS enables singers to adjust to room acoustic conditions more quickly and effectively. This would provide a useful technology for young professional singers, who, as Woszsyk notes [13] often spend more time in rehearsal rooms than on stage.

7.2.1 Vocal health

The research will also contribute to increased understanding of the role of acoustic environment in vocal use which in turn will contribute to a better understanding vocal loading and its implications for vocal health.

7.2.2 Concert hall and other architectural design

A better understanding of performer preferences of concert hall platforms and room acoustics of other performance spaces will contribute to improved concert hall design in the future. In addition, improvements in virtual simulations of room acoustic conditions, will inform better auralizations for concert hall designers, architects and clients.

Not only will this research inform concert hall design, but also general building design for rooms where inhabitants use their voices on a day-to-day basis, for example, council chambers, lecture rooms and teaching classrooms.

7.2.3 Improving SRIR models

The evaluation of room acoustic auralisations through the VSS could be used to help improve the modelling of room impulse responses by identifying which aspects of a room impulse response and resulting sound field are perceptually relevant to the performing musician.This in turn might lead to the production of a simplified spatial room impulse response specifically tailored for use in such systems which could be generated through room acoustic modelling systems, or modified from measured SRIRs, specifically for use in real-time interactive performance systems such as the VSS. Such investigation may seek to find out how simplified the model can be and still be useful for real-time perception of a real performance space.

7.2.4 Real-time convolution

The present research also informs applications which increasingly use real-time auditory scene simulation and real-time convolution for auralisation purposes. For example, recent work by Favrot et al, [5] has used loudspeaker based room auralisation for hearing-aid research. The present study will produce results which might also be applicable to this field of research.

7.2.5 Psychoacoustics

This research project informs the field of psychoacoustics, in that it offer some insights into the perception of sound for the performer, and the psychoacoustics of sound in virtual environments. It also helps to identify the room acoustic parameters which are perceptually relevant for the musician during performance. With future development it would allow this aspect to be further investigated through greater control of the room acoustic parameters involved in the simulation of the performance space.

7.2.6 Application to Virtual Reality Research

This research project also adds to our understanding of how we model and interact with our environment. In virtual reality applications the perceptual evaluation of the virtual simulation is the ultimate aim. Indeed, with relation to the development of computational based auralization Kearney states that " The development of perceptually-based topologies ... warrants further investigation" [44].

7.3 Final Remarks

Chapter 1 outlined the need for improved musical source material for use in future auralisations. The VSS will facilitate the recording of, not anechoic, but "dry enough" ([312]) recordings of singers which still retain the characteristics of a performance as if it were produced in a real performance space, i.e. through the use of close microphone recordings in a virtual space. The VSS could easily be extended to be useful for other instrumentalists, and with further development eventually for groups of musicians.

The VSS does indeed bring us one good step closer to what Gade requested:

> *The primary requirement for carrying out relevant experiments [into musicians' perception] is that room acoustic sound fields of proper realism and with the possibility of changing variables of potential importance can be presented to musicians while playing* [2].

The VSS has been shown to provide a plausible (realistic) sound field whilst the musician is playing - or in this case whilst the singer is singing. Since it has been proven that the simulated sound field replicates a real performance space well enough to elicit the relevant changes in singing performance attributes, it can now be developed in order to allow "variables of potential importance" to be altered whilst the singer is performing. This would allow us to investigate further the question of musician's perception and subsequent action as it relates to room acoustic conditions and musical expression.

A cognitive model, drawn up by Ueno and Tachibana [107, 3] of a musician's perception in a concert hall was presented in Section 4.4.6 and is reproduced in Figure 7.1

Within the schematic, "personal skill" includes perception not only of the concert hall (performance venue) acoustics, but also of musical expression. Results of the analyses undertaken in this thesis suggest that there might be a more complex relationship between these two parts of musical and acoustic perception.

In a study of solo piano playing, Repp [222] argued that changes in musical tempo and timing follow a hierarchical model which is framed by the musical structure of the piece. Major tempo changes indicate larger structural components, and variation of expressive timing for individual phrases (subject of a model first proposed by Todd [156]) operates within these structural constraints. At lower levels of this hierarchical structure Repp found much individual variation between pianists.

Evidence of the interplay between vocal performance attribute variations in response to room acoustic conditions and variations used to convey emotional expression might suggest that a similar hierarchical structure exists here between "perception of hall acoustics" and "perception of musical expression". This is an area which future research might seek to address.

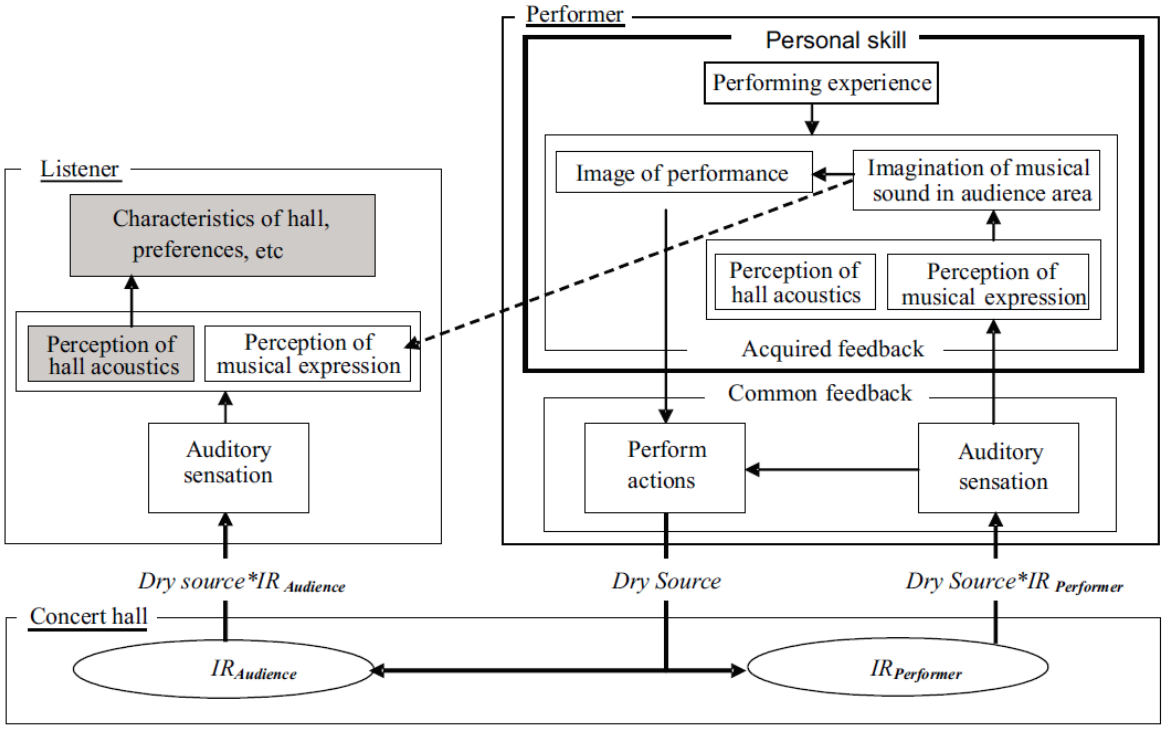


Figure 7.1: *Schematic model of a performer and a listener in a concert hall, from [3]*

The investigations in this thesis do also add weight to the proposed distinction of two types of circulative feedback system between performer and room : one being the common feedback system of action and feedback, in automatic response, the other is acquired and related to the musician's own skill and learned responses.

Differences in tempo and vibrato extent, for example, might be learned responses to changes in Stage Support and Reverberation Time. On the other hand, one might suppose that no singer would actively attempt to sing with less precision or accuracy of intonation and thus changes in tuning patterns could be a reflexive action stemming from the automatic response system.

The VSS can be developed and optimised to help investigate such questions, by designing experiments whereby room acoustic parameters are altered and ensuing changes in singing performance captured, analysed and compared. Knowledge gained through future investigations of this kind will inform conceptual models, such as the one described above [107, 3]. What is more, this research, whilst grounded in a musical context, will ultimately enrich our understanding of human cognition, perception and action.

Appendix A

Index of Supporting Media DVD

Supporting media is organised in folders on the disc accompanying this thesis according to chapters, as follows :

Chapter 2

Deconvolution_Code Matlab code for deconvolution of recorded sine sweeps used to measure SRIR of the real performance space. (Section 3.4.1)

Chapter 3

Microphone_Recordings Recordings of phrase "peter piper..." recorded via head-mounted, overhead and baseball cap microphones (Section 3.4.3)

Input_Impulse_Responses_Performer SRIRs recorded in performer positions (A-D) in the *real performance space* (Section 3.4.1)

Input_Impulse_Responses_Listener SRIRs recorded in listener position in the *real performance space* (Section 3.4.1)

Output_Impulse_Responses SRIRs recorded in the *virtual performance space* (Section 5.5.1)

Chapter 5

Quartet_Scores Scores of "Audivi Vocem" and "Remember Not" sung by the quartet in the real performance space (Section 5.4 and Section 5.5)

ABX_Fragments Fragments used as stimuli for pilot listening test (Section 5.4)

Pilot_Quartet_Test Fragments of quartet singing used in pilot listening test (Section 5.5)

Pilot_Test_Instructions Instructions for participants in the pilot listening test (Section 5.5)

Chapter 6

Solo_Scores Musical scores of pieces performed by singers in the VSS (Section 6.2)

Main_Test_Fragments Fragments of solo singing used as stimuli in main listening test (Section 6.3)

Example_MIDI MIDI files (as in score and as performed) of "De Domo" example fragment (Section 6.4.1)

Ampact_Additions Matlab scripts for additional functionality written by the author for use with AMPACT Toolbox [150] `http://ampact.tumblr.com/` (Section 6.4)

Main_Test_Instructions Instructions for participants in the main listening test (Section 6.3)

Conference Papers

Conference_Papers Copies of conference papers by the author on related work (2001-2014)

Appendix B

Protocol and questionnaire for initial experiment

VIRTUAL ACOUSTIC – PILOT PROJECT

Thank you for agreeing to take part in this experiment

Your singing will be recorded via the microphone and used for acoustic analysis. At the same time, the electrolaryngograph (EGG) data will be recorded for later analysis.

Your data and recordings will remain anonymous at all times, and will only be used for research purposes.

You are able to withdraw from the experiment at any time and you do not have to give a reason. If you decide to withdraw any recorded data or audio will be destroyed.

In the experiment I would like to record your singing as well as data from the electrolaryngograph.

The experiment will run in 2 sections:

In each section you will be asked to sing a small number of vocal tasks followed by a short extract from a piece of your choice.

At the end of each section please complete the questionnaire overleaf . The questionnaire is designed to capture your subjective response to the acoustic characteristics of the acoustic you heard in the experiment. Notes on the definitions of the acoustic characteristics are given on the reverse side of the questionnaire pages.

Notes on the characteristics used in the acoustic assessment

Loudness	The volume or level of sound
Clarity	The extent to which individual notes are clearly distinguishable one from another
Reverberance	Persistence of sound after the interruption of the music ('dry' means little reverberance – 'live' is more reverberant)
Envelopment	Sense of immersion in the sound field
Intimacy	Perception of the spatial dimensions of the space
Warmth	The strength of low frequencies(>350Hz) in relation to the medium tones (350-1400 Hz)
Brilliance	The effect of vivacity arising from the harmonic richness
Timbre	The quality of sound that distinguishes one voice or musical instrument from another

PARTICIPANT Number: **PART ONE** **QUESTIONNAIRE**

Volume

Loud		Moderate			Subdued			Quiet	
1	2	3	4	5	6	7	8	9	10

Notes:

Clarity

Muddy		Blurred			Distinct			Clear	
1	2	3	4	5	6	7	8	9	10

Notes:

Reverb-erance

Dry		Medium Dry			Medium Live			Live	
1	2	3	4	5	6	7	8	9	10

Notes:

Envelop-ment

Frontal		Direct			Diffused			Enveloping	
1	2	3	4	5	6	7	8	9	10

Notes:

Intimacy

Remote		Distant			Close			Intimate	
1	2	3	4	5	6	7	8	9	10

Notes:

Warmth

Harsh/Thin		Moderate			Balanced			Warm	
1	2	3	4	5	6	7	8	9	10

Notes:

Brilliance

Dull		Average			Crisp			Bright	
1	2	3	4	5	6	7	8	9	10

Notes:

Timbre

Unpleasing		Balanced			Pleasant			Beautiful	
1	2	3	4	5	6	7	8	9	10

Notes:

Overall impression of the acoustics

Poor		Satisfactory			Good/very good			Excellent	
1	2	3	4	5	6	7	8	9	10

Notes:

PARTICIPANT Number: **PART TWO** **QUESTIONNAIRE**

Volume

Loud		Moderate			Subdued			Quiet	
1	2	3	4	5	6	7	8	9	10

Notes:

Clarity

Muddy		Blurred			Distinct			Clear	
1	2	3	4	5	6	7	8	9	10

Notes:

Reverb-erance

Dry		Medium Dry			Medium Live			Live	
1	2	3	4	5	6	7	8	9	10

Notes:

Envelop-ment

Frontal		Direct			Diffused			Enveloping	
1	2	3	4	5	6	7	8	9	10

Notes:

Intimacy

Remote		Distant			Close			Intimate	
1	2	3	4	5	6	7	8	9	10

Notes:

Warmth

Harsh/Thin		Moderate			Balanced			Warm	
1	2	3	4	5	6	7	8	9	10

Notes:

Brilliance

Dull		Average			Crisp			Bright	
1	2	3	4	5	6	7	8	9	10

Notes:

Timbre

Unpleasing		Balanced			Pleasant			Beautiful	
1	2	3	4	5	6	7	8	9	10

Notes:

Overall impression of the acoustics

Poor		Satisfactory			Good/very good			Excellent	
1	2	3	4	5	6	7	8	9	10

Notes:

Appendix C

Instructions for Participants in Main Listening Test

Singing Perfromance - Listening Tests
24th October – 8th November 2013

Investigator Name: Jude Brereton, AudioLab, Dept of Electronics, University of York, York, YO10 5DD

Title of Study: Listening tests on recordings of singing voice.

Brief Description of Study:

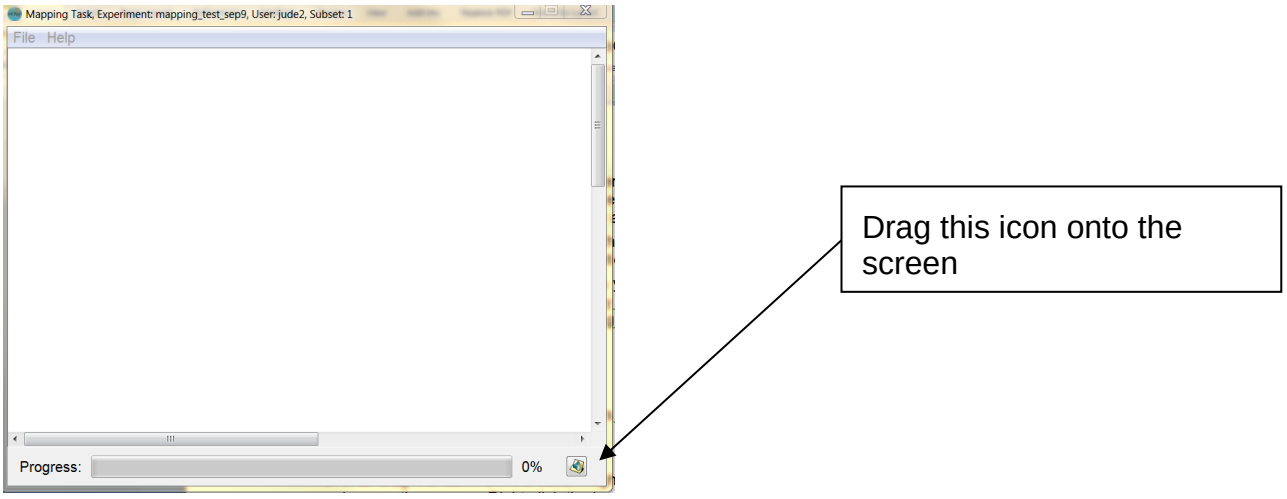
The purpose of this study is to evaluate a number of singing performances which have been recorded by a solo singer on different occasions.

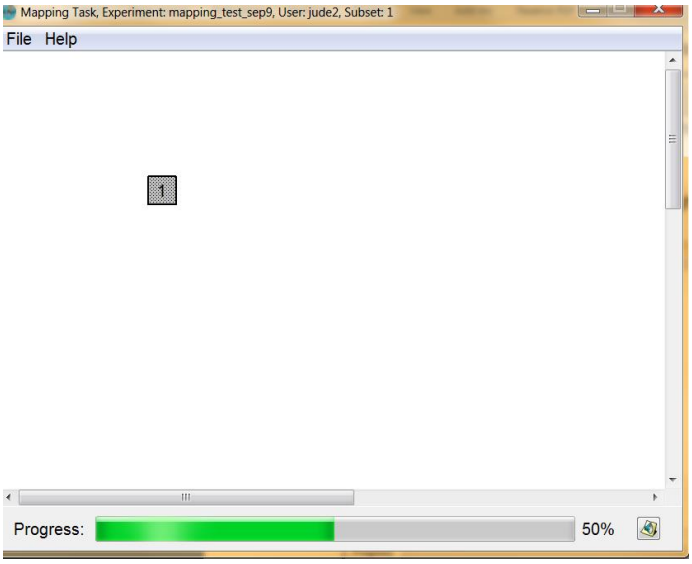
You will be asked to listen to a number of short recordings of sung phrases and to evaluate how similar or dissimilar they are.

Instructions to participants

In this test you will hear the some fragments of singing performance, and each fragment (sound item) will have its own square icon. You are asked to place sound items within a two-dimensional space according to their similarity.

First of all drag the item from *sound item dispenser* in the lower right corner of the window into the workspace.





(When all sounds have been dragged into the workspace, the dispenser will disappear.)

Individual sounds can be played by right-clicking on their icons and selecting the "Play" popup menu item.

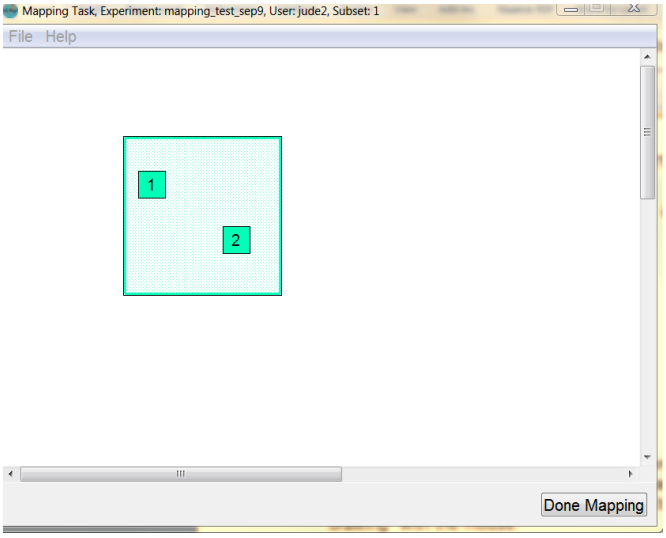
Within the workspace, sound items can be dragged at will.

- When a sound item is selected (by clicking on it), it will become shaded and its border will darken.
- Multiple sound items can be selected by depressing the SHIFT key when clicking on them, or by "drag-drawing" a selection rectangle around them.
- Multiple selected items can be moved together by click-dragging on any of the selected items.
- All selected items can be de-selected by clicking on any blank area of the workspace.
- A single item can be de-selected by SHIFT-clicking it, without affecting the state of other selected items.

Sound items can be labelled by right-clicking on their icons and selecting the "Label" popup menu item. When the cursor is over a sound item, the label will be displayed as a tooltip, as well as in the status bar at the bottom of the window.

Grouping sound items

If you think that one or more sound items are highly similar for a particular reason (e.g. they all sound ‘out-of-tune’) please put a grouping box around them and label the box. This can be done by pressing CTRL key while “drag-drawing” with the mouse.



Subgroups can be created within groups. Items and groups are grouped/ungrouped simply by dragging them into or out of group boxes or by creating group boxes around them. If you mouse-click on a group border, that group and all its members are selected.

Notes about evaluating similarity of performances

Please try to listen to the **performance** of the music itself – rather than any other audio quality such as room reverberation, reproduction noise, microphone blips etc.

When you are happy that you have listened to all items and grouped them accordingly please press "done mapping".

Please note, there are no "right" or "wrong" answers – we are interested in your own evaluation/judgements of similarity – you might think that all fragments are highly similar, or that they are all completely different.

Please do not hesitate to ask the investigator for further instruction either before or during the test.

Please feel free to take a break at any time during the test.

Thank you for your participation.

Appendix D

Room Acoustic Parameters of Real Performance Space

Real Performance Space - Large Choral Setting (LC)								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean Across All Octave Bands**
EDT (s)	1.192	1.159	1.672	1.914	1.744	1.286	0.835	1.400
T30 (s)	1.746	2.038	2.251	2.235	2.070	1.608	0.969	1.845
ST_{early} (dB)	-10.237	-14.163	-13.610	-10.973	-9.820	-8.523	-7.986	-10.759
ST_{late} (dB)	-10.132	-12.024	-12.103	-9.844	-9.384	-9.440	-12.330	-10.751
ST_{total} (dB)	-7.115	-9.908	-9.772	-7.344	-6.546	-5.911	-6.589	-7.598
RR_{160} (dB)	-15.634	-16.972	-16.181	-14.223	-13.885	-14.436	-18.494	-15.689

Table D.1: *Mean values of EDT, T30, -ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions the Large Choral Setting of the Real Performance Space*

Real Performance Space - Music Recital Setting (MR)								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean across all octave bands**
EDT (s)	1.038	1.024	1.400	1.503	1.404	1.117	0.786	1.182
T30 (s)	1.641	1.842	1.919	1.901	1.744	1.369	0.866	1.612
ST_{early} (dB)	-10.410	-14.326	-13.785	-11.268	-10.135	-8.686	-8.487	-11.014
ST_{late} (dB)	-10.563	-12.505	-12.933	-11.100	-10.387	-10.432	-13.510	-11.633
ST_{total} (dB)	-7.414	-10.273	-10.315	-8.155	-7.206	-6.431	-7.269	-8.152
RR_{160} (dB)	-16.074	-17.244	-16.849	-15.389	-14.950	-15.597	-20.021	-16.589

Table D.2: *Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Music Recital Setting of the Real Performance Space*

Real Performance Space - Speech Setting (SP)								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean across all octave bands**
EDT (s)	1.341	0.934	1.042	1.072	1.090	0.962	0.707	1.021
T30 (s)	1.852	1.818	1.491	1.411	1.325	1.092	0.751	1.392
ST_{early} (dB)	-10.422	-14.278	-13.842	-11.326	-10.229	-8.866	-8.380	-11.049
ST_{late} (dB)	-9.807	-12.881	-14.294	-12.844	-12.209	-11.735	-14.270	-12.577
ST_{total} (dB)	-7.033	-10.485	-11.040	-8.994	-8.067	-7.036	-7.364	-8.574
RR_{160} (dB)	-15.337	-17.463	-18.187	-17.232	-16.668	-17.146	-21.043	-17.582

Table D.3: *Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Speech Setting of the Real Performance Space*

Appendix E

Room Acoustic Parameters of Virtual Performance Space

Virtual Performance Space - Large Choral Setting								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean across all octave bands**
EDT (s)	0.279	1.543	1.690	1.702	0.231	0.912	0.619	0.997
T30 (s)	1.604	2.199	2.420	2.276	1.916	1.455	0.932	1.829
ST_{early} (dB)	-7.598	-7.979	-10.295	-10.260	-13.723	-9.135	-7.854	-9.549
ST_{late} (dB)	-16.107	-9.071	-10.401	-10.322	-18.078	-12.612	-14.555	-13.021
ST_{total} (dB)	-6.999	-5.429	-7.301	-7.265	-12.345	-7.510	-7.005	-7.693
RR_{160} (dB)	-20.922	-14.647	-14.902	-14.964	-22.766	-17.590	-20.692	-18.069

Table E.1: *Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Large Choral Setting of the Virtual Performance Space*

Virtual Performance Space - Music Recital Setting (MR)								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean across all octave bands**
EDT (s)	0.245	1.388	1.361	1.324	0.178	0.850	0.608	0.850
T30 (s)	1.532	1.996	2.072	1.927	1.557	1.252	0.828	1.595
ST_{early} (dB)	-7.632	-8.186	-10.366	-10.626	-13.740	-9.409	-7.927	-9.698
ST_{late} (dB)	-16.418	-9.605	-11.457	-11.403	-18.876	-13.281	-14.976	-13.717
ST_{total} (dB)	-7.065	-5.787	-7.803	-7.972	-12.559	-7.902	-7.133	-8.032
RR_{160} (dB)	-21.040	-14.830	-15.727	-15.920	-23.668	-18.390	-21.575	-18.736

Table E.2: *Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Music Recital Setting of the Virtual Performance Space*

Virtual Performance Space - Speech (SP)								
Octave Bands	**125 Hz**	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**8000 Hz**	**Mean across all octave bands**
EDT (s)	0.323	1.406	1.147	0.977	0.186	0.755	0.568	0.766
T30 (s)	1.805	1.887	1.602	1.439	1.154	1.015	0.727	1.375
ST_{early} (dB)	-7.669	-7.753	-10.360	-10.394	-13.679	-8.606	-7.520	-9.426
ST_{late} (dB)	-15.635	-9.504	-12.049	-12.604	-19.933	-13.625	-15.290	-14.091
ST_{total} (dB)	-6.977	-5.507	-8.082	-8.328	-12.744	-7.401	-6.838	-7.982
RR_{160} (dB)	-20.545	-14.609	-16.196	-17.248	-24.610	-19.219	-22.147	-19.225

Table E.3: *Mean values of EDT, T30,ST_{early}, ST_{late}, ST_{total} and RR160 averaged across the four performer positions in the Speech Setting of the Virtual Performance Space*

Appendix F

Comments by singers about the performance spaces

SP Speech setting, short reverberation time

MR Music Recital setting, medium reverberation time

LC Large Choral setting, longest reverberation time

For further details of room acoustic parameters of the above settings referred to in the comments below please see Appendix D

The Real Performance Space

Q. Do you think your singing performance changes between the Virtual Singing Studio and the Real Space - if so, please describe as fully as you can

Singer 201 - Soprano Visuals influence performer to communicate (with face/movement) more as in concert- harder to do without visuals

Singer 211 - Alto *no comments*

Singer 212 - Alto yes - actually perhaps say worse in real because trying to fill what I could see - consonants different

Singer 213 - Alto a much 'easier' space to sing in as setting 1 (LC) although the clarity of setting 2 (SP) was enjoyable

Singer 221 - Tenor 'sense of performance in the real space rather than a rehearsal in the virtual space

Singer 231 - Bass Think my diction was more in virt1(LC) and real2 (LC) - Handel went flat in virt1 (lc) and real2 (lc) yes - I often channel my voice visually (psychological) into the acoustic and I was looking around more here than at the curtain. However, I remember trying to imagine I was in the physical space I was virtually hearing , so perhaps not too much. Hard to say

Singer 232 - Bass just a bit - all settings in the low registers (especially 1 (lc) and 3(MR) felt easier to sing in the real venue than in the virtual one - Keeping pitch in LC mein Traures sohn C/Bb/A - D low D!! - use floor reflection more in low register, in LC tend to push loud, tempo not so good, scattered uncontrolled, timbre or reverb diffuse - in SP easier to grab timbre, in MR lower register is good.

The Virtual Performance Space

Q. Do you think your singing performance changes between settings? - if so, describe as fully as you can

Singer 201 - Soprano Yes - where venues give greater aural response I feel I can sing quieter to greater effect (you can bring the audience in to your expressive ideas of the performance, rather than project the idea to the audience). When I can hear the response of the room to my singing I feel more secure that my technique is working effectively and producing the effect I wish it to, and can then give a more expressive performance - better for audience and performer.

In the last setting (*SP*) I felt I could perform recitative more effectively as the clarity was better, and the resonance of the venue meant I was sure the audience would hear me clearly and I could create an intimate and expressive performance.

LC felt like a Cathedral, but with less clarity than *SP* (like the audience were further away, and I had to project the performance to them). (Ripon Cathedral, singing from the edge of the choir with the audience on the sides of the altar - like in their lunchtime recitals series) *MR* felt like a smaller venue, with good resonance for high notes but not a huge amount of aural feedback in general, so harder work to sing in. (Parish Church) *SP* felt like a cathedral, but where all pitches resonated well and were clear, meaning it felt more intimate - like the audience were nearer and could be drawn in to the performance. (York Minster - as if singing in the choir to audience in the stalls)

Singer 211 - Alto Yes, though I was making an effort to keep overall singing technique the same. The biggest changes I notice are tempo and phrasing choices. A dryer or clearer acoustic prompts me to take faster tempos, particularly with a very ornamental line that would be smeary with more reverberation. In a dry acoustic I find myself either choosing to sing a phrase in a more continuous, sustained manner, or working hard to create through my own singing the same kinds of blooming and tapering shapes that appear naturally when a note is released into a wetter acoustic.

MR was the most pleasurable in terms of receiving some feedback from the acoustic, which always makes it easier to produce the voice and is flattering to the sound, making imperfections less exposed. But if I were rehearsing a very detailed piece in an ensemble,

I might choose *SP* from time to time in order to enable the singers to hear exactly what is going on, with no help from the room.

Singer 212 - Alto Within the setting it took a while to get used to being in an artificial acoustic but after not long I forgot that I wasn't actually in the acoustic space and enjoyed each of them.

Singer 213 - Alto Apart from becoming more comfortable with the surroundings as the process developed (I don't often sing surrounded by curtains - but very relaxing experience!)...

Singer 221 - Tenor I think the space definitely effects how well pitch is maintained during a piece. It also has effects on interpretation especially in terms of tempos which are possible without loosing a sense of the music in the space. Space *SP* made singing quite hard work because there was little feedback, while *LC* and *MR* gave support and something to respond to, which made my voice relax more.

Singer 231 - Bass Yes.

Singer 232 - Bass The setting A made me feel in a quite dry acoustic, so I had to "search" my voice, but it might be better to find technically the right way to produce this or this passage in a song, since we are not biased by the reverberation of a room. So, to sum up, I enjoyed less my sound, but would like this to work technically. Setting *SP* felt quite comfortable although here, I had a tendency to enjoy my sound, the dark sound reverberation, so I'm pretty sure I tended to push a bit on my voice.

Setting *LC* would then be a bit "over the top" : tending to push a bit because of the acoustic. I would add that even if reverberation of setting *SP* and *LC* are felt like such, it felt that the reverberating sound would envelop us more in a natural environment, and the natural church acoustic usually gives a perception of being around the altar, i.e. not too large distance behind, on the left and on the right between us and the walls, but very large distance (and therefore different feeling for the reverberation) in front, in the direction where the voice is projected.

Q. Do you think the Virtual Singing Studio can be improved? If so - how?

Singer 201 - Soprano possible visual element would be interesting. Would help with immersion in virtual environment

Singer 211 - Alto expand to accommodate a group of singers. Address tinny quality.

Singer 212 - Alto yes - loudness of reverb? Don't know

Singer 213 - Alto no - it was much more realistic than I had thought it would be (I was surprised by the acoustic of the real venue!)

Singer 221 - Tenor addition of a visual element

Singer 231 - Bass re the above as a visual species even the teen musicians - something to see that relates to the space might have an effect

Singer 232 - Bass matches well, except for bubble effect! Real space close to vss except for lower register, in NCEM - trust sound in VSS - need to make sound, feel need to push, in VSS feel less support - lower registers most different

Appendix G

Lyrics of recorded fragments

Test 232a

Lyrics "Que le pardon, et la clemence"

Piece "Si la riguer" Aria from the opera La Juive by Halévy

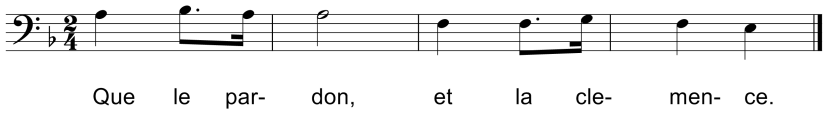


Figure G.1: *Test 232a fragment*

Test 232b

Voice Bass

Lyrics "Why, Why has thou robbed me of my rest"

Piece from "Saul" by Henry Purcell

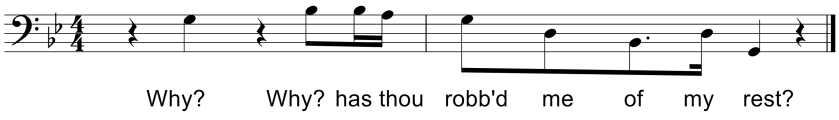


Figure G.2: *Test 232b fragment*

Test 221

Voice Tenor

Lyrics "de domo pulsus regali" , French medieval

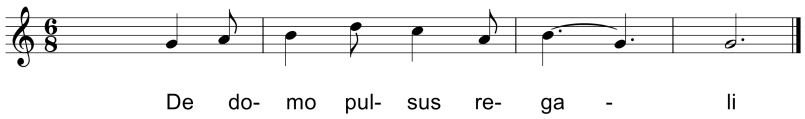


Figure G.3: *Test 221 fragment*

Test 212

Voice Mezzo-soprano

Lyrics "Down a down a down hey down" from "The Three Ravens" an English folk ballad

Verses

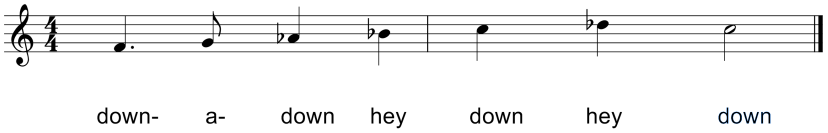


Figure G.4: *Test 212 fragment*

Lines in italic refer to the fragments used in the listening test.

Verse A:
There were three ravens sat on a tree
Down-a-down, hey down, hey,down
They were as black as they might be,
With a down
The one of them said to his mate
"What shall we for our breakfast take?"
With a down, derry, derry derry down, down

Verse B
Down in yonder green field
Down-a-down, hey down, hey,down
There lies a knight slain under his shield,
With a down
Down there comes a fallow doe
As great with young as she might go
With a down, derry, derry derry down, down

Verse C
His hawks they fly so eagerly

Down-a-down, hey down, hey,down
No other fowl dare him come nigh,
With a down
Down there comes a fallow doe
As heavy with young as she might go.
With a down, derry, derry derry down, down

Verse D
She lifted up his bloody head
Down-a-down, hey down, hey,down
And kissed his wounds that were so red
With a down
She got him up across her back
And carried him to the earthen lack.
With a down, derry, derry derry down, down

Verse E
She buried him before his prime
Down-a-down, hey down, hey,down
She was dead herself before even time
With a down
God send every gentlemen
Fine hawks, fine hounds and such a loved one.
With a down, derry, derry derry down, down

Appendix H

Comments on fragments by listeners

TEST 232a	LCRealV1	LCVirutalV1	MRVirtualV1	SPRealV1	SPVirtualV1
1	powerful	warmer, more saturated sound	powerful	warmer, more saturated sound	powerful
2					
3	Powerful first half, second half more gentle	Powerful first half, second half more gentle	Second half not as quiet as others	Second half not as quiet as others	Powerful first half, second half more gentle
4	bright, nasal vowels	darker vowels	darker vowels	bright, nasal vowels	bright, nasal vowels
5	All examples sound the same - No outstanding features (Except the singer is very good/beautiful tone.)	More Dramatic performance	All examples sound the same - No outstanding features (Except the singer is very good/beautiful tone.)	All examples sound the same - No outstanding features (Except the singer is very good/beautiful tone.)	All examples sound the same - No outstanding features (Except the singer is very good/beautiful tone.)
6	Contrast between delivery of phrases (theatrical/cautious)	Contrast between delivery of phrases (theatrical/cautious)	More uniform delivery	More uniform delivery	More uniform delivery
7	forcing on the 1st part.more time before the second part of the phrase, which is sung softer)		Forcing a bit on the 1st phrase, then comforting the medium-low register with the acoustic	Forcing a bit on the 1st phrase, then comforting the medium-low register with the acoustic	comforting the medium-low register with the acoustic, for both parts of the phrase)
8			no comments		
9			no comments		
10			no comments		

Test 232b	LCRealV1	LCVirtualV1	MRVirtuaV1	MRVirtuaV2	MRVirtualV3	SPRealV1	SPRealV2	SPVirtualV1
1	More saturated	average, slower	average, slower	Soften 2nd sentence	softer,	softer,	More saturated	Breathing later on
2	no comments							
3	most similar, 'rest' sounds like 'hest	most similar, 'rest' sounds like 'hest	most similar, 'rest' sounds like 'hest	most similar, 'rest' sounds like 'hest	most similar, 'rest' sounds like 'hest	'w' at beginning more aspirated than others, 'r' less like 'h'	most similar, 'rest' sounds like 'hest	most similar, 'rest' sounds like 'hest
4	no coments							
5	The pitching/intonation seems to be slightly different in every example, but otherwise no outstanding features.							
6	no comments							
7	The first "why" is more vocative	sounds a bit brighter, though the timbre remains homogenic on the whole phrase, but we feel it needs more effort when the phrase goes low	sounds a bit brighter, though the timbre remains homogenic on the whole phrase, but we feel it needs more effort when the phrase goes low			Deeper timbre, all over the phrase, comforted by the acoustic	The first "why" is more vocative	sounds a bit brighter, though the timbre remains homogenic on the whole phrase, but we feel it needs more effort when the phrase goes low
8	no comments							
9	no comments							
10	no comments							
11	This performance is very dramatic. Resembles the performance of an actor in theatre	This performance is very dramatic. Resembles the performance of an actor in theatre		This performance resembles more the performance pop singers have nowadays.		This performance resembles more the performance pop singers have nowadays.	This performance resembles more the performance pop singers have nowadays.	

Test 221	lc_real_v1	lc_real_v2	lc_real_v3	lc_vss_v1	lc_vss_v2	lc_vss_v3
1	Soft and continuous	Strong, powerful	Average group, not too continuous, normal breathing.	Average group, not too continuous, normal breathing.	Close to lc_real_v1, lots of vibrato later	Discontinuos, breathing in middle
2	a bit more legato	very legato	more legato		more legato	
3						
4	less vib	more vib	more vib	more vib	more vib	more vib
5	Vibratoless final note	Disruptive Rubato	Some Rubato	Occasional Dodgy notes	Some Rubato	No outstanding features
6						NO COMMENTS
7	Intermediate tempo, more englobing sound, sounds a bit more legato than the faster group	Slower tempo / more rubato / but the intonation of the syllable 'si's of the word 'pulsis' then gets sharper (Slow tempo, although the enlgobing acoustic is quite missing in comparision with the 2 others of the group)	Intermediate tempo, more englobing sound, sounds a bit more legato than the faster group	Faster tempo	Intermediate tempo, more englobing sound, sounds a bit more legato than the faster group	Faster tempo (phrase a bit shorter, less rrubato, less playing with the acoustic)
8						no comments
9						no comments
10	similar to lc_real_v1 but less bright sound	out of tune	similar to lc_real_v1 but less bright sound	similar to lc_real_v1 but less bright sound	nc	out of tune
11	The steady ending of this performance reminds me of church hymns	nc	Opera-like: similarilty to lc_real_v1	Soft-Tender	extended finish	extended finish
13	very dry performance.	Vibrato or tremolo applied in performance	slower pace - less performance characteristics	more vibrato than other group - tendancy to drift sharp	more vibrato than other group - tendancy to drift sharp	sharp note - out of tune - vibrato applied so similar to the group to the left
14	all at the speed and similar frequency	speed is lower than the others, and lower frequency	They are different in the middle	all at the speed and similar frequency	all at the speed and similar frequency	very short in the middle
15	greater emphasis on 'per sis' (slows the consonants down?)	Slower group	faster tempo - height of the placement = pitching of last note	faster tempo - height of the placement = pitching of last note	faster tempo - height of the placement = pitching of last note, insecure last note	faster tempo - height of the placement = pitching of last note, insecure last note
16						no comments
17	Direct, clean interpretation	More free and operatic	Dotted / pointed / unequal rhythm	Direct, clean interpretation	Dotted / pointed / unequal rhythm	Dotted / pointed / unequal rhythm
18	longer phrases - showing more performing skills	longer phrases - showing more performing skills	no long phrases - less performing skills	no long phrases - less performing skills	longer phrases - showing more performing skills	staccatto at the end of the first phrase - seperate the two phrases with a breath
19						NO COMMENTS

212	LCvirtualA1	LCvirtualA2	LCvirtualA5	LCvirtualB3	LCvirtualB4
1	Average	Average	Fast attack, breathing. discontinuous	Fast attack, breathing. discontinuous	Average
2	no comments				
3				Very pronounced Ds; Different tuning to others	
4	more rich tone, more vib		richer tone	more rich tone, more vib	more rich tone, more vib
5		some rubato	wiffy tuning	wiffy tuning	no outstanding features
6	NO COMMENTS				
7	1st syllable, then phrase with a direction towards the 3rd last syllable / not too much vibrato / an "i-brightness" in the sound	more of a"o-brightness"	1st syllable, then phrase with a direction towards the 3rd last syllable / not too much vibrato / an "i-brightness" in the sound	more of a"o-brightness"	direction towards the 3rd last syllable, without any coma between 1st and 2nd syllable / more intimate / more sad / slower tempo
8	NO COMMENTS				
9	NO COMMENTS				
10	not bright			sound out of tune	
11			Unwilling to accept the incident that hurt her. It seems to me that even without an effort from the other party she excuses it	Anxious	She was wrong. She has acted in a very determined way in the past and now she realised she has to restore the balance
12	nO COMMENTS				
1	Average	Average	Fast attack, breathing. discontinuous	Fast attack, breathing. discontinuous	Average
13	Less acceleration in tempo towards end of phrase	Softer performance - change in tempo less apparent	Least use of vibrato out of all excerpts	Compared with other excerpts - this one is flat and drifts even flatter towrds the end of the phrase	From left to right - increase in volume range and perceived acceleration in pace towards the end of phrase: level 6
14	NA				
15	Husky	Husky	faster group brighter vowels (especially final)	SLow Husky voice quality (needs to clear throat?)	SLow
16					
17	Gentle and swoopy	Gentle and soft	Gentle and soft	Flat	Gentle and soft
18	more allegro	more allegro	more allegro rit. at the end	more allegro with accent at the first syllable	soft and dolce and dramatic, softer than sp_vss_v4
19	NO COMMENTS				

Appendix I

Goodness of fit of MDS solutions in Chapter 6

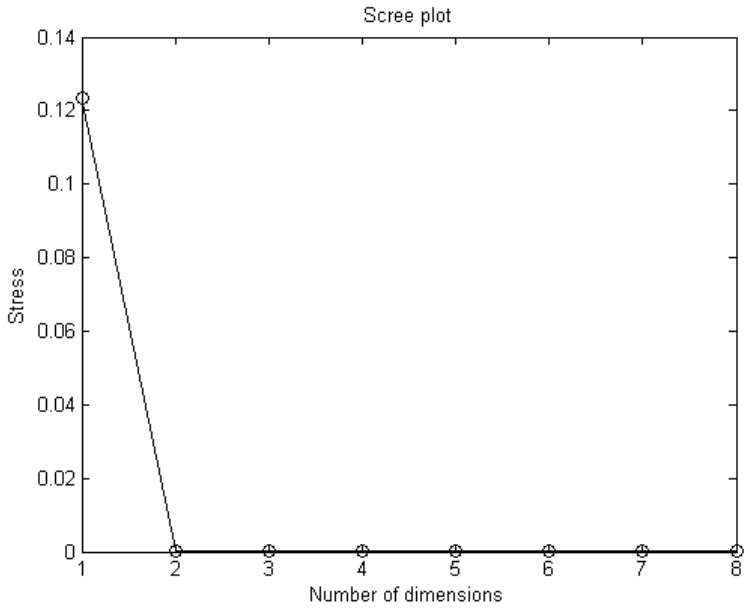


Figure I.1: *Scree plot of stress measure with increasing dimensions used to model test 232a dissimilarity data*

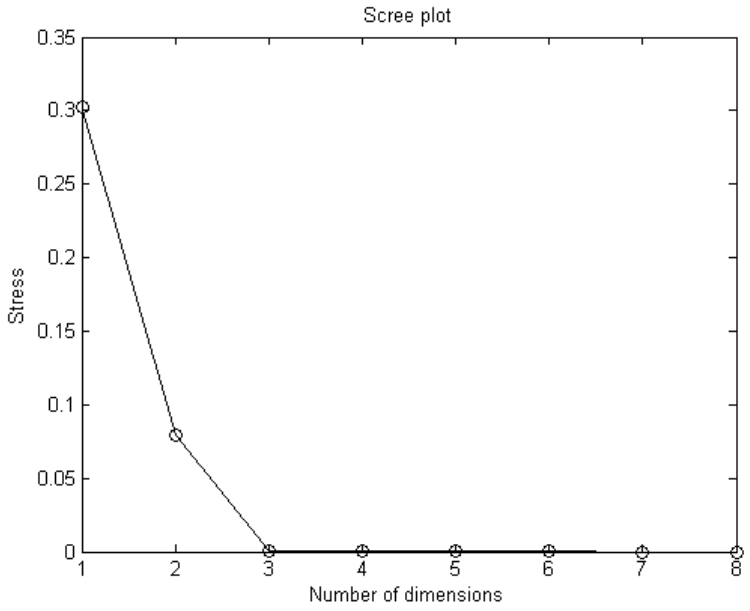


Figure I.2: *Scree plot of stress measure with increasing dimensions used to model test 232b dissimilarity data*

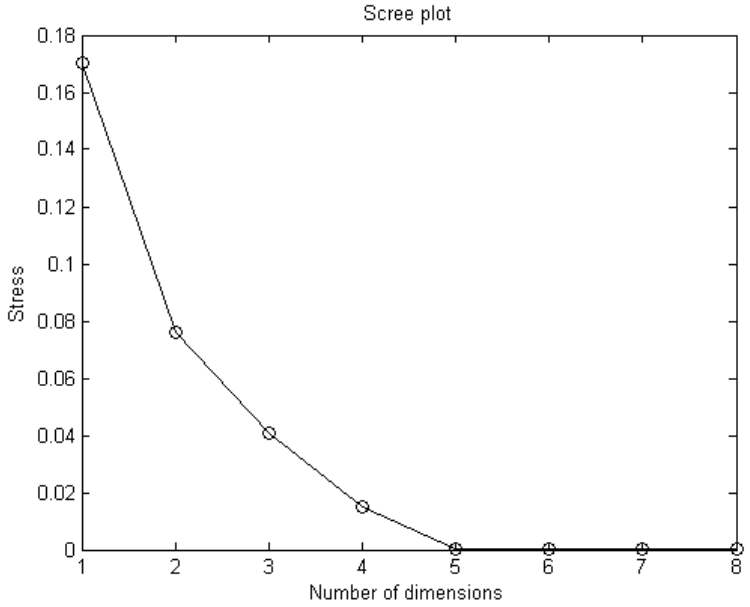


Figure I.3: *Scree plot of stress measure with increasing dimensions used to model test 221 dissimilarity data*

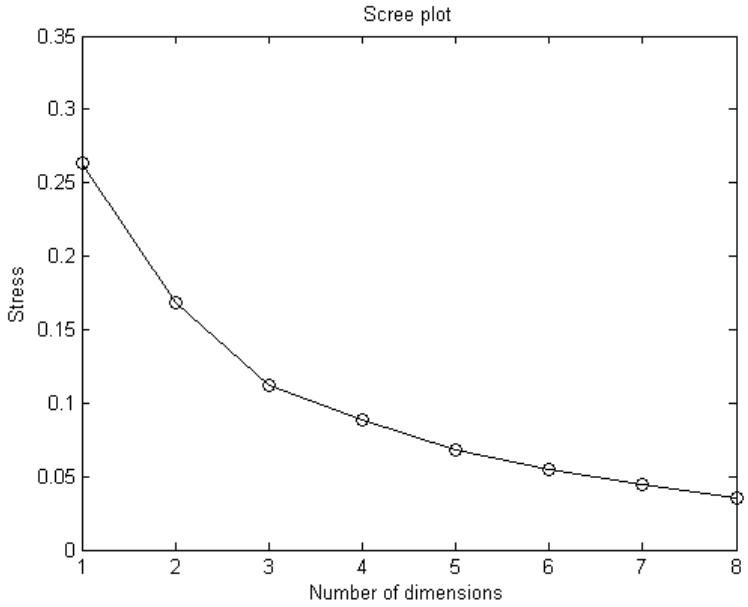


Figure I.4: *Scree plot of stress measure with increasing dimensions used to model test 212 dissimilarity data*

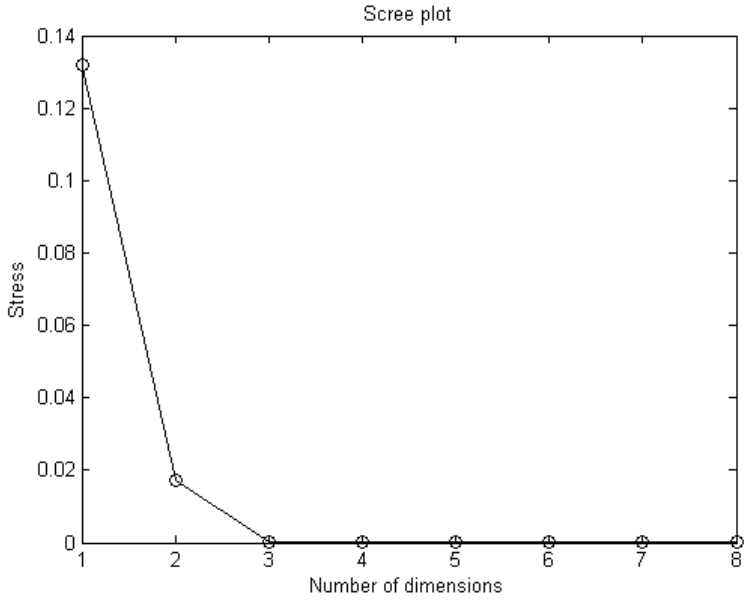
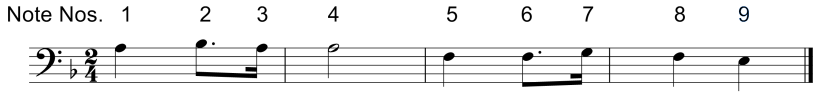


Figure I.5: *Scree plot of stress measure with increasing dimensions used to model test 212 Verse B dissimilarity data*

Appendix J

Vibrato Analysis

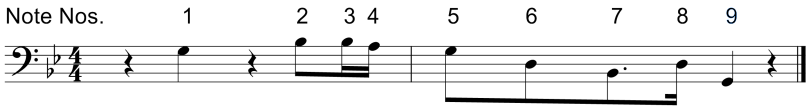
Test 232a



Vibrato Rate (Hz)	***Note number***									
Name of Fragment	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	Mean (Std. dev)
'LCrealv1'	5.8	6.0	13.8	5.8	6.5	7.0	10.1	5.9	6.1	7.1 (2.5)
'LCvirtualV1'	5.6	8.9	7.5	6.9	11.0	6.9	12.9	5.9	7.0	7.5 (1.7)
'MRvirtualV1'	5.6	6.8	13.6	5.6	6.5	8.0	9.4	6.9	7.9	6.5 (0.9)
'SPrealV1'	5.5	6.3	11.0	5.4	7.9	10.6	11.0	6.4	5.5	6.8 (1.7)
'SPvirtualV1'	6.2	6.7	13.2	6.4	6.5	8.9	9.7	6.7	7.7	6.7 (0.5)
Vibrato Extent (cents)	***Note number***									
Name of Fragment	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	Mean (Std. dev)
'LCrealv1'	23.9	47.3	28.1	130.4	53.3	28.0	20.2	102.6	50.5	58.0 (36)
'LCvirtualV1'	78.6	30.4	16.7	152.7	76.3	25.0	17.1	32.1	30.5	58.8 944.7)
'MRvirtualV1'	38.4	28.9	12.0	190.3	127.4	4.7	16.1	58.1	57.0	94.2 (56.8)
'SPrealV1'	64.6	27.9	17.7	101.3	75.3	8.1	10.8	86.6	49.8	59.1 (30.5)
'SPvirtualV1'	46.5	20.1	18.7	313.4	101.5	36.7	35.6	63.0	45.6	114.0 9101.7)

Table J.1: *Vibrato Rate (Hz) and Vibrato Extent (Cents) for notes of phrase Test 232a; Mean values include only notes in the phrase with 4 vibrato cycles or more*

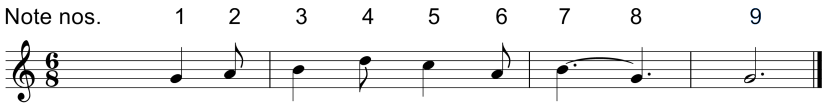
Test232b



Vibrato Rate (Hz)	Note number									
Name of Fragment	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	**Mean (Std Dev)**
'LCrealV1'	6.0	8.6	16.0	7.6	7.8	9.6	7.3	5.9	7.4	7.1 (0.8)
'LCvirtualV1'	6.0	12.5	12.0	9.3	8.1	7.1	7.0	9.5	5.6	8.5 (2.3)
'MRvirtualV1'	5.9	6.8	10.8	13.5	7.6	6.9	7.7	4.9	6.5	6.7 (0.7)
'MRvirtualV2'	5.8	8.6	11.2	10.7	7.7	9.9	8.0	7.0	7.5	7.2 (1.0)
'MRvirtualV3'	5.6	13.1	24.9	14.5	6.7	7.6	9.4	7.0	7.5	9.3 (3.7)
'SPrealV1'	6.2	9.2	14.7	10.2	7.3	9.2	8.5	5.1	6.8	8.0 (1.3)
'SPrealV2'	*11.4*	*7.5*	*7.9*	*11.7*	*7.0*	*8.3*	*9.4*	*7.0*	*7.2*	9.3 (2.1)
'SPvirtualV1'	6.2	11.0	12.7	7.0	8.8	9.9	6.4	10.3	6.7	8.6 (1.9)
Vibrato Extent (Cents)	**Note number**									
Name of Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Mean (Std. Dev)**
'LCrealV1'	110.1	6.8	2.0	47.2	152.2	84.8	359.1	51.2	95.5	119.3 (24)
'LCvirtualV1'	185.1	39.5	10.7	45.5	33.2	88.8	188.3	91.1	57.5	75.3 (52.6)
'MRvirtualV1'	143.3	4.5	18.3	25.7	77.1	24.2	82.6	22.2	91.0	103.8 (28.5)
'MRvirtualV2'	74.5	10.1	8.5	3.2	37.9	13.0	126.8	60.3	86.6	57.9 (29.1)
'MRvirtualV3'	107.6	4.4	4.6	17.1	39.8	51.8	462.3	16.7	89.9	56.0 (51.6)
'SPrealV1'	89.1	3.9	2.1	135.6	46.7	20.0	426.7	13.1	51.2	39.9 (32.3)
'SPrealV2'	183.4	11.5	15.4	11.0	33.1	5.0	105.0	21.6	86.4	134.9 (48.5)
'SPvirtualV1'	104.6	6.1	3.4	13.3	33.1	14.9	364.9	13.6	72.2	45.9 (37.2)

Table J.2: *Vibrato Rate (Hz) and Vibrato Extent (Cents) for each note of phrase in Test 232b; Mean values include only notes in the phrase with 4 vibrato cycles or more*

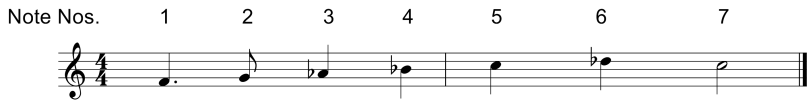
Test221



Vibrato Rate (Hz)	Note number									
Name of Fragment	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	**Mean (Std. Dev)**
'LCrealV1'	7.4	8.4	7.8	12.0	14.7	9.3	8.5	5.8	5.8	8.7 (3.6)
'LCrealV2'	7.3	9.5	5.3	13.5	11.0	13.6	5.0	6.7	5.4	7.6 (3.4)
'LCrealV3'	10.8	9.7	5.7	10.5	9.5	15.7	7.3	5.7	5.6	7.9 (2.1)
'LCvirtualV1'	10.1	8.4	7.7	9.8	9.2	12.7	4.9	7.2	5.0	5.7 (1.1)
'LCvirtualV2'	7.3	8.1	6.6	12.2	9.8	11.3	7.4	5.0	5.9	6.1(1.0)
'LCvirtualV3'	9.0	8.8	5.0	17.1	10.2	14.8	6.7	7.2	5.8	8.6 (3.6)
'MRvirtualV1'	9.8	9.2	7.7	15.8	13.8	11.7	7.3	6.8	6.2	6.8 (0.5)
'SPrealV1'	6.5	9.2	6.4	7.5	8.4	14.5	7.4	6.4	5.1	6.3 (1.0)
'SPrealV2'	8.1	7.1	6.4	10.7	6.4	9.8	4.5	6.0	5.6	5.8 (0.2)
'SPrealV3'	8.0	10.3	8.6	10.2	12.9	15.0	7.0	5.6	6.8	6.5 (0.6)
'SPvirtualV1'	10.5	9.6	7.4	11.2	10.8	11.5	7.4	7.3	5.8	8.2 (2.4)
Vibrato Extent (cents)	**Note number**									
Name of Fragment	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	**Mean (Std. Dev)**
'LCrealV1'	27.2	6.9	97.0	18.5	21.7	61.9	19.2	48.9	53.8	35.9 (15.5)
'LCrealV2'	41.2	20.4	80.9	38.6	30.2	27.8	13.3	38.7	168.2	64.7 (60.7)
'LCrealV3'	27.1	21.9	43.4	46.1	103.8	20.9	21.1	49.2	140.8	63.0 (55.1)
'LCvirtualV1'	23.8	51.7	55.3	28.3	31.8	16.6	22.8	80.4	203.2	102.1 (75.2)
'LCvirtualV2'	40.3	38.1	74.9	34.2	41.5	47.3	39.0	62.2	231.7	111.0 (85.9)
'LCvirtualV3'	59.0	27.0	35.7	20.5	58.6	51.9	39.9	60.9	211.3	91.0 (69.9)
'MRvirtualV1'	27.3	7.4	35.3	20.8	17.3	31.1	17.1	75.2	122.2	71.5 (43.0)
'SPrealV1'	24.9	17.2	134.1	38.2	35.2	27.3	15.4	26.9	82.2	41.5 (29.1)
'SPrealV2'	34.3	55.9	30.7	67.5	241.1	35.1	28.7	71.0	265.1	168.0 (97.0)
'SPrealV3'	54.3	21.8	36.1	39.8	60.1	50.5	30.3	55.6	136.4	74.1 (45.2)
'SPvirtualV1'	25.3	35.6	65.4	34.4	50.8	84.4	33.8	58.2	328.5	148.9 (128.7)

Table J.3: *Vibrato Rate (Hz) and Vibrato Extent (Cents) for notes of phrase Test 221; Mean values include only notes in the phrase with 4 vibrato cycles or more*

Test212



Vibrato Rate (Hz)	Note Number							
Name of Fragment	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***Mean (Std. Dev)***
LCrealA4	7.1	12.4	7.9	11.1	8.3	6.1	6.0	6.55 (0.5)
LCrealC6	8.2	10.1	9.1	9.4	9.6	5.0	6.8	8.23 (1.2)
LCrealD7	7.4	7.0	8.6	6.7	9.4	8.5	6.8	8.02 (1.0)
LCrealD8	6.8	10.8	9.6	6.9	8.8	6.0	6.5	7.93 (1.3)
LCrealE9	6.9	5.7	6.2	6.6	4.6	5.3	6.7	6.84 (0.1)
LCvirtA1	6.0	12.2	7.5	7.5	6.9	6.1	6.8	6.41 (0.4)
LCvirtA2	7.5	12.6	6.9	4.8	6.1	6.0	6.7	7.11 (0.4)
LCvirtA5	6.5	12.2	13.0	4.9	5.5	5.5	6.8	8.79 (3.0)
LCvirtB3	7.0	9.3	6.7	9.6	7.0	6.0	8.0	7.54 (0.5)
LCvirtB4	6.9	8.9	7.9	4.9	8.3	4.5	6.8	7.32 (0.7)
LCvirtB6	7.0	12.5	6.7	5.4	6.1	8.0	7.2	7.38 (0.4)
LCvirtC7	7.8	12.3	7.0	10.5	5.8	5.5	6.0	6.89 (0.9)
LCvirtD8	7.6	11.2	9.6	9.5	6.0	10.1	6.9	8.56 (1.4)
LCvirtE10	7.8	7.4	8.9	5.2	7.4	6.4	6.8	7.47 (0.9)
LCvirtE9	7.2	12.1	15.0	5.5	6.1	4.9	7.1	10.36 (3.4)
MRrealA1	6.7	7.0	6.8	6.9	5.9	5.1	7.0	6.86 (0.1)
MRvirtA1	6.5	8.2	12.5	4.7	6.7	7.2	7.3	8.36 (2.4)
MRvirtA2	6.5	12.0	10.9	4.8	5.8	7.2	7.0	6.90 (0.3)
MRvirtA3	8.0	8.9	9.9	7.6	5.4	5.5	7.4	7.70 (0.3)
MRvirtA4	6.6	9.1	7.6	8.8	7.7	7.7	7.5	7.05 (0.5)
MRvirtB5	7.7	12.3	10.1	5.6	5.7	6.0	6.3	7.74 (0.0)
MRvirtB6	7.1	9.9	6.6	5.1	6.4	4.8	7.2	7.14 (0.1)
MRvirtB7	7.9	8.7	8.3	5.0	5.6	8.7	7.0	7.86 (0.7)
MRvirtC8	8.1	9.4	8.2	7.3	7.5	5.1	7.6	7.86 (0.3)
MRvirtD9	7.4	11.3	10.2	6.9	5.1	4.5	6.1	6.74 (0.7)
MRvirtE10	7.5	7.6	5.6	7.5	5.6	6.7	6.7	6.96 (0.4)
MRvirtE11	7.9	10.9	8.5	10.4	5.2	7.9	7.2	8.36 (1.2)
SPrealB2	6.8	9.4	9.0	5.2	6.6	4.2	7.8	7.29 (0.5)
SPrealC3	7.5	10.0	8.8	11.6	5.7	5.5	6.4	6.94 (0.5)
SPvirtA1	6.9	11.7	8.5	5.3	6.2	5.2	7.9	7.44 (0.5)
SPvirtA2	7.4	10.1	8.2	7.6	5.4	8.7	6.0	7.36 (1.1)
SPvirtB3	7.1	12.9	5.3	7.0	5.5	5.4	7.8	7.45 (0.4)
SPvirtC4	7.0	7.1	7.4	5.2	7.4	5.0	7.5	7.26 (0.2)
SPvirtD5	7.1	11.1	7.9	7.1	10.0	9.6	7.4	8.39 (1.2)

Table J.4: *Vibrato Rate (Hz) for each note of phrase in Test 212; Mean values include only notes in the phrase with 4 vibrato cycles or more*

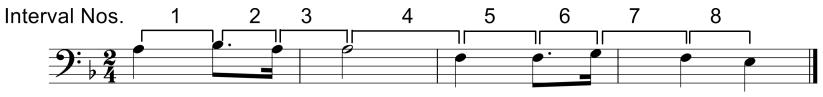
Vibrato Extent	**Note Number**							
Name of Fragment	***1***	***2***	***3***	***4***	***5***	***6***	***7***	**Mean (St Dev)**
'LCrealA4'	74.5	16.0	10.0	272.7	11.0	26.8	76.4	75.5 (1.0)
'LCrealC6'	115.8	61.5	95.3	204.7	35.4	40.1	119.3	90.2 (38.7)
'LCrealD7'	60.4	16.2	21.1	54.6	13.9	10.9	107.0	48.0 (39.3)
'LCrealD8'	68.2	14.6	27.9	63.7	9.9	22.0	106.7	53.2 (37.4)
'LCrealE9'	46.8	9.9	8.5	31.9	15.4	20.1	74.8	60.8 (14.0)
'LCvirtA1'	44.7	16.3	20.5	112.0	25.9	22.8	98.7	71.7 (27).0
'LCvirtA2'	89.9	18.2	32.9	84.7	18.8	24.0	88.5	89.2 (0.7)
'LCvirtA5'	51.3	16.7	12.5	89.6	22.2	17.2	107.8	57.2 (39.1)
'LCvirtB3'	96.1	11.4	29.7	84.8	24.2	19.2	85.2	90.6 (5.5)
'LCvirtB4'	79.7	12.2	13.1	48.6	22.5	22.0	76.2	59.5 (26.2)
'LCvirtB6'	48.4	19.0	30.2	71.2	19.8	62.7	89.8	67.0 (17.1)
'LCvirtC7'	60.4	21.6	41.8	116.0	27.9	24.4	114.4	87.4 (27.0)
'LCvirtD8'	74.7	30.0	34.2	72.4	28.0	51.3	100.0	65.1 (24.8)
'LCvirtE10'	39.7	16.3	11.7	19.4	20.1	39.0	74.1	36.9 (21.5)
'LCvirtE9'	32.3	35.5	6.8	30.9	20.7	13.1	159.2	58.5 (59.2)
'MRrealA1'	129.4	15.2	19.8	91.8	38.5	35.8	114.9	122.1 (7.3)
'MRvirtA1'	92.4	22.5	7.8	74.9	20.1	22.4	141.6	66.1 (54.1)
'MRvirtA2'	73.0	18.1	8.5	69.8	19.2	21.8	82.6	59.1 (26.7)
'MRvirtA3'	85.5	30.4	14.0	121.3	22.9	25.4	47.3	66.4 (19.1)
'MRvirtA4'	115.7	33.0	26.2	193.1	35.0	13.1	116.9	116.3 (0.6)
'MRvirtB5'	79.9	26.3	14.5	97.5	21.0	16.7	93.3	79.9 (0.0)
'MRvirtB6'	71.3	28.5	32.5	54.2	44.1	23.0	142.0	106.7 (35.4)
'MRvirtB7'	102.0	38.3	48.8	54.0	26.0	23.6	100.6	75.4 (36.7)
'MRvirtC8'	118.7	42.8	58.0	161.5	100.3	16.6	152.2	107.3 (34)
'MRvirtD9'	81.5	17.5	16.8	97.4	21.8	20.6	83.7	82.6 (1.1)
'MRvirtE10'	63.8	13.8	32.1	63.0	17.1	36.6	63.5	54.6 (12.7)
'MRvirtE11'	48.2	11.5	21.3	47.8	20.6	17.1	112.0	56.3 (34.6)
'SPrealB2'	80.7	34.6	18.3	59.0	32.8	22.6	93.6	87.1 (6.5)
'SPrealC3'	93.5	24.5	42.7	73.4	26.3	16.4	129.9	111.7 (18.2)
'SPvirtA1'	63.6	23.9	29.6	92.1	15.9	19.9	66.3	65.0 (1.4)
'SPvirtA2'	89.4	23.1	31.9	54.0	27.0	10.4	101.7	67.2 (40.4)
'SPvirtB3'	120.3	23.4	11.7	69.8	27.5	22.5	125.4	122.8 (2.6)
'SPvirtC4'	68.6	12.5	38.2	83.6	18.7	29.3	77.2	72.9 (4.3)
'SPvirtD5'	50.2	20.0	32.6	81.7	21.5	8.4	106.5	43.9 (34.2)

Table J.5: *Vibrato Extent (Cents) for each note of phrase in Test 212; Mean values include only notes in the phrase with 4 vibrato cycles or more*

Appendix K

Intonation Metrics

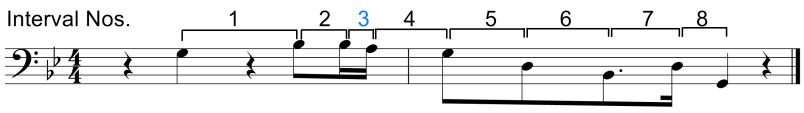
Test 232a



Name of Fragment	**MAIE**	**MAPP**	**MAIP**
'LCrealv1'	21.9	20.4	30.6
'LCvirtualV1'	44.3	16.2	26.4
'MRvirtualV1'	49.6	21.2	33.4
'SPrealV1'	24.8	17.2	9.5
'SPvirtualV1'	83.1	41.8	62.6
Mean (StDev)	**44.7 (24.6)**	**23.4 (10.5)**	**32.5 (19.2)**

Table K.1: *Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 232a*

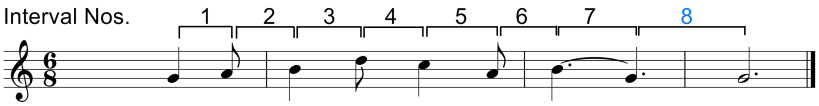
Test232b



Name of Fragment	MAIE	MAPP	MAIP
'LCrealV1'	83.5	52.6	59.8
'LCvirtualV1'	45.9	28.7	47.6
'MRvirtualV1'	37.2	18.9	36.4
'MRvirtualV2'	46.6	19.7	35.5
'MRvirtualV3'	67.3	25.0	39.8
'SPrealV1'	50.0	34.0	49.4
'SPrealV2'	100.9	44.6	74.3
'SPvirtualV1'	30.3	13.7	24.5
Mean (StDev)	**57.7 (24.3)**	**29.7 (13.4)**	**45.9 (15.6)**

Table K.2: *Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 232b*

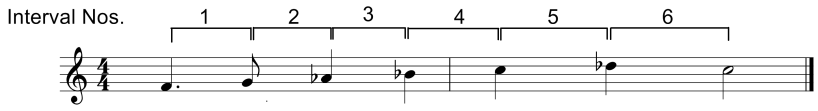
Test221



Name of Fragment	**MAIE**	**MAPP**	**MAIP**
'LCrealV1'	23.43	10.45	14.40
'LCrealV2'	22.0	15.3	10.9
'LCrealV3'	23.5	10.0	18.3
'LCvirtualV1'	26.6	18.4	23.3
'LCvirtualV2'	23.4	28.0	18.4
'LCvirtualV3'	35.7	19.5	22.3
'MRvirtualV1'	19.9	13.7	16.4
'SPrealV1'	11.4	18.1	18.2
'SPrealV2'	19.4	9.5	15.9
'SPrealV3'	22.6	24.4	13.9
'SPvirtualV1'	38.7	19.8	36.8
Mean (StDev)	**24.2 (7.5)**	**17.0 (6)**	**19.0 (6.9)**

Table K.3: *Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 221*

Test212



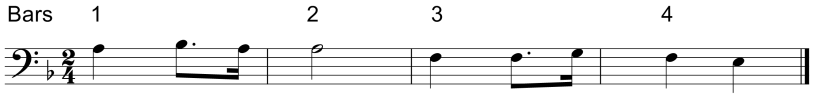
Name of Fragment	MAIE	MAPP	MAIP
'LCrealA4'	181.4	17.6	22.6
'LCrealC6'	247.9	18.7	32.3
'LCrealD7'	14.6	28.3	10.5
'LCrealD8'	16.3	60.2	28.7
'LCrealE9'	25.7	27.5	32.0
'LCvirtA1'	17.2	14.8	18.7
'LCvirtA2'	18.0	37.3	33.0
'LCvirtA5'	10.0	11.6	23.4
'LCvirtB3'	32.0	34.9	13.9
'LCvirtB4'	27.2	18.6	31.7
'LCvirtB6'	16.1	20.6	20.6
'LCvirtC7'	28.6	21.5	26.4
'LCvirtD8'	18.3	17.7	22.4
'LCvirtE10'	15.1	42.1	80.3
'LCvirtE9'	36.2	20.6	23.3
'MRrealA1'	22.0	15.5	12.4
'MRvirtA1'	18.9	22.4	13.1
'MRvirtA2'	17.5	40.0	63.5
'MRvirtA3'	18.9	35.2	23.7
'MRvirtA4'	80.4	17.0	31.6
'MRvirtB5'	21.8	22.8	25.6
'MRvirtB6'	11.2	18.2	29.3
'MRvirtB7'	12.5	79.8	171.7
'MRvirtC8'	64.3	20.0	33.5
'MRvirtD9'	30.9	20.0	17.5
'MRvirtE10'	27.6	120.3	236.6
'MRvirtE11'	24.4	14.2	9.9
'SPrealB2'	24.4	20.7	13.8
'SPrealC3'	16.5	16.8	18.1
'SPvirtA1'	37.2	22.8	37.5
'SPvirtA2'	25.2	33.6	38.1
'SPvirtB3'	23.8	15.2	28.1
'SPvirtC4'	10.0	13.2	16.4
'SPvirtD5'	10.4	30.2	19.6
mean (StDev)	**35.4 (48.4)**	**28.5 (21.4)**	**37.1 (45.3)**

Table K.4: *Mean Absolute Interval Error (MAIE), Mean Absolute Pitch and Interval Precision (MAPP and MAIP) measured in cents for Test 212*

Appendix L

Tempo Analysis

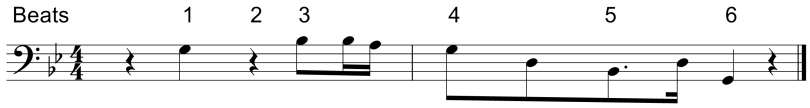
Test 232a



Tempo (bpm)	Beat Number				
Name of Fragment	***1***	***2***	***3***	***4***	***Global Tempo (SD)***
'LCrealv1'	46.8	49.9	44.0	55.5	48.7 (5.0)
'LCvirtualV1'	42.0	67.6	46.4	60.5	52.1 (12.0)
'MRvirtualV1'	56.9	52.6	53.4	58.6	55.3 (2.9)
'SPrealV1'	50.6	51.2	52.5	54.7	52.2 (1.8)
'SPvirtualV1'	58.5	43.3	49.1	63.3	53.5 (9.0)

Table L.1: *Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 232a*

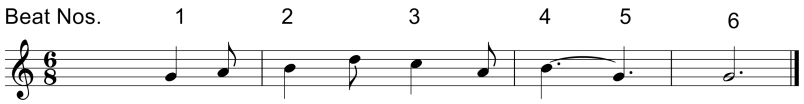
Test 232b



Tempo (bpm)	**Beat number**						
Name of Fragment	***1***	***2***	***3***	***4***	***5***	6	Global Tempo (SD)
'LCrealV1'	39.37	95.64	49.89	35.29	55.28	74.78	58.38 (24.04)
'LCvirtualV1'	36.52	68.42	37.16	55.20	47.87	42.76	47.99 (13.35)
'MRvirtualV1'	36.47	75.95	42.67	46.01	52.17	47.24	50.09 (15.24)
'MRvirtualV2'	37.43	60.79	39.63	42.99	48.52	46.82	46.03 (9.33)
'MRvirtualV3'	39.16	102.74	45.22	43.31	54.45	52.91	56.30 (26.19)
'SPrealV1'	32.33	96.93	42.37	39.47	54.55	46.58	52.04 (25.77)
'SPrealV2'	30.96	66.89	45.45	45.98	55.56	40.16	47.50 (13.32)
'SPvirtualV1'	33.84	76.24	42.25	49.59	51.02	46.37	49.88 (15.88)

Table L.2: *Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 232b*

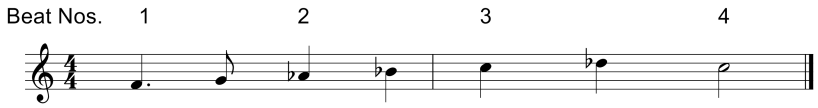
Test 221



Tempo (bpm)	**Beat number**						
Name of Fragment	***1***	***2***	***3***	***4***	***5***	6	Global Tempo (SD)
'LCrealV1'	70.23	60.36	56.59	54.60	57.78	63.45	60.50 (5.6)
'LCrealV2'	60.79	48.87	61.94	51.92	80.61	49.99	59.02 (12.0)
'LCrealV3'	70.70	60.42	64.86	53.89	85.03	61.64	66.09 (10.8)
'LCvirtualV1'	67.77	78.77	60.02	58.98	107.59	75.33	74.74 (18.0)
'LCvirtualV2'	64.03	77.25	61.50	55.00	60.38	54.80	62.16 (8.3)
'LCvirtualV3'	72.35	85.11	59.33	64.35	86.50	49.31	69.49 (14.7)
'MRvirtualV1'	73.47	82.72	71.68	63.42	82.12	62.13	72.59 (8.8)
'SPrealV1'	56.02	57.86	63.94	55.03	76.56	50.88	60.05 (9.1)
'SPrealV2'	60.02	60.44	48.60	53.22	71.60	60.58	59.08 (7.8)
'SPrealV3'	79.75	85.11	66.99	62.63	33.72	135.95	77.36 (33.8)
'SPvirtualV1'	73.11	80.79	54.15	73.89	109.29	63.35	75.76 (18.9)

Table L.3: *Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 221*

Test 212



Tempo (bpm)	Beat number				
Name of Fragment	***1***	***2***	***3***	***4***	***Global Tempo (SD)***
'LCrealA4'	43.6	55.5	40.8	60.3	50.1 (9.3)
'LCrealC6'	49.3	51.9	44.0	52.3	49.4 (3.8)
'LCrealD7'	48.1	53.7	37.4	54.8	48.5 (8.0)
'LCrealD8'	47.5	50.3	37.1	49.9	46.2 (6.2)
'LCrealE9'	38.2	47.4	35.8	53.8	43.8 (8.3)
'LCvirtA1'	50.4	65.1	41.1	77.2	58.4 (15.9)
'LCvirtA2'	54.9	60.8	41.8	63.1	55.2 (9.6)
'LCvirtA5'	50.4	59.6	39.8	55.2	51.2 (8.5)
'LCvirtB3'	55.7	60.7	49.2	61.5	56.7 (5.7)
'LCvirtB4'	45.8	57.0	40.3	64.3	51.9 (10.8)
'LCvirtB6'	48.9	55.4	45.9	69.3	54.9 (10.4)
'LCvirtC7'	52.2	59.7	45.3	42.6	49.9 (7.6)
'LCvirtD8'	51.7	48.2	38.2	65.4	50.8 (11.2)
'LCvirtE10'	41.4	36.4	26.0	27.5	32.8 (7.3)
'LCvirtE9'	33.0	60.4	33.8	41.0	42.1 (12.8)
'MRrealA1'	51.0	57.7	34.7	66.4	52.5 (13.4)
'MRvirtA1'	51.9	58.7	37.1	67.9	53.9 (13.0)
'MRvirtA2'	50.5	62.1	41.0	79.2	58.2 (16.4)
'MRvirtA3'	53.3	62.4	40.2	70.0	56.5 (12.8)
'MRvirtA4'	46.9	66.2	45.0	80.3	59.6 (16.8)
'MRvirtB5'	52.9	67.7	42.4	87.6	62.6 (19.6)
'MRvirtB6'	48.6	62.9	43.3	51.6	51.6 (8.3)
'MRvirtB7'	49.1	58.9	38.4	56.6	50.7 (9.2)
'MRvirtC8'	57.8	52.6	41.2	52.4	51.00 (7.0)
'MRvirtD9'	50.7	58.3	39.8	49.2	49.5 (7.6)
'MRvirtE10'	43.6	46.7	34.6	46.0	42.8 (5.6)
'MRvirtE11'	43.7	50.0	36.7	49.9	45.1 (6.3)
'SPrealB2'	43.3	55.1	38.0	72.0	52.1 (15.1)
'SPrealC3'	45.0	58.8	38.6	37.2	44.9 (9.8)
'SPvirtA1'	52.1	61.8	45.7	63.4	55.8 (8.3)
'SPvirtA2'	43.4	58.7	41.0	68.0	52.8 (12.8)
'SPvirtB3'	47.2	62.4	43.9	88.5	60.5 (20.4)
'SPvirtC4'	47.7	49.7	41.8	69.4	52.1 (12.0)
'SPvirtD5'	48.4	55.7	39.7	60.8	51.1 (9.2)

Table L.4: *Tempo expressed in Beats Per Minute (bpm) for each note of phrase, global (average) tempo across the phrase and standard deviation for Test 212*

Acronyms

BR Bass Ratio.

Br Brilliance.

BRIR Binaural Room Impulse Response.

C_{80} Speech Clarity.

C_{80} Clarity.

EDT Early Decay Time.

ESS Exponential Swept Sine.

G Strength.

G_e Early Strength.

G_{late} Late Strength.

HATS Head and Torso Simulator.

IACC Inter Aural Cross-correlation.

ILD Inter-aural Level Delay.

ITD Inter-aural Time Difference.

ITDG Initial Time Delay Gap.

JND Just Noticeable Difference.

LTAS Long-term Average Spectrum.

MDS Multi-dimensional Scaling.

OBRIR Oral-binaural room impulse response.

PCA Principal Components Analysis.

RES Reverberation Enhancement System.

RIR Room Impulse Response.

Room Gain Room Gain.

RR160 Running Reverberation.

RRAS Real-time Room Acoustic Simulation.

RT60 Reverberation Time.

SIL Sound Intensity Level.

SIRR Spatial Impulse Response Rendering.

SPL Sound Pressure Level.

SRIR Spatial Room Impulse Response.

ST Stage Support.

$\mathbf{ST}_{early}$ Early Stage Support.

$\mathbf{ST}_{late}$ Late Stage Support.

$\mathbf{ST}_{total}$ Total Stage Support.

$\mathbf{ST}_{v}$ Voice Support.

STI Speech Transmission Index.

T30 Reverberation Time.

VAE Virtual Acoustic Environment.

VBAP Vector Based Amplitude Panning.

VSS Virtual Singing Studio.

References

[1] A. H. Marshall, "Acoustical determinants for the architectural design of concert halls," *Architectural Science Review*, vol. 11, no. 3, pp. 81–87, 1968.

[2] A. C. Gade, "Investigations of musicians' room acoustics conditions in concert halls. part i: Methods and laboratory experiments," *Acustica*, vol. 69, pp. 193–203, 1989.

[3] K. Ueno, K. Kato, and K. Kawai, "Effect of room acoustics on musicians' performance. part i: Experimental investigation with a conceptual model," *Acta Acustica united with Acustica*, vol. 96, no. 3, pp. 505–515, 2010.

[4] G. Zarlino, *Le istitutioni harmoniche*. facs. edn bologna 1966 ed., 1588.

[5] S. Favrot and J. Buchholz, "Loudspeaker-based room auralisation in auditory perception research," in *International Workshop on the Principles and Applications of Spatial Hearing*, 2009.

[6] S. Favrot, *A loudspeaker-based room auralization system for auditory research*. Phd, DTU, Denmark, 2010.

[7] K. Ueno and H. Tachibana, "A consideration on acoustic properties on concert-hall stages," in *Proceedings of the International Symposium on Room Acoustics*, (Melbourne, Australia), 2010.

[8] K. Ueno and H. Tachibana, "A consideration on acoustic properties on concert-hall stages," *Building Acoustics*, vol. 18, no. 3, pp. 221–235, 2011.

[9] K. Ueno and H. Tachibana, "Experimental study on the evaluation of stage acoustics by musicians using a 6-channel sound simulation system," *Acoust. Sci. & Tech.*, vol. 24, no. 3, pp. 130–138, 2003.

[10] Z. S. Kalkandjiev and S. Weinzierl, "The influence of room acoustics on solo music performance: An empirical case study," *Acta Acustica united with Acustica*, vol. 99, pp. 433–441, 2013.

[11] Z. S. Kalkandjiev and S. Weinzierl, "Room acoustics viewed from the stage: Solo performers' adjustments to the acoustical environment," in *International Symposium on Room Acoustics*, 2013.

[12] W. Woszczyk, D. Ko, and B. Leonard, "Virtual stage acoustics: A flexible tool for providing useful sounds for musicians," in *Proceedings of the International Symposium on Room Acoustics*, 2010.

[13] W. Woszczyk and W. L. Martens, "Evaluation of virtual acoustic stage support for musical performance," in *Acoustics 08 Paris*, pp. 1041–1046, 2008.

[14] K. Kato, K. Ueno, and K. Kawai, "Musicians' adjustment of performance to room acoustics, part ii: Acoustical analysis of performed sound signals.," in *19th International Congress on Acoustics, Madrid, 2-7 September 2007*, 2007.

[15] K. Kato, K. Ueno, and K. Kawai, "Musicians' adjustment of performance to room acoustics, part III: understanding the variations in musical expressions," in *Acoustics 08, Paris, June 29 - July 4*, 2008.

[16] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, p. 839?855, 2007.

[17] T. L. Bonebright, "An investigation of data collection methods for auditory stimuli: Paired comparisons versus a computer sorting task," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 2, p. 275?278, 1996.

[18] T. Lokki, "Tasting music like wine: Sensory evaluation of concert halls," *Physics Today*, vol. 67, no. 1, pp. 27–32, 2014.

[19] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3148–3161, 2012.

[20] D. M. Howard and J. Angus, *Acoustics and psychoacoustics*. Oxford: Focal, 2006.

[21] T. Rossing, P. Wheeler, and F. Moore, *The science of sound*. Addison Wesley, 2002.

[22] M. Vorländer, *Auralization*. Springer, Berlin, 2007.

[23] H. Kuttruff, *Room acoustics*. New York: Spon Press, 1973.

[24] M. Barron, *Auditorium acoustics and architectural design*. E & FN Spon, 1993.

[25] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," *Preprints-Audio Engineering Society*, 2000.

[26] P. Fausti and A. Farina, "Acoustic measurements in opera houses: Comparison between different techniques and equipment," tech. rep.

[27] J. Merimaa, T. Peltonen, T. Lokki, and A. O. C. Engineers, "Concert hall impulse responses pori, finland," *Available onlin e at: http://www. acoustics. hut. fi/projects/poririrs [Las t viewed 25-Jul-2008]*, 2005.

[28] A. Farina, P. Martignon, A. Azzali, and A. Capra, "Listening tests performed inside a virtual room acoustic simulator," in *I seminario Música Ciênca e Tecnologia "Acústica Musical"*, Nov. 2004.

[29] "ISO 3382acoustics - measurement of the reverberation time of rooms with reference to other acoustical parameters.," 1997.

[30] S. Wallace, *Collected Papers on Acoustics*. Cambridge: Harvard University Press, 1923.

[31] S. Campanini and A. Farina, "A new audacity feature: room objective acoustical parameters calculation module," in *Proc. Linux audio conference*, 2009.

[32] D. Howard and L. Moretti, *Sound and Space in Renaissance Venice: Architecture, Music, Acoustics*. Yale University Press, 2009.

[33] S. Cerdá, R. M. Cibrián, A. Giménez, S. Girón, and T. Zamarreño, "Mismatches between objective parameters and measured perception assessment in room acoustics: a holistic approach," *Building and Environment*, pp. 0360–1323, 2014.

[34] A. H. Marshall, D. Gottlob, and H. Alrutz, "The acoustical conditions preferred for ensemble," *The Journal of the Acoustical Society of America*, vol. 63, no. S1, pp. S35–S36, 1978.

[35] J. S. Bradley, "Using ISO 3382 measures, and their extensions, to evaluate acoustical conditions in concert halls," *Acoustical Science and Technology*, vol. 26, no. 2, pp. 170–178, 2005.

[36] M. Barron, "Objective assessment of concert hall acoustics," *Proceedings of the Institute of Acoustics*, vol. 28, no. 2, pp. 70–78, 2006.

[37] D. Bonsi, M. Longair, P. Garsed, and R. Orlowski, "Acoustic and audience response analyses of eleven venetian churches," in *Acoustics 08 Paris, June 29- July 4,*, pp. 3088–3092, 2008.

[38] T. J. Cox, W. J. Davies, and Y. Lam, "The sensitivity of listeners to early sound field changes in auditoria," *Acta Acustica united with Acustica*, vol. 79, no. 1, 1993.

[39] A. C. Gade, "Investigations of musicians' room acoustic conditions in concert halls. {ii:} field experiments and synthesis of results," *Acustica*, vol. 69, pp. 249–262, 1989.

[40] D. Griesinger, "Recent experiences with electronic acoustic enhancement in concert halls and opera houses," *material from David Griesinger's Internet Home Page, undated but prior to May 2002*, 2002.

[41] T. Lokki, H. Vertanen, A. Kuusinen, J. Ptynen, and S. Tervo, "Auditorium acoustics assessment with sensory evaluation methods," in *Proc. ISRA*, p. 29?31, 2010.

[42] J. Bradley, G. Soulodre, and S. Norcross, "Factors influencing the perception of bass," *J. Acoust. Soc. Am*, vol. 101(5), p. 3135, 1997.

[43] L. Beranek, *Concert Halls and Opera Houses: Music, Acoustics, and Architecture.* Springer, NY, 2nd edition, 2004.

[44] G. Kearney, *Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction*. PhD thesis, Trinity College, Dublin, Ireland, 2010.

[45] J. J. Dammerud, *Stage Acoustics for Symphony Orchestras in Concert Halls*. PhD thesis, University of Bath, 2009.

[46] M. Barron and C. Chinoy, "1:50 scale acoustic models for objective testing of auditoria," *Applied Acoustics*, vol. 12, no. 5, pp. 361–375, 1979.

[47] J. Meyer, "Understanding the orchestral stage environment from the musician's, singer's and conductor's point of view.," in *Wallace Clement Sabine Centennial Symposium, Cambridge, Massachusetts.*, pp. 93–96, 1994.

[48] K. Ueno and K. Tachibana H., "Analysis of musicians' evaluation of acoustics in concert halls based on the individual-scale method," *J. Acoust. Soc. Jpn*, vol. 59, pp. 591–602, 2003.

[49] Y. H. Kim, J. Y. Jeon, and D. Cabrera, "Evaluation of stage support for musicians' performance in a concert hall," in *20th International Congress on Acoustics*, 2010.

[50] M. Cederlöf, "Podium Acoustics for the symphony orchestra an investigation of correlations between subjective and objective acoustic parameters," Master's thesis, Department of Speech, Music and Hearing (TMH) Royal Institute of Technology, 2006.

[51] J. Y. Jeon and M. Barron, "Evaluation of stage acoustics in seoul arts center concert hall by measuring stage support," *The Journal of the Acoustical Society of America*, vol. 117, p. 232, 2005.

[52] J. J. Dammerud and M. Barron, "Early subjective and objective studies of concert hall stage conditions for orchestral performance," in *19th International Congress on Acoustics*, (Madrid, Spain), 2007.

[53] J. Brunskog, A. C. Gade, G. P. Bellester, and L. R. Calbo, "Increase in voice level and speaker comfort in lecture rooms," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2072–2082, 2009.

[54] D. Pelegrín-García, "Comment on "increase in voice level and speaker comfort in lecture rooms"," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1161–1164, 2011.

[55] D. Griesinger, "Further investigation into the loudness of running reverberation," in *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, vol. 17, p. 35?35, 1995.

[56] P. Holman, *Henry Purcell*. Oxford studies of composers, Oxford ; New York: Oxford University Press, 1994.

[57] M. A. Poletti, "Active acoustic systems for the control of room acoustics," 2010.

[58] E. B. Brixen and C. Wolter, "Optimising stage acoustics by the aid of electro acoustics.," in *Proc.I.o.A 28,*, pp. 13–17, 2006.

[59] "LARES." http://www.lares-lexicon.com/.

[60] "Systems for improved acoustic performance B.V.." http://www.siap.nl/.

[61] R. Schwenke and S. Ellison, "Objective assessment of active acoustic system performance," in *Proc. International Symposium on Room Acoustics*, 2010.

[62] "Active Field Control (Reverberation Enhancement System) - YAMAHA global gateway." http://www.yamaha-afc.com/.

[63] T. Lokki and J. Hiipakka, "A time-variant reverberation algorithm for reverberation enhancement systems," in *Proceedings of COST G-6 Conference on Digital Audio Effects*, pp. 6–8, 2001.

[64] J. Pätynen, "Virtual acoustics in practice rooms," Master's thesis, Helsinki Unviersity of Technology, Department of Electrical and Communications Engineering, 2007.

[65] M. Kleiner, B. Dalenbäck, and P. Svensson, "Auralization- an overview," *J. Audio Eng. Soc*, vol. 41, no. 11, pp. 861–875, 1993.

[66] N. Xiang and J. Blauert, "Binaural scale modelling for auralisation and prediction of acoustics in auditoria," *Applied Acoustics*, vol. 38, no. 2-4, pp. 267–290, 1993.

[67] J.-D. Polack, X. Meynial, and V. Grillon, "Auralization in scale models: Processing of impulse response," *J. Audio Eng. Soc*, vol. 41, no. 11, pp. 939–945, 1993.

[68] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating interactive virtual acoustic environments," *J. Audio Eng. Soc*, vol. 47, no. 9, pp. 675–705, 1999.

[69] F. Otondo, J. Rindel, and C. Christensen, "Directional patterns and recordings of musical instruments in auralizations," in *Proc. Workshop on Current Research Directions in Computer Music, Barcelona, Spain*, pp. 230–232, 2001.

[70] M. C. Vigeant, L. M. Wang, and J. Holger, "Investigations of mult-channel auralization technique for solo instruments and orchestra," (Madrid, Spain), 2-7 September, 2007.

[71] T. Lokki and J. Pätynen, "Applying anechoic recordings in auralization," in *Proc. of the EAA Symposium on Auralization, Espoo, Finland, 15-17 June 2009*, 2009.

[72] M. Vigeant, L. Wang, and J. Rindel, "Room acoustics computer modeling: Study of the effect of source directivity on auralizations," *Architectural Engineering aculty Publications, University of Nebraska, Lincoln*, 2006.

[73] M. C. Vigeant, L. M. Wang, and J. H. Rindel, "Objective and subjective evaluations of the multi-channel auralization technique as applied to solo instruments," *Applied Acoustics*, vol. 72, pp. 311–323, 2011.

[74] L. Wang and M. Vigeant, "Objective and subjective evaluation of the use of directional sound sources in auralizations," 2004.

[75] J. H. Rindel, F. Otondo, and C. L. Christensen, "Sound source representation for auralization," in *Proceedings of International Symposium on Room Acoustics: Design and Science*, 2004.

[76] D. Cabrera, D. Lee, R. Collins, B. Hartmann, W. L. Martens, and H. Sato, "Characterising the variation in oral-binaural room impulse responses for horizontal rotations of a head and torso simulator," *Building Acoustics*, vol. 18, no. 1, pp. 227–252, 2011.

[77] D. Cabrera, P. J. Davis, and A. Connolly, "Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment," *Journal of Voice*, vol. 25, no. 6, pp. e291–e303, 2011.

[78] M. Beeson, A. Moore, D. Murphy, S. Shelley, and A. Southern, "Renderair room acoustics simulation using a hybrid digital waveguide mesh approach," in *Audio Engineering Society Convention 124*, May 2008.

[79] D. Murphy, M. Beeson, S. Shelley, A. Moore, and A. Southern, "Hybrid room impulse response synthesis in digital waveguide mesh based room acoustics simulation," in *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08)*, pp. 129–136, Citeseer, 2008.

[80] A. Southern, S. Siltanen, D. Murphy, and L. Savioja, "Room impulse response synthesis and validation using a hybrid acoustic model," *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

[81] A. Foteinou and D. Murphy, "Evaluation of the psychoacoustic perception of geometric acoustic Modeling-Based auralization," in *Audio Engineering Society Convention 130*, 2011.

[82] A. Southern, J. Wells, and D. , "Rendering walk-through auralisations using wave-based Acoustical models," in *17th European Signal Processing Conference (EUSIPCO 2009) Glasgow, Scotland, August 24-28, 2009*, 2009.

[83] A. J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc*, vol. 36, no. 12, pp. 977–995, 1988.

[84] M. A. Gerzon, "Practical periphony the reproduction of full-sphere sound," in *Audio Engineering Society Convention 65*, 1980.

[85] B. Wiggins, "Has ambisonics come of age," *Proc of the Institute of Acoustics*, vol. 30, no. 6, 2008.

[86] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[87] V. Pulkki and M. J., "Spatial impulse response rendering: A tool for reproducing room acoustics for Multi-Channel listening," *Journal of the Audio Engineering Society*, vol. 55,, no. 6, pp. 503–516, 2007.

[88] V. Pulkki, "Spatial sound reproduction with directional audio coding | mendeley," *Journal of the Audio Engineering Society*, vol. 46, no. 6, pp. 458–466, 1997.

[89] D. Barry and G. Kearney, "Localization quality assessment in source separation-based upmixing algorithms," 2009.

[90] T. Lokki, J. Patynen, T. Peltonen, and O. Salmensaari, "A rehearsal hall with virtual acoustics for symphony orchestras," in *the 126th Audio Engineering Society (AES) Convention*, 2009.

[91] "Harpex - technology." http://harpex.net.

[92] "WigWare the blog of bruce." http://www.brucewiggins.co.uk.

[93] "Ambisonic decoders." http://www.dmalham.freeserve.co.uk.

[94] C. Guastavino, V. Larcher, G. Catusseau, and P. Boussard, "Spatial audio quality evaluation: comparing transaural, ambisonics and stereo," in *Proceedings of the 13th International Conference on Auditory Display, Montral, Canada*, 2007.

[95] Favrot and J. Buchholz, "LoRA: A loudspeaker-based room auralization system," *Acta Acustica united with Acustica*, vol. 96, no. 2, pp. 364–375, 2010.

[96] J. Pätynen and T. Lokki, "Evaluation of concert hall auralization with virtual symphony orchestra," in *Proceedings of the International Symposium on Room Acoustics, ISRA, 29-31 August 2010, Melbourne, Australia*, 2010.

[97] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *Journal of Sound and Vibration*, vol. 77, no. 2, 1981.

[98] S. Cerdá, A. Giménez, and R. M. Cibrián, "An objective scheme for ranking halls and obtaining criteria for improvements and design," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 419–430, 2012.

[99] T. Lokki, *Physically-based auralization*. PhD thesis, Helsinki University of Technology, Helsinki, 2002.

[100] T. Lokki and L. Savioja, "Evaluation of auralization results," in *Forum Acusticum*, 2005.

[101] A. Kuusinen, J. Pätynen, S. Tervo, and T. Lokki, "Relationships between preference ratings, sensory profiles, and acoustical measurements in concert halls," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 239–250, 2014.

[102] T. Lokki, "Sensory evaluation of concert hall acoustics," in *Proceedings of Meetings on Acoustics*, vol. 19, p. 032004, Acoustical Society of America, 2013.

[103] W. Chiang and J. Huang, "Subjective evaluation of acoustical environments for solo performance," *Building Acoustics*, vol. 6, no. 1, pp. 17–36, 1999.

[104] D. Cabrera, A. Azzali, A. Capra, A. Farina, and P. Martignon, "Perceived room size and source distance in five simulated concert auditoria," in *Proc. 12th International Congress on Sound and Vibration*, 2005.

[105] T. Lokki, R. Kajastila, and T. Takala, "Virtual acoustic spaces with multiple reverberation enhancement systems," in *AES30th International Conference*, 2007.

[106] R. Kajastila, T. Lokki, and L. Savioja, "Interactive multi-channel auralization with camera-based tracking," in *19th International Congress on Acoustics, Madrid, 2-7 September*, 2007.

[107] K. Ueno and T. Hideki, "Cognitive modeling of musician's perception in concert halls," *Acoust. Sci. & Tech.*, vol. 26, pp. 156–161, 2005.

[108] W. Chiang, S. Chen, and C. Huang, "Subjective assessment of stage acoustics for solo and chamber music performances," *Acta Acustica united with Acustica*, vol. 89, no. 5, pp. 848–856, 2003.

[109] W. L. Martens and W. Woszczyk, "Virtual acoustic reproduction of historical spaces for interactive music performance and recording," *The Journal of the Acoustical Society of America*, vol. 116, p. 2484, 2004.

[110] S. Yokoyama, K. Ueno, S. Sakamoto, and H. Tachibana, "6-channel recording/reproduction system for 3-dimensional auralization of sound fields," *Acoustical Science and Technology*, vol. 23, no. 2, pp. 97–103, 2002.

[111] E. Battenberg and R. Avižienis, "Implementing real-time partitioned convolution algorithms on conventional operating systems," in *International conference on Digital Audio Effects*, 2011.

[112] J. Miller, M. Anderson, E. Wenzel, and B. McClain, "Latency measurement of a real-time virtual acoustic environment rendering system," in *Proceedings of the International Conference on Auditory Display (ICAD 2003)*, 2003.

[113] E. Wenzel, "The impact of system latency on dynamic performance in virtual acoustic environments," *Target*, vol. 135, p. 180, 1998.

[114] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan, "Effect of time delay on ensemble accuracy," in *Proceedings of the International Symposium on Musical Acoustics*, vol. 31, 2004.

[115] A. Libeaux, T. Lent, D. Houben, and M. Kob, "Voice assessment in choir singers a virtual choir environment," in *19th Internationl Congress on Acoustics Madrid, 2-7 September 2007*, (Madrid), 2007.

[116] A. Foteinou, D. Murphy, and A. Masinton, "Verification of geometric Acoustics-Based auralization using room acoustics measurement techniques," in *Audio Engineering Society Convention 128*, 2010.

[117] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Audio Engineering Society Convention 129*, Nov. 2010.

[118] "Liquidsonics reverberate: Convolution reverb for vst." http://www.liquidsonics.com.

[119] "Microphone data website." www.microphone-data.com.

[120] M. Brookes, "Voicebox: A speech processing toolbox for matlab." Imperial College, Software Library, www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[121] B. B. Monson, *HIGH-FREQUENCY ENERGY IN SINGING AND SPEECH*. PhD thesis, THE UNIVERSITY OF ARIZONA, 2011.

[122] P. G. Singh and I. J. Hirsh, "Influence of spectral locus and f0 changes on the pitch and timbre of complex tones," *Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2650–2661, 1992.

[123] "Genelec commercial install."

[124] C. Rioux, "Reaper." `https://github.com/codyrioux/reaper`, 2014.

[125] I. Laird and D. Murphy, "Energy-based calibration of virtual performance systems," in *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, (York, UK, September 17-21, 2012), 2012.

[126] I. Bork, "A comparison of room simulation software - the 2nd round robin on...," *Acta Acustica united with Acustica*, vol. 86, no. 6, pp. 943–956, 2000.

[127] M. Queiroz, F. Iazzetta, F. Kon, M. H. Gomes, F. L. Figueiredo, B. Masiero, L. K. Ueda, L. Dias, M. H. Torres, and L. F. Thomaz, "AcMus: an open, integrated platform for room acoustics research," *Journal of the Brazilian Computer Society*, vol. 14, no. 3, pp. 87–103, 2008.

[128] "AcMus - room acoustic parameters - file exchange - MATLAB central."

[129] B. G. Churcher, "The acoustics of churches," *British Journal of Applied Physics*, vol. 15, p. 249, 1964.

[130] M. Planck, "Die natürliche stimmung in der modernen vokalmusick," *Vierteljahrsschrift für Musikwissenschaft*, vol. 9, no. 4, 1893.

[131] D. C. Miller, *The Science of Musical Sounds*. New York: The Macmillan Company, 1916.

[132] L. Vernon, "Synchronization of chords in artistic piano music.," in *In: Carl E. Seashore (ed.). Objective Analysis of Musical Performance*, Studies in the Psychology of Music, Iowa: University Press, IV, 306–345. ed., 1937.

[133] C. Seashore, "The natural history of the vibrato," *Proceedings of the National Academy of Sciences*, vol. 17, no. 12, 1931.

[134] C. E. Seashore, *Psychology of Music*. Courier Dover Publications, 1938.

[135] A. Lerch, "An introduction to audio content analysis: Applications in signal processing," 2012.

[136] C. Shackford, "Some aspects of perception. i: Sizes of harmonic intervals in perfrmance," *Journal of Music Theory*, vol. 5, no. 2, pp. 162–202, 1961.

[137] S. Bolzinger and J. C. Risset, "A preliminary study on the influence of room ACoustics on piano performance," *Le Journal de Physique IV*, vol. 02, no. C1, pp. C1–93–C1–96, 1992.

[138] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–30, 2002.

[139] J. Devaney and D. Ellis, "An empirical approach to studying intonation tendencies in polyphonic vocal performances," *Journal of Interdisciplinary Music Studies*, vol. 2, no. 1-2, pp. 141–156, 2008.

[140] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J. L. Collette, "Vibrato: detection, estimation, extraction, modification," in *Proceedings of Digital Audio Effects Workshop*, vol. 99, 1999.

[141] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle, "Automatic characterisation of musical signals: Feature extraction and temporal segmentation," *Journal of New Music Research*, vol. 28, pp. 281–295, Dec. 1999.

[142] P. Monahan, C. Gobl, B. Foley, and A. N. Chasaide, "Semi-automatic identification of vocal source-filter," in *Proc. Of the 6th Irish DSP and Control Colloqium*, (Belfast), pp. 127–134, 1995.

[143] E. Molina, E. Gomez, and I. Barbancho, "Automatic scoring of singing voice based on melodic similarity measures," Master's thesis, MS thesis, Universitat Pompeu Fabra, Music Technology Group, 2012.

[144] Y. Horii, "Automatic analysis of voice fundamental frequency and intensity using a visi-pitch," *Journal of Speech and Hearing Research*, vol. 26, pp. 467–71, 1983.

[145] J. Devaney, M. I. Mandel, and I. Fujinaga, "Study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (AMPACT)," in *Proceedings of the 2012 International Society on Music Information Retrieval Conference (ISMIR)*, pp. 511–6, 2012.

[146] T. Eerola and P. Toiviainen, *MIDI Toolbox: MATLAB Tools for Music Research.* Kopijyv, Jyvskyl, Finland.: University of Jyvskyl:, 2004.

[147] "Download sonic visualiser." http://www.sonicvisualiser.org/download.html.

[148] N. Cook and D. Leech-Wilkinson, "A musicologist's guide to sonic visualiser," *London: Centre for the History and Analysis of Recorded Music. URL (retrieved March 2010): http://www. charm. rhul. ac. uk/analysing/p9_1. html*, 2009.

[149] C. Cannam, C. Landone, and M. Sandler, "Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference*, (Firenze, Italy), pp. 1467–1468, 2010.

[150] J. Devaney, M. I. Mandel, D. Ellis, and I. Fujinaga, "Automatically extracting performance data from recordings of trained singers," *Psycholomusicology: Music, Mind and Brain*, vol. 21, no. 1 & 2, pp. 108–136, 2011.

[151] J. Devaney, *An empirical study of the influence of musical context on intonation practices in solo singers and SATB ensembles*. Phd, McGill University, Canada, 2011.

[152] A. Gabrielsson and P. N. Juslin, "Emotional expression in music performance: Between the performer's intention and the listener's experience," *Psychology of Music*, vol. 24, pp. 68–91, Apr. 1996.

[153] B. H. Repp, "Patterns of note onset asynchronies in expressive piano performance," *Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3917–3932, 1996.

[154] C. Krumhansl, "A perceptual analysis of mozart's piano sonata k. 282: Segmentation, tension, and musical ideas," *Music Perception: An Interdisciplinary Journal,*, vol. 13, no. 3, pp. 401–432, 1996.

[155] R. Timmers, R. Ashley, P. Desain, and H. Heijink, "The influence of musical context on tempo rubato," *Journal of New Music Research*, vol. 29, no. 2, pp. 131–158, 2000.

[156] N. Todd, "A model of expressive timing in tonal music," *Music Perception: An Interdisciplinary Journal*, vol. 3, no. 1, pp. 33–57, 1985.

[157] E. F. Clarke, "The perception of expressive timing in music," *Psychological Research*, vol. 51, no. 1, pp. 2–9, 1989.

[158] A. Lerch, *Software-Based Extraction of Objective Parameters from Music Performances*. GRIN Verlag, Apr. 2009.

[159] B. H. Repp, "Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study," *Psychological research*, vol. 56, no. 4, p. 269284, 1994.

[160] D. M. Howard, "Variation of electrolayngographically derived closed quotient for trained and untraned adult female singers," *Journal of Voice*, vol. 9, no. 2, pp. 163–72, 1995.

[161] V. M. O. Barrichelo, R. J. Heuer, C. M. Dean, and R. T. Staloff, "Comparison of singer's formant, speaker's ring and LTA spectrum among classical singers and untrained normal speakers," *Journal of Voice*, vol. 15, no. 3, pp. 344–350, 2001.

[162] D. Mürbe, G. Hofmann, F. Pabst, and J. Sundberg, "Auditory and kinesthetic feedback in singing - significance and effects of training on pitch control," in *3rd International Workshop MAVEBA 2003*, (Firenze), pp. 183–186, 2003.

[163] D. Rossiter, D. M. Howard, and M. DeCosta, "Voice development under training with and without the influence of real-time visually presented feedback," *Journal of the Acoustical Society of America*, vol. 99, no. 5, pp. 3253–3256, 1996.

[164] D. M. Howard, H. Daffern, and J. S. Brereton, "Quantitative voice quality analyses of a soprano singing early music in three different performance styles," *ScienceDirect - Biomedical Signal Processing and Control :*, 2011.

[165] T. F. Cleveland, J. Sundberg, and R. Stone, "Long-term-average spectrum characteristics of country singers during speaking and singing," *Journal of Voice*, vol. 15, no. 1, pp. 54–60, 2001.

[166] R. Timmers, "Perception of music performance on historical and modern commercial recordings," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, p. 2872, 2007.

[167] J. A. Bowen, "Tempo, duration, and flexibility: Techniques in the analysis of performance," *Journal of Musicological Research*, vol. 16, no. 2, pp. 111–156, 1996.

[168] J. Sundberg, *The Science of the Singing Voice*. Dekalb, Illinois: Northern Illinois University Press, 1987.

[169] M. Atkinson and S. McHanwell, *Basic medical science for speech and language therapy students*. London; Philadelphia: Whurr Publishers, 2002.

[170] M. Garcia, "A complete treatise on the art of singing," in *A complete treatise on the art of singing*, ed. tr paschke ed., 1841.

[171] J. Sundberg, *Research Aspects on Singing*. Stockholm: Royal Swedish Academy of Music, 1981.

[172] T. F. Cleveland, "A clearer view of singing voice production: 25 years of progress," *Journal of Voice*, vol. 8, no. 1, pp. 18–23, 1994.

[173] S. Ternström, "Choir acoustics: An overview of scientific research published to date," *International Journal of Research in Choral Singing*, vol. 1, no. 1, 2003.

[174] M. Kob, N. Henrich, H. Herzel, D. M. Howard, I. Tokuda, and J. Wolfe, "Analysing and understanding the singing voice: Recent progress and open questions," *Current Bioinformatics*, vol. 6, pp. 362–374, 2011.

[175] J. Sundberg, P. Gramming, and J. Loverti, "Comparisons of pharynx, source, formant and pressure characteristics in operatic and musical theatre singing," *Journal of Voice*, vol. 7, no. 4, pp. 301–310, 1993.

[176] M. Echternach, M. Markl, and B. Richter, "Vocal tract configurations in yodelling : prospective comparison of two swiss yodeller and two non-yodeller subjects," *Logopedics Phoniatrics Vocology*, vol. 36, no. 3, pp. 109–113, 2011.

[177] K. L. Reid, P. Davis, J. Oates, D. Cabrera, S. Ternström, M. Black, and J. Chapman, "The acoustic characteristics of professional opera singers performing in chorus versus solo mode," *Journal of Voice*, vol. 21, no. 1, pp. 35 – 45, 2007.

[178] T. D. Rossing, J. Sundberg, and S. Ternström, "Acoustic comparison of voice use in solo and choir singing," *Journal of the Acoustical Society of America*, vol. 79, pp. 1975–1981, 1986.

[179] T. D. Rossing, J. Sundberg, and S. Ternström, "Acoustic comparison of soprano solo and choir singing," *Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 830–836, 1987.

[180] T. F. Cleveland, "Acoustic properties of voice timbre types and their influence on voice classification," *Journal of the Acoustical Society of America*, vol. 61, 1997.

[181] R. Miller and H. K. Schutte, "Spectral analysis of several categories of timbre in a professional male (tenor) voice," *Journal of Research into Singing 7*, pp. 6–10, 1986.

[182] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1417–1430, 2005.

[183] J. Brereton, "A voice source analysis of chest and head registers in singing," in *The 3rd International Physiology and Acoustics of Singing Conference* (D. M. Howard, J. Brereton, and H. Daffern, eds.), 2006.

[184] I. R. Titze, "A framework for the study of vocal registers," *Journal of Voice*, vol. 2, pp. 183–94, 1988.

[185] M. Echternach, J. Sundberg, T. Baumann, M. Markl, and B. Richter, "Vocal tract area functions and formant frequencies in opera tenors : modal and falsetto registers," *The Journal of the Acoustical Society of America*, vol. 129, no. 6, p. 3955, 2011.

[186] M. Echternach, J. Sundberg, M. Zander, and B. Richter, "Perturbation measurements in untrained male voices' transitions from modal to falsetto register.," *Journal of Voice*, pp. 663–669, 2014.

[187] A. Libeaux, "The EGG spectrum slope in speakers and singers: Variations related to voice sound pressure level, vowel and fundamental frequency," Master's thesis, The Department of Speech, Music and Hearing, KTH Royal Institute of Technology,, Stockholm, Sweden, 2010.

[188] M. F. Pederson, S. Moller, S. Krabbe, P. Bennet, and P. Kitzing, "Change of voice in puberty in choir girls," *Acta Otolaryngol*, vol. suppl 412, pp. 46–9, 1984.

[189] C. Barlow and J. LoVetri, "Closed quotient and spectral measures of female adolescent singers in different singing styles," *Journal of Voice*, vol. 24, no. 3, pp. 314 – 318, 2010.

[190] J. Williams, G. F. Welch, D. M. Howard, and E. Himonides, "A baseline study on male chorister vocal behaviour and development in an intensive professional context.," in *A STINT on Voice Research*, National Centre for Early Music, York: British

Voice Association (BVA), Royal Institute of Technology (KTH), The University of York, Sponsored by the Swedish Foundation for International Cooperation in Research and Higher Education, 2005.

[191] J. Williams, *Teaching Singing to Children and Young Adults*. Oxford: Compton Publishing Ltd, 2013.

[192] A. Gabrielsson, "Music performance research at the millennium," *Psychology of Music*, vol. 31, pp. 221–272, July 2003.

[193] A. Gabrielsson, "The performance of music," in *The psychology of music*, pp. 501–602, New York: Academic Press, 2nd ed.

[194] D. M. Howard, J. Brereton, and H. Daffern, "Case study of voice quality differences in a soprano singing in different early music performance styles," in *6th International Workshop on: Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2009)*, (Firenze, Italy), 2009.

[195] B. Karrick, "An examination of the intonation tendencies of wind instrumentalists based on their performance of selected harmonic musical intervals," *Journal of Research in Music Education*, vol. 46 (1), pp. 112–127, 1998.

[196] A. Stevenson, *Oxford Dictionary of English*. Oxford reference online premium, OUP Oxford, 2010.

[197] H. v. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green, 1885.

[198] J. Brereton, "A voice source analysis of chest and head registers in singing," unpublished MPhil thesis, Trinity College, Dublin, Ireland, 2000.

[199] D. Rossiter, D. M. Howard, and R. Comins, "Objective measurement of voice source and acoustic output change with a short period of vocal tuition," *Journal of Voice*, vol. 4, no. 1, pp. 16–31, 1995.

[200] J. Sundberg, M. Andersson, and C. Hultqvist, "Effects of subglottal pressure variation on professional baritone singers' voice sources," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1965–1971, 1999.

[201] C. Gobl, *The Voice Source in Speech Communication: Production and Perception Experiments Involving Inverse Filtering and Synthesis*. PhD thesis, Dept of Speech, Music and Hearing, KTH, Stockholm, 2003.

[202] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, "Singing power ratio: quantitative evaluation of singing voice quality," *Journal of Voice*, vol. 10, no. 3, pp. 228–235, 1996.

[203] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton, "The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers," *Journal of Voice*, vol. 20, no. 1, pp. 82–88, 2006.

[204] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton & Co, 1960.

[205] S. Ternström and G. Kalin, "Formant frequency adjustment in barbershop quartet singing," in *International Congress on Acoustics, Madrid, 2-7 September 2007*, 2007.

[206] S. Ternström, "Long-term average spectrum characterisitcs of different choirs in different rooms," *Voice*, vol. 2, pp. 55–77, 1993.

[207] D. M. Howard, "Quantifiable aspects of different singing styles - a case study," *Voice*, vol. 1, pp. 47–62, 1992.

[208] D. Noson, K. Kato, and Y. Ando, "Singers' preferred acoustic condition in performance in an opera house and self-perception of the singer's voice.," *Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2436–2437, 2001.

[209] M. Chudy, A. P. Carrillo, and S. Dixon, "On the relation between gesture, tone production and perception in classical cello performance," in *Proceedings of Meetings on Acoustics*, vol. 19, 2013.

[210] T. Cleveland and J. Sundberg, "Acoustic analysis of three male voices of different quality," in *Proceedings of the Stockholm Music Acoustics Conference*, pp. 143–56, 1983.

[211] M. Södersten, A. Hakansson, and B. Hammarberg, "Comparison between automatic and manual inverse filtering procedures for healthy female voices," *Logopedics Phoniatrics Vocology*, vol. 24, pp. 26–38, 1999.

[212] C. Gobl, A. N. Chasaide, and P. Monahan, "Intrinsic voice source characteristics of selected consonants," 1995.

[213] D. M. Howard, *Electroglottography/electrolaryngography*. No. 1 in G - Reference, Information and Interdisciplinary Subjects Series, Plural Pub., 2009.

[214] R. Timmers, "Vocal expression in recorded performances of Schubert songs," *Musicae Scientiae*, vol. 11, no. 2, p. 237?268, 2007.

[215] S. Dalla Bella and C. Palmer, "Tempo and dynamics in piano performance: The role of movement amplitude," in *8th International Conference on Music Perception & Cognition*, (Evanston, IL), 2004.

[216] T. Eerola and P. Toiviainen, "MIR in matlab: The midi toolbox," in *Proceedings of the International Conference on Music Information Retrieval*, pp. 22–27, 2004.

[217] J. Sundberg, A. Friberg, and R. Bresin, "Musician's and computer's tone inter-onset interval in mozart's piano sonata k 332, 2nd mvt, bar 1-20," *TMH-QPSR, KTH*, vol. 45, no. 3, pp. 47–59, 2003.

[218] L. H. Shaffer, "Timing in solo and duet piano performances," *The Quarterly Journal of Experimental Psychology Section A*, vol. 36, no. 4, pp. 577–595, 1984.

[219] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 2005.

[220] J. BEAUCHAMP, "Analysis and synthesis of musical instrument sounds," in *Analysis, Synthesis, and Perception of Musical Sounds* (J. Beauchamp, ed.), Modern Acoustics and Signal Processing, pp. 1–89, Springer New York, 2007.

[221] A. D. Patel and J. R. Iversen, "Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: An empirical study of sound symbolism," in *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona*, pp. 925–928, 2003.

[222] B. Repp, "Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "träumerei"," *JOURNAL of the ACOUSTICAL SOCIETY OF AMERICA*, vol. 92, pp. 2546–2546, 1992.

[223] S. Dixon and W. Goebl, "Pinpointing the beat: Tapping to expressive performances," in *7th International Conference on Music Perception and Cognition*, vol. 1, 2002.

[224] R. Timmers, "Predicting the similarity between expressive performances of music from measurements of tempo and dynamics," *The Journal of the Acoustical Society of America*, vol. 117, no. 1, p. 391, 2005.

[225] S. Dixon, "MATCH: music alignment tool chest," May 2005.

[226] D. Schulenberg, "Tempo relationships in the prelude of J. S. Bach's sixth english suite: A performance studies approach," *Journal of Musicological Research*, vol. 18, pp. 139–160, Jan. 1999.

[227] N. P. McAngus Todd, "The dynamics of dynamics: A model of musical expression," *The Journal of the Acoustical Society of America*, vol. 91, no. 6, pp. 35–40, 1992.

[228] A. Friberg and J. Sundberg, "Time discrimination in a monotonic, isochronous sequence," *Journal of the Acoustical Society of America*, vol. 98, no. 5, pp. 2524–2531, 1995.

[229] C. Palmer, "Music performance," *Annual Review of Psychology*, vol. 48, pp. 115–138, Feb. 1997.

[230] A. Gabrielsson, "Studying emotional expression in music performance," in *Bulletin of the Council for Research in Music Education*, vol. 141, pp. 47–53.

[231] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, 1975.

[232] D. Howard and J. Brereton, "Music that can exhibit intonation drift in a capella SATB quartet singing if sung in-tune," in *Proceedings of Pan European Voice Conference, PEVOC8*, 2009.

[233] P. Q. Pfordresher, S. Brown, K. M. Meier, M. Belyk, and M. Liotti, "Imprecise singing is widespread," *The Journal of the Acoustical Society of America*, vol. 128, pp. 2182–2190, Oct. 2010.

[234] S. Dalla Bella, J. Giguiere, and I. Peretz, "Singing proficiency in the general population," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 11–82, 2007.

[235] M. Berkowska and S. Dalla Bella, "Uncovering phenotypes of poor-pitch singing: the sung performance battery (SPB)," *Frontiers in Psychology*, vol. 4, 2013.

[236] P. Boersma, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–5., 2001.

[237] W. Leukel and T. Stoffer, "The influence of harmonic context on the tuning of thirds played by professional flautists," *Psychology of Music*, vol. 32 (1), pp. 75–88, 2004.

[238] J. M. Barbour, "Just intonation confuted," *Music and Letters*, vol. 19, no. 1, p. 48, 1938.

[239] E. R. Guthrie and H. Morrill, "The fusion of non-musical intervals," *The American Journal of Psychology*, vol. 40, p. 624, Oct. 1928.

[240] L. S. Lloyd, "The myth of equal temperament," *Music & Letters*, vol. 21, no. 4, pp. pp. 347–361, 1940.

[241] D. M. Howard, "Equal or non-equal temperament in a capella SATB singing," *Logopedics, phoniatrics, vocology*, vol. 32, no. 2, pp. 87–94, 2007.

[242] D. M. Howard, "Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation," *Journal of Voice*, vol. 21, pp. 300–315, May 2007.

[243] E. Prame, "Vibrato extent and intonation in professional western lyric singing," *Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 616–621, 1997.

[244] B. Hagerman and J. Sundberg, "Fundamental frequency adjustment in barbershop singing," *Dept. for Speech, Music and Hearing Quarterly Progress and Status Report*, vol. 21, pp. 028–042, 1980.

[245] S. Ternström and J. Sundberg, "Acoustical factors related to pitch precision in choir singing," *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR)*, vol. 2, no. 3, p. 1982, 1982.

[246] A. Vurma and J. Ross, "Production and perception of musical intervals," *Music Perception*, vol. 23, no. 4, pp. 331–44, 2006.

[247] A. Friberg, *A Quantative Rule System for Musical Perfromance*. PhD thesis, KTH, Sweden, 1995.

[248] J. Sundberg, "How can music be expressive?," *Speech Communication*, vol. 13, no. 12, pp. 239 – 253, 1993.

[249] A. Vurma, "Mistuning in two-part singing," *Logopedics Phoniatrics Vocology*, vol. 35, pp. 24–33, 2010.

[250] S. Ternstrm and J. Sundberg, "Intonation precision of choir singers," *Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 59–69, 1988.

[251] M. Mauch, K. Frieler, and S. Dixon, "Intonation in unaccompanied singing: Accuracy, drift and a model of reference pitch memory," *Journal of the Acoustical Society of America*, vol. 136, pp. 401–11, 2014.

[252] R. M. van Besouw, J. S. Brereton, and D. M. Howard, "Range of tuning for tones with and without vibrato," *Music Perception*, vol. 26, no. 2, pp. 145–155, 2008.

[253] R. Bowman Macleod, *Influences of Dynamic Level and Pitch Height on the Vibrato Rates and Widths of Violin and Viola Players.* PhD thesis, Tallahassee, Florida State University, College of Music, 2006.

[254] E. Prame, "Measurement of the vibrato rate of ten singers," *Journal of the Acoustical Society of America*, no. 96, 1994.

[255] M. P. Lynch, R. E. Eilers, K. D. Oller, R. C. Urbano, and P. Wilson, "Influences of acculturation and musical sophistication on perception of musical interval patterns," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 17, no. 4, pp. 967–975, 1991.

[256] J. Sundberg, *The Science of Musical Sounds (Cognition and Perception).* London: Academic Press, 1991.

[257] L. Yoo, D. J. Sullivan, S. Moore, and I. Fujinaga, "The effect of vibrato on the response time in determining the pitch relationship of violin tones," in *Proceedings of the 5th International Conference on Music Perception and Cognition*, (Seoul National University, Korea.), pp. 209–2011, 1998.

[258] R. Stowell, *Violin technique and performance practice in the late eighteenth and early nineteenth centuries.* Cambridge: Cambridge University Press, 1985.

[259] S. Iwamiya, K. Kosugi, and O. Kitamura, "Perceived principal pitch of vibrato tones," *Journal of the Acoustic Society of Japan*, vol. E, no. 4, pp. 73–82, 1983.

[260] J. Sundberg, "Pitch of synthetic sung vowels," in *Speech Transmission Laboratory - Quarterly Progress and Status Report,*, vol. 1, (Stockholm: Royal Institute of Technology.), pp. 34–44, 1972.

[261] J. Sundberg, "Effects of vibrato and the "singing formant" on pitch," *Musicologica Slovatca*, vol. 6, pp. 51–69, 1978.

[262] J. I. Shonle and K. Horan, "The pitch of vibrato tones," *Journal of the Acoustical Society of America,*, vol. 67, pp. 246–252, 1980.

[263] C. d'Alessandro and M. Castellengo, "The pitch of short-duration vibrato tones.," *Journal of the Acoustical Society of America*, vol. 95, pp. 1617–1630, 1994.

[264] C. d'Alessandro and M. Castellengo, "Etude de la perception des notes courtes chantes en presence de vibrato.," in *Proceedings of the XIIth International Congress of Phonetic Sciences,*, pp. 86–89, 5, 1991.

[265] R. M. van Besouw and D. M. Howard, "Effects of carrier and phase on the pitch of long-duration vibrato tones.," *Musicae Scientiae*, vol. 13, no. 1, pp. 139–161., 2009.

[266] H. Gockel, B. C. J. Moore, and R. Carlyon, "Influence of rate of change of frequency on the overall pitch of frequency-modulated tones," *Journal of the Acoustical Society of America*, vol. 109, pp. 701–12, 2001.

[267] T. Dart, *The Interpretation of Music*. London: Hutchinson, 1954.

[268] B. Blesser and L.-R. Salter, *Spaces speak, are you listening?* MIT Press, 2007.

[269] H. Bagenal, "Bach's music and church acoustics," *Music and Letters*, vol. 11, no. 2, 1930.

[270] P. Q. Pfordresher, "Coordination of perception and action in music performance," *Advances in Cognitive Psychology*, vol. 2, no. 2-3, pp. 183–198, 2006.

[271] N. J. Eyring, T. W. Leishman, K. M. Sorensen, and N. G. Eyring, "Methods for automating multichannel directivity measurements of musical instruments in an anechoic chamber.," *J Acoust Soc Am*, vol. 130, pp. 23–99, Oct. 2011.

[272] "Personal webpages of will kimball, trombonist."

[273] I. Nakayama, "Preferred time delay of a single reflection for performers," *Acta Acustica united with Acustica*, vol. 54, no. 4, pp. 217–221, 1984.

[274] D. Ko, W. Woszczyk, and S. H. Chon, "Evaluation of a new active acoustics system in performances of five string quartets," in *Audio Engineering Society Convention 132*, Audio Engineering Society, 2012.

[275] A. H. Marshall and F.-J. Meyer, "The directivity and auditory impressions of singers," *Akustika*, vol. 28, pp. 181–184, 1985.

[276] Z. Schärer and S. Weinzierl, "Room acoustics viewed from the stage: Solo performers' adjustments to the acoustical environment," (Toronto, Canada), June 2013.

[277] K. Kato, T. Nagao, T. Yamanaka, K. Kawai, and K. Sakakibara, "Study on effect of room acoustics on timbral brightness of clarinet tones. part II: an acoustic interpretation and synthesis of analytical results," in *Proceedings of the 20th International Congress on Acoustics, Sydney*, 2010.

[278] D. Pelegrin Garcia, B. Smits, J. Brunskog, and C.-H. Jeong, "Vocal effort with changing talker-to-listener distance in different acoustic environments," *J. Acoust. Soc. Am*, vol. 129, no. 4, pp. 1981–1990, 2011.

[279] M. Kob, G. Behler, A. Kamprolf, O. Goldschmidt, and C. Neuxchaefer-Rube, "Experimental investigations of the influence of room acoustics on the teacher's voice," *Acoustical Science and Technology*, vol. 29, no. 1, 2008.

[280] V. Lyberg Åhlander, D. Pelegrí GarcÍa, S. Whitling, R. Rydell, and A. Löfqvist, "Teachers' voice use in teaching environments: A field study using ambulatory phonation monitor," *Journal of Voice*, pp. 841.e5–841.e15, June 2014.

[281] M. K. Miller and K.Verdolini, "Frequency and risk factors for voice problems in teachers of singing and control subjects," *Journal of Voice*, vol. 9, pp. 348–357, 1995.

[282] D. M. Howard and J. Angus, "Room acoustics: How they affect vocal production and perception," in *Occupational voice - care and cure* (P. H. Dejonckere, ed.), pp. 29–46, The Hague, The Netherlands: Kugler Publications, 2001.

[283] E. F. Chang, C. A. Niziolek, R. T. Knight, S. S. Nagarajan, and J. F. Houde, "Human cortical sensorimotor network underlying feedback control of vocal pitch," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 2653–2658, Feb. 2013.

[284] J. M. Zarate and R. J. Zatorre, "Experience-dependent neural substrates involved in vocal pitch regulation during singing," *NeuroImage*, vol. 40, pp. 1871–1887, May 2008.

[285] E. Borg, D. Gustafson, C. Bergkvist, and C. Wikstr om, "On the problem of listening while talking," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, 2009.

[286] S. Ternström, "Hearing myself with others: Sound levels in choral performance measured with separation of one's own voice from the rest of the choir," *Journal of Voice*, vol. 8, pp. 293–302, Dec. 1994.

[287] B. Blesser and L. Salter, *Spaces speak, are you listening* . MIT Press, 2007.

[288] A.-M. Laukkanen, N. P. Mickelson, M. Laitala, T. Syrj, A. Salo, and M. Sihvo, "Effects of hearfones on speaking and singing voice quality," *Journal of Voice*, vol. 18, no. 4, pp. 475–487, 2004.

[289] "National centre for early music." http://www.ncem.co.uk/.

[290] S. Bech, *Perceptual audio evaluation: theory, method and application.* Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2006.

[291] G. Martin, *Introduction to Sound Recording.* online, 2011.

[292] C. Fog and T. Pederson, "Tools for product optimisation," in *Human Centered Processes.*, (Brest. France,), 1999.

[293] G. Scavone, S. Lakatos, P. Cook, and C. Harbke, "Perceptual spaces for sound effects obtained with an interactive similarity rating program," in *Proceedings of International Symposium on Musical Acoustics*, 2001.

[294] G. P. Scavone, S. Lakatos, and C. R. Harbke, "The sonic mapper: an interactive program for obtaining similarity ratings with auditory stimuli," in *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan*, 2002.

[295] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of some listening test methods: A case study," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 356–364, 2005.

[296] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis: with Readings*. Englewood: Simon & Schuster Company, 4th edition ed., 1995.

[297] S. Choisel and F. Wickelmaier, "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *Journal of the Audio Engineering Society*, vol. 54, p. 815?826, 2006.

[298] J. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[299] R. Timmers, "From objective measurements to subjective ratings of similarity between expressive performances of music," in *Proceedings of the 8th International Conference on music Perception & Cognition*, 2004.

[300] Z. Schärer and S. Weinzierl, "Empirische fallstudie zum einfluss der raumakustik auf die musikalische interpretation," in *DAGA 2012 - Darmstadt*, 2012.

[301] S. Choisel and F. Wickelmaier, "Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference," *Journal of the Acoustical Society of America*, vol. 121, pp. 388–400, 2007.

[302] A. Hedblad, "Evaluation of musical feature extraction tools using perceptual ratings," 2011.

[303] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustic correlates of polyphonic timbre," *Music Perception*, vol. 27, no. 3, pp. 223–241, 2010.

[304] "Visual virtual microphone." http://david-mcgriffy.software.informer.com/.

[305] "Statistics toolbox." http://www.mathworks.co.uk/products/statistics/.

[306] J. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[307] K. Murphy, "HMM matlab toolbox."

[308] "Dynamic time warp in matlab." http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/.

[309] J. Gleiser, A. Friberg, and S. Granqvist, "A method for extracting vibrato parameters applied to violin performance," *TMH-QPSR*, vol. 4, pp. 39–44, 1998.

[310] J. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[311] H. Jers and S. Ternström, "Intonation analysis of a multi-channel choir recording.," *TMH-Quarterly Progress and Status Report*, vol. 47, no. 1, pp. 1–6, 2005.

[312] A. Buen, "How dry do the recordings for auralization need to be?," in *Proceedings of the Institute of Acoustics*, vol. 30, pp. 107–114, 2008.