UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Development of a Modulation Transfer Function-Based Method for Evaluating Bass Reproduction Accuracy in Professional Monitoring Loudspeakers

by

Lara Elizabeth Harris

Thesis for the degree of Doctor of Philosophy

July 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Doctor of Philosophy

DEVELOPMENT OF A MODULATION TRANSFER FUNCTION-BASED METHOD FOR EVALUATING BASS REPRODUCTION ACCURACY IN PROFESSIONAL MONITORING LOUDSPEAKERS

by Lara Elizabeth Harris

This study develops a measure which allows visual and quantitative assessment of bass reproduction accuracy in professional studio monitors. This type of loudspeaker must present mix engineers in particular with a faithful impression of recordings; they can then create an optimum musical balance between instruments that will translate well to other reproduction systems. Inaccurate monitors can lead to expensive remixing or a degraded musical experience for the end consumers, especially if the fundamental rhythm section has been poorly adjusted.

Existing work suggested that the Modulation Transfer Function (MTF) might be a more informative descriptor of bass reproduction accuracy than typical steady-state measures; it might therefore provide a grading system of performance that engineers could use when selecting suitable monitors for their work. The purpose of this project was to investigate the technique and i) develop an algorithm to summarise the critical aspects of monitor performance at low frequencies and ii) see whether it predicted the subjective impression of reproduction accuracy.

An algorithm was developed, considering different calculation methods and parameters to optimise it for low-frequency application with musical signals. It was applied to groups of loudspeaker models, simulating the responses of real mix monitors; then listening tests were conducted with these models reproducing music. The subjective and objective results were compared to see whether the algorithm would be a useful measure of monitor performance.

The algorithm successfully summarised behaviour of simulated and measured monitor responses; it described important factors such as extension and smoothness, and how the system responded to temporally-varying input signals. Based on ordinal data, the algorithm was found to predict all statistically significant judgements from listeners. These participants had demonstrated that they were accurate and consistent listeners, but found it difficult to reach consensus in some evaluations where the listening task required more skilled judgements of overall performance. It was concluded that the algorithm in its current form is effective and suitable for the intended application, but subjective evaluations of more complex alignments are needed from professional mix engineers; this might allow the numerical MTF scores to be graded against perceived bass reproduction accuracy, therefore enhancing the predictive power of the technique.

Contents

1 Introduction **21**

1.1 Project Background and Motivation . . . 21
1.1.1 Loudspeakers for Professional Monitoring . . . 21
1.1.2 Requirements for Accurate Bass Reproduction . . . 22
1.1.3 Assessing a Monitor's Performance . . . 23
1.1.4 Motivation for a New Measure . . . 26
1.2 Thesis Outline . . . 29
1.3 Original Contributions . . . 30

2 Developing an Algorithm **31**

2.1 The Modulation Transfer Function . . . 31
2.1.1 Defining 'The MTF' . . . 31
2.1.2 Developing an MTF-Based Method . . . 32
2.1.3 Use of the MTF for Loudspeaker Evaluation . . . 34
2.2 Aspects of Modulation . . . 35
2.2.1 Definition of Amplitude Modulation . . . 35
2.2.2 Requirements for Successful Amplitude Modulation . . . 36
2.3 Methods for MTF Calculation . . . 39
2.3.1 Method 1: Schroeder Expression . . . 39
2.3.2 Method 2: Formal Simulation . . . 41
2.3.3 Method 3: Filter-Modulate . . . 42
2.3.4 Method 4: Modulate-Filter . . . 43
2.4 Method Selection . . . 43
2.4.1 Test Parameters and Systems . . . 44
2.4.2 Evaluation of Test Systems . . . 45
2.4.3 Method Selection Through Further Investigation . . . 47
2.5 Optimisation I: Frequency Bands . . . 51
2.5.1 Defining the Range . . . 51
2.5.2 Defining the Band Parameters . . . 52
2.5.3 Selecting a Band Arrangement . . . 53
2.6 Optimisation II: Modulation Frequencies . . . 56
2.6.1 Spectrum Calculation . . . 56
2.6.2 Selected Modulation Frequencies . . . 59
2.7 Optimisation III: Modulating Function . . . 64
2.8 Comparing Parameters With Previous Work . . . 65
2.9 Algorithm Output . . . 65
2.9.1 Visual Representation . . . 66
2.10 Averaging for Consistency of Results . . . 66
2.11 Validation with Test Systems . . . 71
2.12 Summary of Algorithm Development . . . 75

3 Creating Virtual Loudspeakers **79**
3.1 Choosing a Virtual Approach . . . 79
3.1.1 Advantages of Simulation . . . 79
3.1.2 Response Equalisation Using DSP . . . 81
3.2 Woofer Modelling . . . 83
3.2.1 Lumped-Parameter Loudspeaker Model . . . 83
3.2.2 Model Validation . . . 84
3.3 Virtual Reproduction . . . 88
3.3.1 Measurement of the Experimental Loudspeaker . . . 88
3.3.2 Creating an Inverse Filter . . . 94
3.3.3 Creating a Response for Analysis . . . 99
3.3.4 Creating a Response for Listening . . . 100
3.4 Summary . . . 101

4 Objective Evaluation of Loudspeaker Models **103**
4.1 Model Groups . . . 103
4.1.1 Group Development . . . 103
4.1.2 Group Composition . . . 104
4.1.3 Design Strategy for Groups I and II . . . 105
4.2 Group I Models . . . 106
4.2.1 Response Measurement: Group I . . . 107
4.2.2 MTF Assessment: Group I . . . 110
4.2.3 Discussion of Group I Objective Results . . . 111
4.3 Group II Models . . . 113
4.3.1 Response Measurement: Group II . . . 114
4.3.2 MTF Assessment: Group II . . . 115
4.3.3 Discussion of Group II Objective Results . . . 116
4.4 Collated Results: Groups I and II . . . 117
4.5 Group III Models . . . 118
4.5.1 Generating Artificial Systems . . . 119
4.5.2 Response Measurement: Group III . . . 121
4.5.3 MTF Assessment: Group III . . . 122
4.5.4 Discussion of Group III Objective Results . . . 123
4.6 Summary of Objective Evaluation . . . 124

5 Listening Test Design **127**
5.1 Experimental Aims and Requirements . . . 127
5.2 Developing a Test Strategy . . . 129
5.2.1 Method Definition . . . 129
5.2.2 Response Bias . . . 131
5.2.3 Ease of Participation . . . 132
5.2.4 Duration and Efficiency . . . 133
5.2.5 Chosen Methodology . . . 133
5.2.6 Assumptions in the Chosen Procedure . . . 136
5.3 Music Selection and Processing . . . 136

5.3.1 Selection and Extraction . . . 136
5.3.2 Loudness Matching . . . 140
5.3.3 Selected Extracts . . . 140
5.4 Software Implementation . . . 146
5.4.1 Generating User Playlists . . . 147
5.4.2 Experimental Interface . . . 148
5.5 Execution: Hardware and Practical Matters . . . 149
5.5.1 Switchbox Testing . . . 149
5.5.2 Calibration of Music Levels . . . 152
5.6 Summary of Listening Test Design . . . 152

6 Statistical Methods for Analysis of Subjective Data 155
6.1 Primary Aims for Data Analysis . . . 155
6.2 Identifying an Appropriate Class of Methods . . . 156
6.2.1 Classification . . . 157
6.2.2 Selection . . . 157
6.3 Investigating Programme Dependence . . . 157
6.3.1 The Chi-Square (χ^2) Test for Independence . . . 158
6.3.2 Calculating the Value of χ^2 . . . 159
6.3.3 Summary of Procedure for Detecting Programme Dependence . . . 160
6.4 A/B Pair Results . . . 161
6.4.1 Hypothesis Testing Using the Binomial Distribution . . . 161
6.4.2 Analysing One Pair . . . 162
6.4.3 Extending to Multiple-Pair Hypothesis Tests . . . 162
6.4.4 Directional Hypothesis Testing . . . 164
6.4.5 Summary of Procedure for Multiple A/B Pair Analysis . . . 164
6.5 Post-Screening by Individual Performance . . . 165
6.5.1 Requirements and Assumptions . . . 165
6.5.2 Summary of Procedure for Post-Screening . . . 166
6.6 Summary of Statistical Methods . . . 167

7 Subjective Evaluation of Loudspeaker Models 169
7.1 Recap of Procedure . . . 169
7.2 Listening Test I: Evaluation of Group I Models . . . 170
7.2.1 Extract Analysis (Programme Dependence) . . . 170
7.2.2 Post-Screening (Intra-Listener Performance) . . . 171
7.2.3 Pairwise Results Based on Total Sample (All Listeners) . . . 171
7.2.4 Pairwise Results Based on Post-Screened Sample . . . 174
7.2.5 Listening Test I Results Summary . . . 175
7.3 Listening Test II: Evaluation of Group II Models . . . 177
7.3.1 Extract Analysis (Programme Dependence) . . . 178
7.3.2 Post-Screening (Intra-Listener Performance) . . . 178
7.3.3 Pairwise Results Based on Total Sample (All Listeners) . . . 179
7.3.4 Pairwise Results Based on Post-Screened Sample . . . 181
7.3.5 Listening Test II Results Summary . . . 182

7.4 Listening Test III: Evaluation of Group III Models . . . 184
7.4.1 Extract Analysis (Programme Dependence) . . . 184
7.4.2 Post-Screening (Intra-Listener Performance) . . . 186
7.4.3 Pairwise Results Based on Total Sample (All Listeners) . . . 187
7.4.4 Pairwise Results Based on Post-Screened Sample . . . 189
7.4.5 Listener Comments . . . 190
7.4.6 Listening Test III Results Summary . . . 191
7.5 Summary and Discussion of Subjective Evaluation . . . 192

8 Assessment of the MTF-Based Method for Loudspeaker Evaluation 197
8.1 Comparing Significant Pair Results (Listeners vs Algorithm) . . . 197
8.1.1 Pairwise Outcomes . . . 197
8.1.2 Indirect Ranking . . . 199
8.2 Comparison With Non-Significant Pair Results . . . 201
8.2.1 Two Pairs from Listening Test I . . . 202
8.2.2 Two Pairs from Listening Test II . . . 205
8.2.3 Conclusions Following Investigation of Non-Significant Results . . . 207
8.3 Subjective Modification of Band Scores . . . 208
8.3.1 Adjustment for Hearing Threshold . . . 208
8.3.2 Adjustment for Programme Balance . . . 211
8.4 Comparison With Other Methods . . . 213
8.4.1 Possible Alternatives . . . 213
8.4.2 Demonstrating Alternatives . . . 215
8.4.3 Summary of Alternative Methods . . . 222
8.5 Summary . . . 223

9 Conclusions and Suggestions for Further Work 227
9.1 Conclusions . . . 227
9.1.1 Algorithm Development and Objective Validation . . . 228
9.1.2 Collection of Subjective Data and Comparison with Algorithm Results . . . 233
9.2 Suggestions for Further Work . . . 239

Appendices 242

A Results from Comparison of MTF Methods 243

B Modulation Spectrum: Tracklist and Genre Classifications 244

C Derivation of the Lumped Parameter Model 249

D Comparison of Large- and Small-Chamber Loudspeaker Measurements 252

E List of Experimental Equipment 254

F Model Parameters 255

G MTF Results Listing 256

H High-Pass Filter Transfer Functions **260**

I Extract Spectrograms **261**

J Listening Test vs MTF Results: No Post-Screening **263**

K Additional Comparison of Non-Significant Pair Results **264**

L Group II Normalised Schroeder Results **268**

References **271**

List of Figures

1 Illustrating the principle of the MTF in acoustic measurement 32
2 AM envelopes for three modulation factors 37
3 Amplitude modulation of a band-limited signal and carrier sinusoid 38
4 Test systems: Complex frequency responses 45
5 Comparing Methods 3 and 4: Input and output 50
6 Six frequency band arrangements 54
7 Comparison of MTF results across test band sets 55
8 Musical extract envelope example 57
9 Single-segment modulation spectrum 58
10 Envelope spectra from two different musical extracts 58
11 Distribution of genres used to calculate musical modulation frequency spectrum . . 59
12 Musical modulation frequency spectrum 60
13 Genre-specific envelope spectra 61
14 Generic musical envelope spectrum up to 20 Hz. 62
15 Standard deviation of mean score from very low modulation frequencies 63
16 Modulating functions and modulated waveforms 64
17 Comparing modulation frequencies and analysis range with previous work 65
18 Illustrating the averaging process 68
19 Input and output envelopes formed from a single noise iteration 69
20 Input and output envelopes formed from 100 noise iterations 69
21 Mean-score standard deviation across 30 evaluations of the same system 70
22 Algorithm validation: MTF results for four loudspeakers 72
23 Waterfall plots for validation systems 73
24 Frequency response magnitude and phase for validation systems 74
25 Example screenshot from the *WooferMaker* GUI 84
26 Comparing two models with real monitors 85
27 Effects of integer- and fractional-sample delays in frequency and time domain . . . 87
28 Equipment layout for loudspeaker measurement 89
29 Experimental loudspeaker and measurement microphone positions 90
30 Recording equipment 91
31 Measurement system response 91
32 Three measurements of the experimental loudspeaker 93
33 Inverse filter creation stage 1: Experimental loudspeaker FRF 94
34 Inverse filter creation stage 2: Time-of-flight removed 95
35 Effect of windowing with linear magnitude 96
36 Effects of smoothing principal phase 97
37 Inverse filter creation stage 3: Smoothed response 98
38 Inverse filter creation stage 4: Inverse filter with gain capping 99
39 Virtual loudspeaker response simulation 100
40 Group I: Measured and simulated responses 108
41 Group I MTF results (mean-band) 110
42 Group I MTF results (intensity image) 111

43 Group II: Measured and simulated responses . . . 114
44 Group II MTF results (mean-band) . . . 115
45 Group II MTF results (intensity image) . . . 116
46 Changes in mean score with low frequency extension . . . 118
47 Group III Impulse response comparison . . . 120
48 Group III: Measured and simulated responses . . . 121
49 Group III MTF results (mean-band) . . . 122
50 Group III MTF results (intensity image) . . . 123
51 Listening test I extract characteristics . . . 142
52 Listening test II extract characteristics . . . 144
53 Listening test III extract characteristics . . . 146
54 Example playback matrices . . . 148
55 Example screenshots from the listening test GUI . . . 149
56 Three-way switchbox . . . 150
57 Switchbox analysis spectra . . . 151
58 Example crosstabulation . . . 159
59 Types of listening test trial . . . 169
60 Group I extract distribution . . . 170
61 Group I hidden reference response distribution . . . 171
62 Group I pair result plots . . . 173
63 Group I post-screened pair result plots . . . 175
64 Group II extract distribution . . . 178
65 Group II hidden reference response distribution . . . 179
66 Group II pair result plots . . . 180
67 Group II post-screened pair result plots . . . 181
68 Group III extract distribution . . . 185
69 Chosen arrangement for collapsing extract groups . . . 186
70 Group III hidden reference response distribution . . . 187
71 Group III pair result plots . . . 188
72 Group III post-screened pair result plots . . . 189
73 Diagraphs for all pair results in listening tests I and II . . . 200
74 Band-mean levels for averaged extract spectra used in listening tests I and II . . . 202
75 Group I: R vs F. Magnitude and phase . . . 203
76 Group I: C vs G. Magnitude and phase . . . 204
77 Group II: D vs G. Magnitude and phase . . . 205
78 Group II: F vs G. Magnitude and phase . . . 206
79 Comparison of the MAF against R and F assuming two reproduction SPLs . . . 210
80 Loudspeaker D and G (Group II): Comparing magnitude responses . . . 216
81 Loudspeaker D and G (Group II): Comparing phase responses . . . 217
82 Loudspeaker D and G (Group II): Comparing impulse responses . . . 218
83 Loudspeaker D and G (Group II): Comparing waterfall plots . . . 219
84 Loudspeaker D and G (Group II): Comparing intensity images . . . 220
85 Normalised Schroeder method: Comparison of intensity images . . . 221
86 Three small-chamber loudspeaker measurements . . . 252

87 Comparing large- and small-chamber loudspeaker measurements 253
88 Spectrograms for listening test I extracts 261
89 Spectrograms for listening test II extracts 262
90 Spectrograms for listening test III extracts 262
91 Listening test I responses: C vs F 264
92 Listening test II responses: D vs E 265
93 Listening test II responses: E vs F 266
94 Listening test II responses: E vs G 266
95 Listening test II responses: D vs F 267
96 Normalised Schroeder method: Group II MTF results (mean-band) 268
97 Normalised Schroeder method: Group II MTF results (intensity image) 269

List of Tables

1 Comparison of results from four MTF methods 45
2 Method 1 results after normalisation 48
3 Summary of test band set parameters 53
4 Key design features for Group I virtual loudspeakers 107
5 Design (target) features for Group II virtual loudspeakers 113
6 Approximations of filter orders to real loudspeaker systems 119
7 Description of listening test I extracts 141
8 Description of listening test II extracts 143
9 Description of listening test III extracts 145
10 Extract playback SPLs 152
11 Group I pair results 174
12 Group I post-screened pair results 175
13 Group II pair results 180
14 Group II post-screened pair results 182
15 Group III pair results 189
16 Group III post-screened pair results 190
17 Comparison of listening test pair results with MTF mean scores 198
18 Difference in $\overline{M}$ scores for pairs with no directional outcome 199
19 Arrow counts derived from diagraphs 201
20 Comparison of algorithm and listener-derived loudspeaker rankings 201
21 MAF-corrected $\overline{M}$ scores for Group I and II 209
22 Comparison of corrected values assuming a reproduction level of 70 dB SPL 210
23 Comparison of mean MTF scores before and after programme weighting 211
24 Summary of method comparisons 222
25 MTF results for System 1 243
26 MTF results for System 2 243
27 Extracts used for analysis of musical modulation frequencies 248
28 Equipment used for measurement and experimentation 254
29 Group I model input parameters 255
30 Group II model input parameters 255

31 MTF matrix results for Group I models 257
32 MTF matrix results for Group II models 258
33 MTF matrix results for Group III models 259
34 Significant pair outcomes: listening tests vs algorithm 263

Author's Declaration

I, Lara Elizabeth Harris, declare that the thesis entitled:

Development of a Modulation Transfer Function-Based Method for Evaluating Bass Reproduction Accuracy in Professional Monitoring Loudspeakers

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Parts of this work have been published as:

- L. Harris and K. Holland. Evaluating Loudspeaker Quality at Low Frequencies: Optimisation of a Music-Focussed Modulation Transfer Function Technique. *Proceedings of the Institute of Acoustics.* Vol.30(4)*, 2009*; pp.76–88.
- L. Harris and K. Holland. Using Statistics to Analyse Listening Test Data: Some Sources of Advice for Non-Statisticians. *Proceedings of the Institute of Acoustics.* Vol. 31(4), *2010*; pp.294–309.

Signed:

Dated:

Acknowledgements

This list is not exhaustive, but I must limit my thanks and acknowledgements to a small number of people who had a direct influence on the project and this thesis. These are listed in no particular order, except for the first three. As is tradition, I must mention my supervisor Dr Keith Holland first; without his guidance, expertise, encouragement, and support, I could not have conducted the research. Second, I must thank and acknowledge Andrew Harper. Without his unerring reassurance, domestic assistance, and insightful suggestions for improvements, I could not have written the thesis. His penchant for buying expensive textbooks also turned out to be extremely useful. Philip Newell provided such compelling motivation for the project, and such engaging written and verbal accounts of it, that I must acknowledge and thank him for his contribution.

Thomas King gave me confidence (no pun intended) in my statistical methods and specifically helped me with the Sidak correction. I must acknowledge my friend and colleague Dr Delphine Nourzad. She motivated me to make a huge change in my life circumstances and actually write up the work that I'd abandoned for so long. She was always there when I needed advice and consolation about the writing and submission process; she also let me copy her text for the author's declaration. My friend Dr Filippo Fazi let me use some of his experimental equipment, and was rather previous, some might say optimistic, in acknowledging me as Dr Harris in his thesis; I hope that credit does not remain a stain on his otherwise flawless tome (at least I assume it to be, given that I can barely understand the title, let alone the content). Susan Brindle helped make lots of things happen, and she, along with Keith, must be thanked for encouraging me not to quit ISVR within the first semester when I realised I was in well over my head. The technicians and consulting staff at ISVR helped so much with practical matters and, along with other nameless strapping individuals, provided vital assistance in lugging my experimental equipment around. Thanks also to Dr. Rod Self; he provided me with much needed employment that enabled me to hang around the university for a bit longer to conduct my final experimental work. Unfortunately, it gave me a very good excuse to be doing other things when I should have been working on this thesis, but I think it was worth it.

Additional thanks for personal assistance go to Leigh and Zibi Dombek, not least for eventually adopting me and saving me from complete isolation from the outside world whilst writing. I must also thank Irena and John Roden for their support, and apologise to them for not finally finishing this project whilst they were still alive.

This thesis was written in Lyx*with Jabref†. I am so thankful that these programs exist, and that AH recommended them. I think having to produce this document in the ubiquitous Windows-based word processor would have been the straw that broke the camel's back.

*http://www.lyx.org/
†http://jabref.sourceforge.net/index.php

Nomenclature

Variables and abbreviations used regularly throughout the thesis are listed here.

$*$ Convolution

α Significance level for hypothesis testing

$\alpha_{\mathbf{func}}$ Functional value of α, corrected for multiple-pair comparisons

β Bandwidth

β_e Effective bandwidth: apparent bandwidth of test signal after modulation.

c' Critical count

f_c Cut-off frequency (half-power point, -3 dB relative to reference level)

$f_{\mathbf{m}}$ Modulation frequency

$f_{\mathbf{cf}}$ Band centre frequency

m Modulation index

$\overline{m}$ Mean modulation index, single band

$\overline{M}$ Mean modulation index, full matrix

$\overline{M}_{\Delta}$ Difference in mean modulation index

2AFC Two-alternative forced choice

AM Amplitude modulation

DSB-LC Double sideband large carrier

DSB-SC Double sideband suppressed carrier

DSP Digital signal processing

FRF Frequency response function

GUI Graphical user interface

HPF High-pass filter

LF Low-frequency

LPF Low-pass filter

MTF Modulation transfer function

SNR Signal-to-noise ratio

SPL Sound pressure level

STI Speech transmission index

1 Introduction

This chapter gives an introduction to the project and thesis. Section 1.1 outlines the research problem and motivation for the project. Section 1.2 describes the structure for the rest of the thesis, and section 1.3 summarises the original contributions arising from this work.

1.1 Project Background and Motivation

This section draws on relevant publications to present a background to the research question. It is described how loudspeakers are used professionally to make recorded music, why good bass reproduction is important, how to infer from measurements whether a loudspeaker will be good at reproducing bass, and the consequences if it isn't. The discussion is followed by the motivation for this project and its main objectives.

1.1.1 Loudspeakers for Professional Monitoring

Loudspeakers are used by professional audio engineers at three fundamental stages in making recorded music [1]. The first is during recording to capture the initial performance of the musicians. In the next stage, the level of each individual instrument is adjusted to give a balanced mix of sounds. In the final stage, a master mix is created; engineers apply effects, and check the broader frequency balance and overall levels to create a performance that should translate well to a variety of domestic reproduction systems. At each of these stages, the engineers use loudspeakers to monitor the recorded sound and make changes before it goes out to consumers, the listening public.

During recording, the engineer needs loudspeakers that reveal potential issues with the sound whilst they are still in a position to make changes; they need to know if they've obtained the best possible sound 'at source'. Loudspeakers used at this stage are typically large with high power handling capabilities; they must be able to reproduce the recorded content at high levels, in excess of 100 dB SPL, to keep the performers enthused and energetic, whilst still being subtle and revealing enough to allow the recording engineer to judge critical aspects of the material. These monitors may be flush mounted in the wall of the control room, several metres from the listening position.

The mixing of a recording is where the engineer creates a musical and timbral balance of each individual instrument captured at the recording stage. There is some artistic interpretation on the part of the engineer to ensure that the emotion and impact intended by the musicians is conveyed to the final recording; engineers rely on the mix monitors to help them achieve this. Mixing commonly takes place on smaller monitors, at lower sound levels than during recording, around 75-85 dB SPL. It is also common for these monitors to be mounted at a closer distance, around 1.5 m from the listening position, on stands or the console; this allows finer judgement without the interfering influence of added reverberation from the listening environment [2]. Good mix monitors should provide the engineer with a response that is 'flat and fast' in the mounted position [1]; these two characteristics help the engineer in creating a mix that has an accurate and realistic balance between instruments.

Mastering is the last stage of creating a recording before it goes out for production; the engineer makes a final check for overall frequency balance and dynamics, and can make some

limited adjustments if results from the mixing stage are found to require it. Dedicated mastering studios tend to feature monitors more like audiophile high-fidelity loudspeakers, capable of revealing very subtle detail in recordings to the lowest frequencies; they are not expected to reproduce demanding content at very high levels, such as solo kick drums, but do need to be highly revealing of all aspects of the music that may have been recorded or, more typically, mixed on less extended loudspeakers. Mastering engineers therefore typically choose very large, sealed cabinet loudspeakers, or reflex cabinets with very low tuning frequencies, to allow the most critical evaluation. The mastering engineer must be able to hear all aspects of the recording, even elements that the vast majority of the listening public may never detect.

Although studio monitors appear to have slightly different requirements depending on which part of the recording chain they are used in, they are all designed to fundamentally perform the same function; they are professional tools that allow skilled engineers to carry out quality control and achieve aesthetic excellence in the product they are helping to create. Toole [3] described neutral monitors as a 'transparent window into the art'. Given suitable monitors then, recording engineers and producers can use their technical skills and expertise to create a recording that conveys the spirit and emotion of the music, as intended by the musicians [4, 5]. In working towards this goal, engineers are not listening for pleasure; listening for errors is the primary goal of professional monitors. The engineers are highly practised in the assessment of quality, being able to detect subtle change in programme material with great consistency [6]; they are concentrating on specific aspects of the musical performance and recording quality, listening in a way that allows them to evaluate and interpret its characteristics and make decisions regarding any problematic features such as resonances, lack of definition, and imbalances in level or frequency [4, 7]. Loudspeakers chosen by engineers to create a well-balanced mix are rarely used in domestic situations; they are not designed to flatter the audible presentation, but to draw attention to any defects before the product goes out to consumers [8].

1.1.2 Requirements for Accurate Bass Reproduction

This project focuses only on the reproduction quality of low-frequency musical content; this description will be used synonymously with the musical term 'bass'. The region covered by this term is not defined in a standardised way, but Colloms [9] focussed on the region 20-160 Hz. He stressed that this is a full three octaves of the musical spectrum.

For consumers of recorded music, the accurate reproduction of bass content contributes towards involvement, a feeling of energy, musicality, and, ultimately, enjoyment. Harley [10] described the whole-body experience that can be obtained from the synergistic combination of a kick drum's transient with the attack of bass guitar strings. This provides the tonal foundation and rhythmic drive in many types of music [11]; rock and pop could not exist without it. Although much recorded music contains little content below 40 Hz, when music does contain content in the lowest octaves, it can make a huge impact subjectively [9, 12]. But the quantity of bass does not necessarily relate to its quality; the subjective characterisation of bass quality is a combination of both level and time coherence with the rest of the frequency range [9]. When reproduced without definition and articulation, the low frequency content may be heard but seems disconnected from what is happening above it. As a result, the whole musical performance loses its impact [10]. Andrews [13] refers to two types of reproduction: reality, and soft focus; in the latter case, definition and dimension will be missing, and musical instruments will be

confused, presenting the listener with a form of 'audio mush'. If reproduced correctly, the bass will be full and deep, revealing the weight and scale of the musical content whilst still translating the dynamic subtleties that can make a song exciting and engaging, perhaps invoking the desire to dance, or at least nod the head [14, 15]. In order to convey this effect to the end consumer of musical content, the elements that create it must have been adequately monitored and preserved at each stage of the recording chain [16].

1.1.3 Assessing a Monitor's Performance

As discussed in the previous section, accurate monitors are required to ensure correct balancing of a recording's content so that consumers can receive an engaging and powerful musical experience. Specific aspects of studio monitors that determine performance are now discussed.

1.1.3.1 Variety of Alignments The low-frequency alignment of a moving coil loudspeaker is typically characterised by the roll-off frequency, the -3 dB point compared to passband output level, and the steepness of roll-off below this point, the rate at which lower frequencies are attenuated. This is a somewhat simplified description, as the alignment is actually defined by a 'constellation' of parameters that combine in a specific way to determine the overall shape of the response [17]. The parameters that affect a loudspeaker's alignment are discussed in more detail in section 3.2.1, but the discussion here will be limited to one key factor influencing the alignment: the design strategy of the loudspeaker cabinet. The most basic design approach uses a sealed cabinet, giving a 2^{nd}-order roll off, 12 dB/octave [18]. The other common design strategy is reflex loading, where a port is added to the cabinet. This contains a mass of air that resonates at a specific tuned frequency, usually chosen to boost output near the frequency where the pressure level from the diaphragm naturally begins to reduce. This type of alignment produces a steeper roll-off in the low frequencies, as the diaphragm and port output are in antiphase below the resonance point; this is typically 4^{th} order, or 24 dB/octave, for a standard reflex design [19].

Reflex loading is often used because it extends bass output over a narrow region; the magnitude of the pressure response can be seen to extend to a lower frequency, then drops off quickly compared to a sealed-box equivalent. This is a compromise to achieve greater bass output at high sound pressure levels, as is required of studio monitors; it is a common approach in small to medium-sized monitors, typically used for mixing, as manufacturers aim to achieve lots of bass from a small box. A smaller box requires the use of smaller drive units, which have to travel further to move an equivalent volume of air compared to diaphragms with a larger radiating area. In a bass-reflex cabinet, air moves in and out of the port freely below its resonance; the 'air spring' that prevents excessive diaphragm excursion above this frequency is lost, and the drive unit may be damaged if the input signal induces a large displacement, i.e. contains high-level low frequencies. Therefore, this type of monitor may also have added protection filters to stop large excursion at very low frequencies from damaging the suspension and causing excessive distortion. These analogue high-pass filters remove very low frequency content from the signal; the drive unit is thus stopped from moving too far by electrical rather than mechanical means. Such filters add to the inherent steep roll-off of a bass reflex design to further increase phase shift at low frequencies, therefore having a greater impact on the loudspeaker's transient response [20]. Roll-offs can exceed 6^{th} order in monitors that have ported cabinets with added electrical filtering to limit diaphragm displacement [21].

Even from this very basic summary of design strategies, it can be appreciated that professional studio loudspeakers will exhibit a range of different behaviours at low frequencies. The distinction between different roll-off slopes is significant in characterising the main design strategies used in professional monitors for two main reasons. Firstly, it determines the quantity of bass a speaker can reproduce; ported monitors will output low frequencies at a higher level over a narrow range of frequencies; a sealed monitor with the same roll-off frequency will offer less output in this region but will still be able to reproduce content well below this point. The steepness of the roll-off is also significant in understanding another aspect of loudspeaker performance, as, in a minimum phase system, the magnitude and phase response are interdependent [22]; they are uniquely related such that one may be calculated from the other (though both are required to confirm whether the assumption of minimum phase is correct) [23, 24]. Multi-way loudspeakers are accepted to be non-minimum phase systems [16, 25, 26]. The transfer function can be represented as a cascade of a minimum phase and all-pass stage that affects only the system's phase response; introduction of this 'excess phase' can produce audible effects that cannot be predicted from the amplitude response alone [27, 28]. However, loudspeakers are generally assumed to be minimum phase systems if considering only the response at low frequencies [18, 29, 30]. A loudspeaker with a steeper roll-off at low frequencies will therefore have greater phase distortion through the bass region [31].

Preis [22, 29] presented a thorough and detailed theoretical discussion of the implications of phase anomalies in different types of audio system; he concluded that careful consideration of the steady-state magnitude, phase, and phase slope measurements of a system can indicate a lot about its transient behaviour. A linear phase shift, one that is proportional to frequency, is equivalent to a pure delay [32, 33]. Phase shifts that affect some frequencies more than others, such as rapid roll-offs in a loudspeaker's alignment, change the relationship between individual components in a signal. The delay experienced in a band of frequencies due to a non-linear phase function is known as group delay [32], also called envelope delay [23]. This frequency-dependent phase shift does not delay the signal as a whole, but rather, elements within it [28]. In the time domain, the waveform of the signal is seen to have changed; the output from the system is not simply an amplitude-scaled version of the signal that went in. This effect is called dispersion [32, 33], where different parts of a signal arrive at different times. Heyser [23] illustrated this effect as a frequency-dependent placement of sources; regions with higher group delay, such as in the low-frequency roll-off of loudspeakers, could be imagined as having the transducers placed further back in space, therefore increasing the arrival time to listeners relative to the rest of the spectrum. In speech, this leads to a loss of intelligibility; in music, bass notes lag behind the rest of the music.

Thus, it may be concluded that if loudspeakers differ greatly in their low-frequency alignment, they will not affect the musical signals passing through them in the same way.

1.1.3.2 Differing Measured Behaviour of Alignments

In order to produce mixes of a consistently high quality, there must be a degree of consistency across monitors, regardless of their design philosophy [6]. It is rare, however, to find an acceptable level of consistency, even within loudspeakers designed to perform the same professional function. Newell *et al.* [8, 21, 34] conducted in-depth comparative studies of medium-sized professional studio monitors, analysing performance in both the frequency and the time domain through a combination of different

measures. The tested units showed a range of low frequency alignments, and included sealed-box and reflex-loaded systems, both with and without protection filters. It was found that the monitors with a flat and extended frequency characteristic did not necessarily perform well in response to dynamic signals. The loudspeakers with steep roll-offs exhibited long and uneven decays in the low frequencies; waterfall plots clearly showed a dominant resonance, or ringing, around the roll-off frequency. Resonant systems take time to come to rest after excitation; therefore, the reflex enclosures used in smaller monitors to increase output at lower frequencies also compromise transient performance. Impulsive signals will be smeared in time as bass notes arrive late compared to other signal content [26], and low level detail in the musical signal may be masked by the extended ringing on of one part of the response where the resonance occurs [34, 35]. The waterfall plots of the sealed-box monitors under test, having less steep attenuation at low frequencies, did not show such dominant resonant tails [21]. One of the best performing systems in this respect was the Yamaha NS10M, a small and relatively cheap monitor that was extremely popular for mixing pop and rock recordings for over 20 years [8]. Although its anechoic magnitude response showed it to be neither flat nor particularly extended, waterfall plots showed a rapid decay that was nearly uniform with frequency. Consideration of both amplitude and time response of this loudspeaker therefore showed that in one aspect it was a poor performer compared to equivalent models, but in another it shared qualities with much larger, higher-quality studio monitors.

1.1.3.3 Consequences It has been established that monitors with different low-frequency alignments will show a variety of measured behaviours. This suggests that they will present a mixing engineer with audibly different representations of the same recorded material [21]. The implication of these differences to professional monitoring is now considered.

Gaining bass extension in smaller loudspeakers at the expense of transient accuracy may be acceptable for domestic listening, or possibly even as a compromise for recording monitors, where reproduced content needs to be loud and low to enthuse the performers. However, for mixing and mastering, a more critical impression of the musical content is required [35]. A very flat and extended bass response is not necessarily a main requirement for all professional mixing monitors. It has been shown that mix engineers in particular view amplitude uniformity as a secondary priority [36]; although studied in relation to the rooms used for monitoring, rather than the loudspeakers specifically, clarity of the reproduced sound whilst mixing was considered to be paramount. A separate study on a similar topic also showed that fast low-frequency decays were more important than extended magnitude responses for the kind of critical listening conditions that mix engineers require [7]; it was later demonstrated in an environment typical of studio control rooms that the decay time alone of low-frequency resonances can lead to perceived tonal changes or colouration of musical signals [37]. Experienced engineers can mentally compensate for any deficiencies in bass output as long as the deviation from a flat magnitude is smooth, and if, as is usually the case, they are well accustomed to the loudspeakers they are working on. Abrupt response changes introduce colouration to musical content in the region of the disturbance, and are not so easily compensated for; they are usually accompanied by time response errors, and even experienced professionals cannot imagine what the correct reproduction should sound like if presented with material affected in this way [38, 39].

Poor timing in a loudspeaker's low-frequency response is detrimental to creating a

well-balanced mix for a number of reasons. It affects timbre, rhythm, and even pitch of the bass instruments, and there is evidence that perception of these effects is greater at lower frequencies and higher SPLs [26, 28, 40–42]. Late arrival and long decay of low frequencies both compromise the punch of a rhythm section; basslines may lack speed and fluidity, sounding sluggish, sloppy, and lacking in dynamics as the loudspeaker rings on at certain frequencies and therefore masks subsequent waveforms [6, 13, 43]. Small monitors using resonances to extend their low-frequency response colour the timbral balance between bass instruments; musical notes exciting ringing around the roll-off frequency may change the perceived pitch from that of the note to that of the loudspeaker's dominant resonance, sometimes described as 'one-note bass' [10, 21]. This signal-dependent colouration makes it difficult for a mix engineer to make judgements about the correct balance between instruments of a transient and steady-state nature [1]. It is impossible to tell which aspects of the music are due to the instruments, and which are due to time response errors in the monitors. Mixes produced on fast-decaying monitors tend to translate more successfully to other reproduction systems because the engineer is presented with a more faithful representation of the true instrumental balance contained within the recording. Very fast low-frequency decay is needed for mixing rock and pop music in particular, because time response errors particularly affect the perceived balance of bass guitar and bass drum. These instruments are fundamental aspects of the rhythm section in many types of music, but their dominant content is in the region where resonances in smaller monitors are likely to occur due to the tuning of ports around the natural frequency of the drive unit. Monitors with an extended response but poor transient behaviour may therefore yield a mix biased towards the bass drum; the comparatively steady-state notes of the bass guitar will sound louder, and the engineer will try to compensate for this accordingly in the mix [34].

Mastering engineers have limited abilities to adjust recordings once they leave the mixing stage. As mastering engineers seem to favour large sealed cabinets or ported systems with very low-tuned resonances, they are able to detect when inappropriate mixes have been created, but can do little to correct it. If the mix is judged too bass-heavy overall, perhaps due to having been mixed on sealed-box speakers with reduced low-frequency output, this can be fixed; as long as the balance of individual instruments within the same region has been correctly set, some global equalisation (EQ) can restore an appropriate spectral balance [38]. If an imbalance occurs between instruments in the same frequency range, such as bass guitar and bass drum, a broad bass boost cannot adjust one instrument without affecting the other to some extent. Mastering engineers can therefore do little terms of EQ or dynamic control to correct misjudgements in the mix that are due to problems in a monitor's time response at low frequencies [4, 8]. Mistakes cannot be corrected at this stage, only replicated and distributed to consumers. The only options then are to send out a poor product, or to go back and remix the recording. Selection of monitors with this type of response error can therefore lead to very time-consuming, and very expensive, mistakes [4].

1.1.4 Motivation for a New Measure

Toole [3] described the use of inaccurate loudspeakers to make reference recordings as equivalent to making a technical measurement with an uncalibrated instrument. This leads to a 'circle of confusion', where loudspeakers are used to make recordings, which are then used to demonstrate and evaluate other audio products, which are then used to make recordings, and so on. Viewing

the issue in this way helps to illustrate the nature of the problem. A studio engineer is creating the reference; if they are doing so on an 'uncalibrated instrument', they will not know it at the time, and may only realise that something is wrong when comparing against similar recordings created elsewhere. If they are making mixes that do not transfer well to other systems, or need frequent remixing, it would be useful for them to know whether the equipment they are using is part of the problem; armed with all the relevant information, they can then make a decision as to whether they should buy new monitors or consider a change of career. Since the introduction of digital technology, the cost of a full recording system has dropped in price considerably, but the purchase of high-quality monitors remains a substantial investment [1]. An engineer is therefore faced with an important business decision when selecting monitors that will permit accurate mixes to be created. However, manufacturers do not usually provide potential customers with all the information they may need in order to make a considered purchase.

1.1.4.1 Existing Measures As far back as 1948, the British Broadcasting Corporation (BBC) was selecting professional monitors on the basis of both amplitude and time response [44]. From the BBC's experience in broadcasting a wide range of material, it had been observed that poor monitors could lead engineers to produce programme balance that would not translate well to other systems, whereas good monitors all produced similarly balanced recordings. The loudspeakers, and the equipment to measure them, were crude by today's standards, but it was reported that whilst a loudspeaker's magnitude response should be reasonably uniform, it was not a criterion on which to judge loudspeakers; there was, however, a 'good measure of agreement between the transient troubles and the faults which are audible' [45]. More than 65 years later, it is still common for manufacturers to focus on the first of these aspects of performance. Along with figures for distortion and power handling, the 'tech specs' summaries always state the frequency response limits, and reputable manufacturers may even state the magnitude values upon which these are based [46–52]. The more detailed technical specifications and data sheets, if available, rarely provide further insight; along with the already-stated frequency limits, they may contain attractive pictures of the product, reviews, a frequency response (magnitude) plot, and more details relating to power handling and electrical characteristics [53–58]. Whilst unusual, some manufacturers do provide potential customers with detailed technical information [59], but even plots of cumulative spectral decay and group delay are not especially useful when trying to understand the perceived effect a monitor will have on a signal passing through it, at least not without in-depth study of the associated topics.

The magnitude response has become the measure of choice for good reason; it is simple to measure, intuitive, easily depicted on a data sheet, and has performed well under investigation of its relation to subjective preference [60, 61]. For most domestic applications, listening will be for pleasure, and selection of loudspeakers based on personal preference is perfectly appropriate. For professional monitoring, as already discussed, a more informed choice is required. The introduction of resonant elements through reflex loading and added protection filters degrades the phase, and therefore time response of a loudspeaker [21]. Such effects cannot be fully understood from inspection of a magnitude response alone. Even if a studio engineer was aware that steep roll-offs in a magnitude response indicate large, frequency-dependent shifts in phase, and therefore group delay, it would still be almost impossible for them to directly interpret the audible impact this will have when they are mixing [38]. The situation is further complicated if

the frequency response is non-minimum phase, where changes in the phase, and therefore transient response, will not be detectable from deviations in the amplitude response in isolation [23]. The amplitude aspect of accurate bass reproduction is much easier to interpret, where flatter, more extended lines reflect more even, deeper levels of bass. Colloms [9], when describing the subjective importance of good bass reproduction, suggested that even this representation might be misleading; he argued that the bass region is not afforded its rightful visual weight in the industry-standard logarithmic frequency plots, as at very low frequencies, 'every Hz counts'. However, allowing greater scrutiny on a measurement that still fails to show much of the relevant information is likely to be of little use.

Harley [10] described subjective aspects of good bass reproduction in terms of extension, dynamics, articulation, and tonality. The terms for subjective impression of low-frequency quality were developed formally by Wankling *et al.* [62] as 'bass energy', 'resonance', and 'articulation'. The first term related to 'strength' and 'depth', aspects of low-frequency extension and balance relative to the rest of the audio spectrum, as described by a steady-state magnitude response plot. The latter two described the perceived loudness of individual notes, and their definition and distinction, the 'tightness' of the sound. Newell and Holland [38] described eight types of objective analysis that are useful for describing different aspects of a loudspeaker's response. Colloms [63] considered thirteen. Though these measures do not focus only on low frequency behaviour, if considered together they give a comprehensive account of a loudspeaker's performance. With this information, audio engineers could make a fully informed decision when comparing different monitors; this assumes that they could interpret all of the measurements correctly, and that manufacturers made all of this information available in a standardised way. It seems that audio engineers would benefit from a single, intuitive measure of performance that summarises the most critical aspects of a loudspeaker's low frequency behaviour when selecting monitors for professional use [34].

1.1.4.2 Proposals for a New Measure One of the methods described by Newell *et al.* for loudspeaker evaluation was based on the Modulation Transfer Function (MTF) [21, 35, 64]. By convolving an amplitude-modulated noise signal with a monitor's impulse response, it was demonstrated how low-frequency ringing might mask detail in musical signals. It was seen that loudspeakers with steeper roll-offs reduced the modulation depth of the signal more than models with gentler low-frequency attenuation. The MTF was thus suggested as a measure of the extent to which low-frequency resonances in a loudspeaker's alignment, such as those from reflex loading and the addition of protection filters, was degrading the musical reproduction. This aspect of a loudspeaker's performance, the ability to reproduce musical waveforms accurately, was one of the factors referred to by Andrews [13] in determining whether a loudspeaker is defined as reality or soft focus. It is also a fundamental principle of the Speech Transmission Index (STI), the standardised measure for speech intelligibility inside listening spaces [65]. Although that method focusses on a different type of signal and part of the audio spectrum, it seemed that the MTF could perform a similar role in assessment of loudspeakers reproducing music at low frequencies. A method based on this technique might therefore be a useful measure to help engineers decide whether the monitors they are considering using will provide them with an accurate reproduction of the musical signals captured during recording.

1.1.4.3 Project Aims The project had two primary objectives. Firstly, development of an MTF-based algorithm that is seen to be revealing of the different alignment behaviours commonly encountered in professional mix monitors. The focus is on small to medium-sized monitors because there is a vast array of these models on the market to choose from but their performance at low frequencies is known to vary considerably, despite this region containing the fundamental components of most types of recorded music. Selecting suitable monitors using currently accepted measures may therefore be a particularly difficult process for engineers.

The second objective was to investigate whether the algorithm reflects listeners' judgements of music reproduced through these types of loudspeaker. It has been suggested that any effects due to phase distortion may be subtle, only detectable with certain types of signal under well-controlled listening conditions [28]. If this is true, it seems that professional mixing engineers would be the most likely of any listeners to appreciate these differences.

Evidence that the algorithm is successful in both of these respects is required before the method may be considered a truly useful measure of performance when evaluating mix monitors for professional use.

1.2 Thesis Outline

The rest of the thesis describes the process of developing the MTF algorithm and evaluating its suitability for the intended application, first through analysis of loudspeaker models in isolation, then, after executing a series of subjective experiments, through comparison with listener evaluations of the same systems. A brief synopsis of each chapter is presented below.

Chapter 2 - Algorithm development. The process of developing the MTF-based algorithm is described, explaining the key parameters that were investigated to optimise it for evaluation of musical content at low frequencies. Some basic validation is presented, using simplistic loudspeaker simulations and measured responses of real medium-sized professional mix monitors.

Chapter 3 - Creation of loudspeaker models. This explains the process used to create a suitable set of loudspeaker models for experimentation. It is described how realistic low-frequency loudspeaker simulations were generated, representative of the range of alignments that might be observed in real mixing monitors.

Chapter 4 - Objective evaluation. The final loudspeaker models chosen for evaluation are presented. These are divided into three groups for assessment. The simulated alignments are presented along with measured results; findings following their analysis with the MTF algorithm are also presented and discussed.

Chapter 5 - Listening test design. This describes development of the listening test procedure that was used to evaluate the loudspeaker models. Details of the experimental setup, and selection and preparation of suitable musical extracts is also presented.

Chapter 6 - Statistical methods. During the project it became evident that a clear framework for statistical analysis of the listening test data had to be developed in order to fully address aspects of the experimental question. This chapter describes how suitable

methods were selected, determined by the chosen experimental method, and summarises how they were applied to the test data.

Chapter 7 - Subjective evaluation. Results from each set of listening tests, one for each of the three groups of loudspeaker models, are presented and analysed. The findings and conclusions within and across each set of experiments are summarised and discussed in detail.

Chapter 8 - Comparison of objective and subjective results. The MTF results of the loudspeaker models are compared with the findings from subjective evaluation. Conclusions about suitability of the algorithm for the assessment of mix monitor performance at low frequencies are presented.

The final chapter summarises the findings from the project, presenting conclusions and recommendations for further work.

1.3 Original Contributions

There are two main contributions of this work to the topic of using an MTF-based method for evaluation of loudspeakers:

1. Development of an algorithm tailored to the assessment of monitors reproducing music at low frequencies.

   A suitable method for this application was investigated in more detail than had previously been performed. This included a review of modulation theory to understand implications of applying the technique at low frequencies, comparison of possible computation methods, experimentation with key parameters for improved performance, and consideration of the most effective output formats. Usefulness of the final algorithm as an objective metric for its intended application has been demonstrated with both loudspeaker models and measurements of real monitors.

2. Collection of subjective data for comparison with algorithm results.

   Earlier work lacked experimental data to demonstrate that an MTF-based method related to subjective judgements, i.e. could predict the responses of listeners when assessing the accuracy of monitors reproducing music. In this project, carefully controlled listening tests were conducted to assess the subjective usefulness of the new algorithm. It was shown that when listeners were in consensus about a judgement, the algorithm predicted the outcome; when listeners did not agree about a judgement, detail within the algorithm results was useful in understanding why they may have been divided about the loudspeakers they were comparing.

From these two contributions, it is concluded that the algorithm developed and evaluated in this project is useful for its intended application; it is therefore suitable for more detailed subjective experimentation with the intended end-users. These two outcomes from the project are discussed in more detail in section 9.1

2 Developing an Algorithm

This chapter describes how an algorithm was developed to evaluate loudspeakers reproducing low-frequency musical signals. Section 2.1 gives an introduction to the basis of the method, the Modulation Transfer Function (MTF), describing existing techniques that are relevant to the algorithm developed in this project. Section 2.2 summarises important aspects of modulation theory that are referred to later in the chapter. Possible options for calculating the MTF are given in section 2.3; a description of how one of these was selected is then given in section 2.4. Development of key parameters is described in sections 2.5 to 2.7, and details of the resulting algorithm output are given in section 2.9. Validation of the algorithm using multiple test systems is shown in section 2.11.

2.1 The Modulation Transfer Function

2.1.1 Defining 'The MTF'

The definition of 'the MTF' depends on the context and particular details of the application. Variants include the Optical Transfer Function (OTF) [66], a measure of blur in images, the Temporal MTF (TMTF) [67], a measure of temporal resolution in the auditory system, the Complex MTF (CMTF) [68], a formal definition of what is commonly meant by 'the MTF' in acoustics, and a proposed variant of this, the Narrow-band MTF (NMTF) [69]. The principle of the MTF considered in this project has been described many times in relation to its application in speech intelligibility evaluation; the following description is based on some informative sources related to that method [65, 70, 71].

In traditional communications terms, a noise carrier signal is amplitude-modulated by a sinusoidal function. After passing through the system under test, comparing the ratio of output to input envelope depth gives a measure of the system's ability to preserve temporal fluctuations in a signal passing through it. The input and output are described by:

$$I_i(t) = \overline{I}_i(1 + m_i \cos 2\pi f_m t) \tag{2.1}$$

$$I_o(t) = \overline{I}_o(1 + m_o \cos 2\pi f_m (t - \tau)) \tag{2.2}$$

where: I_i is input intensity, I_o is output intensity, with mean input and output intensities given by $\overline{I}_i$ and $\overline{I}_o$. The modulation frequency in Hz is given by f_m, and t is time in seconds; m_i and m_o are the modulation indexes of the input and output signals respectively. The variable τ signifies that the output signal is delayed relative to the input and therefore corresponds to a phase change, but this term is ignored when calculating the MTF for assessment of speech intelligibility [70, 72, 73]. The resulting modulation index, the quantitative measure of reduction depth of the signal envelope, is then calculated as:

$$m(f_m) = \frac{m_o(f_m)}{m_i(f_m)} \tag{2.3}$$

The concept is illustrated in Figure 1.

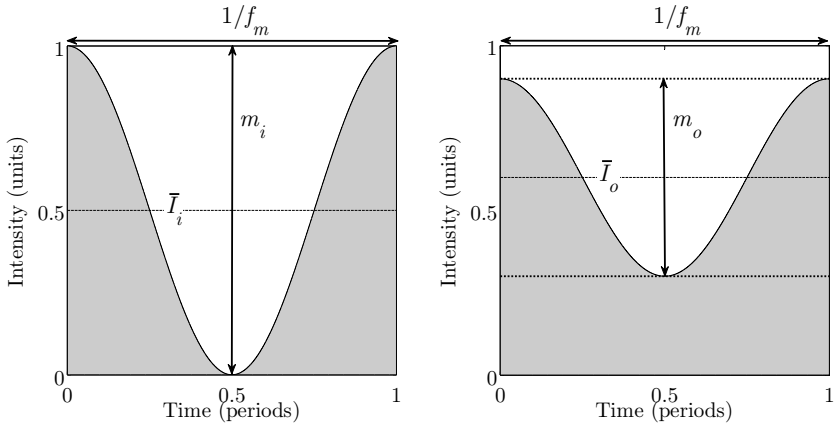


Figure 1: Illustrating the principle of the MTF in acoustic measurement. System input and output signals are shown in the left and right plots respectively. The modulation frequency is f_m. The modulation index is found by calculating the ratio of m_o to m_i. Note that after transmission through the system, the envelope depth has reduced; if the test system is noisy, the mean intensity increases so that $\bar{I}_o$ is greater than $\bar{I}_i$.

The method defined by Eqns. 2.1 to 2.3 was originally developed based on direct measurement of signals before and after acoustic transmission through the system under test. Schroeder [68] later defined an 'indirect' method of MTF calculation, showing that $m(f_m)$ could be derived from the impulse response of the test system. Under this definition, the MTF is a complex quantity; Schroeder therefore provided a formal definition for the complex modulation transfer function (CMTF), which means that it also accounted for phase transfer. As this component is not required for typical acoustic applications it is usually the case that 'the MTF' refers to $|m(f_m)|$, the modulus of the CMTF [68, 74–76]. Schroeder's definition was attractive because, unlike the direct method of calculation from input and output envelopes, it did not require a lengthy measurement procedure with many modulated noise signals; only a single measurement of the system's response was needed to calculate the MTF. However, Schroeder's approach was based on a number of key assumptions; the expression produced results equivalent to direct measurement only if the test system was linear, noise free, time-invariant, and not band-limited [68, 77]. Despite these limitations, the method is still commonly used to calculate the MTF in acoustic applications, and has been extended to allow its use with noise-contaminated and weakly non-linear systems [73, 78].

2.1.2 Developing an MTF-Based Method

Section 2.1.1 described the two fundamental methods used in acoustic applications to calculate the MTF: direct measurement of modulated test signals, and calculation based on the system's impulse response. However, neither of these fully describe the MTF-based method that they relate to. In the 1970s, the MTF was used to form the basis of a method called the Speech Transmission Index (STI) [72]. The creators of the STI wanted a diagnostic method for assessing how much an acoustic space reduced the intelligibility of speech; the motivation for development was to find a way to save the time and effort required for long and tiresome subjective

intelligibility testing. The method produced a single score between 0 and 1, allowing simple interpretation of results by those without extended training in the underlying techniques, and demonstrated good correlation with subjective data, thereby removing the need for time-consuming formal assessment of speech intelligibility through word recognition tests with many listeners [72, 77].

The MTF was chosen as the basis for the method because it was recognised that accurate transmission of a speech signal's envelope is essential in preserving intelligibility. If a system preserves the temporal envelope of a signal, it is inferred that the system preserves intelligibility [73]. The MTF was therefore proposed as a useful way to evaluate an acoustic space (termed the 'enclosure' in the original literature) as an envelope transmission system; it allowed quantification of the smoothing, or smearing, of the signal envelope in the time domain as a function of modulation frequency [72]. It was acknowledged that this use of the MTF did not reveal any additional information about the transmission path than existing measures used at the time; it was the way in which the information was processed and presented which made the STI a useful measure.

An important feature in creating the STI, making it an 'MTF-based method' rather than simply 'the MTF', was the application-specific choice of analysis parameters. With reference to the direct-measurement method upon which the STI was first developed, the frequency band covered by the noise carrier signal was considered to be an important parameter; the use of different bands allowed inspection of the system's behaviour in different regions of the audio spectrum. Seven bands were chosen, each an octave wide as this was deemed to be 'a common degree of frequency selectivity in acoustical measurements' [75]. The point at which filtering into frequency bands is performed can vary according to the method being used for calculation, but in the originally proposed STI method, filtering was performed at the receiving end (modulate–transmit–filter); filtering at the input (modulate–filter–transmit) has the disadvantage that the filter becomes part of the test system [68, 72, 74].

Another key feature in STI development was the selection of specific modulation frequencies. There was a clear rationale for choosing the values for in this application: spectral differences between the constituent components of speech produce fluctuations in the envelope of the signal [70]. A small number of modulation frequencies were selected by analysing the long-term temporal envelope spectrum of speech. This method of modulation frequency selection had two advantages: it reduced computation time due to the limited number of assessment variables – important when the results had to be physically measured – and more importantly, it tailored the method specifically to the application of interest. In this way, the test signals were made to have temporal fluctuations equivalent to individual components of a complex speech signal; if the system being evaluated was seen to transmit the individual components accurately, it was inferred that the system must also preserve intelligibility of a complex speech signal comprising these individual elements [73]. Although a range up to 20 Hz was suggested through analysis of the long-term speech envelope spectrum, the modulation frequencies eventually chosen for use in the STI used 14 third-octave spaced values ranging between 0.6 and 12.5 Hz.

The final crucial stage in development of the STI was formulating a way to process and present the results; the defined method generated a 7–by–14 matrix of m scores, each one being the result of a different combination of bandpass (octave) filtered noise, modulated by individual modulation frequencies. Through conversion of each matrix element to an apparent signal to

noise ratio, band weighting, and averaging, the final result of a single figure between 0 and 1 was produced. Through comparing the results with subjective test results, the 'tailoring' of these final processing stages was refined, and it was demonstrated that the method was a simple and reliable way to predict intelligibility of speech inside an acoustic space without needing to perform subjective testing [70, 73, 75].

Since its development in the 1970s, various extensions, modifications, and limitations of STI as an MTF-based method have been explored to try and improve its usefulness and accuracy, particularly in relation to its ability to predict subjective impression. One aspect of this involved making computation faster, described as 'estimations' of the STI [77]; these were not always successful, due to using the indirect method described in section 2.1.1 without meeting the required assumptions, or by drastically reducing the number of measurement points, equivalent to calculating result from a sparsely populated MTF matrix; RASTI [79] and STIPa [65] are examples of the latter approach. Other modifications include: alternative weighting of bands before averaging to emphasise the contribution of parts of the frequency spectrum most important to speech intelligibility, corrections for masking, and adjustments for redundancy between bands [80–82][‡]. Thus, the relatively simple concept that was the original basis for the STI technique has been refined and developed over several decades by many researchers, and it is now a internationally-used standardised technique for gauging speech intelligibility, typically applied by acoustics consultants when creating new listening spaces or installing public address systems into existing ones [65]. However, even after more than 40 years of research, the method is still subject to ongoing research to improve consistency of application, interpretation of results, and correlation with subjective impression. Known limitations include inconsistency of results for a given system due to the use of a random test signal (noise), and the insensitivity of the final score in reflecting changes in a system's frequency response magnitude [65, 82, 84–86].

2.1.3 Use of the MTF for Loudspeaker Evaluation

As already described, the listening space is considered a key part of the reproduction system in the application that STI was developed for. The algorithm being developed in this project focuses only on the reproduction accuracy of the loudspeaker, and only at low frequencies. Application of the MTF in this context was first demonstrated by Holland *et al.* [21]. Based on the principles of the STI, the method was applied to a number of professional monitors that had differing low-frequency design strategies [64]. The analysis parameters had been selected in STI to be appropriate for analysis of speech; for application to loudspeakers at low frequencies, the bands and modulation frequencies were adjusted accordingly. The method used seven third-octave bands with centre frequencies from 18 to 80 Hz, and seven modulation frequencies similar to those used in the speech application: 3.15, 4.0, 5.0, 6.3, 8.0. 10.0, and 12.5 Hz.

It appeared that loudspeakers expected to demonstrate a better transient response returned a higher MTF score. The method was therefore proposed as a measure of bass reproduction accuracy, also referred to as bass articulation, showing the extent to which resonances or distortions in a loudspeaker have 'blurred' the detail in low-frequency elements of a musical signal [35]. The MTF was considered to be especially useful in the assessment of professional mixing monitors because it offers important information that the industry-standard magnitude

[‡]Comprehensive reviews of development since STI's creation have been compiled and provide interesting reading beyond the scope of this discussion [77, 83].

response plot lacks. This popular measure of loudspeaker performance was likened to the letter count in a paragraph of text; if the order of the letters is scrambled, the letter count remains the same, but the original meaning in the text is lost [38]. This is analogous to evaluating a loudspeaker on the basis of its magnitude response and declaring it to be an accurate transducer when it preserves the relative level of constituent components of a signal but changes their relative phases. As discussed in section 1.1.3.1, for musical signals this means that the individual frequency components may still be reproduced by an inaccurate monitor, but they will not arrive in the correct temporal order. It appeared that the MTF could identify mix monitors that are scrambling the musical message, and are therefore likely to produce errors that cannot be corrected at the mastering stage.

Formal subjective assessment of the method developed by Holland *et al.* was performed by Harris [87]; this study indicated that there was a correlation between MTF score and subjective impression of bass reproduction accuracy. However, the study had limited scope, and the algorithm had not been developed in depth; alternative computation methods and parameters had not been tried, known errors with the technique had not been investigated, and simplistic loudspeaker models had been used. It was therefore the aim of this project to address these issues and develop an MTF-based method that was shown to be useful in evaluating the musical reproduction accuracy of professional monitoring loudspeakers at low frequencies.

2.2 Aspects of Modulation

The MTF technique is based on detection of waveform envelope depth following amplitude modulation. Communication theory is used here to summarise two key aspects of this type of modulation that are relevant to development of the MTF algorithm; these are referred to later in the chapter.

2.2.1 Definition of Amplitude Modulation

Considering the simple case of two sinusoids, the amplitude of message signal $x_{\mathrm{m}}(t)$ modulates a carrier wave $x_{\mathrm{c}}(t)$ to produce $y(t)$:

$$x_{\mathrm{m}}(t) = B\cos 2\pi f_{\mathrm{m}}t \tag{2.4}$$

$$x_{\mathrm{c}}(t) = A\cos 2\pi f_{\mathrm{c}}t \tag{2.5}$$

$$\begin{aligned} y(t) &= x_{\mathrm{m}}(t)\cos 2\pi f_{\mathrm{c}}t + x_{\mathrm{c}}(t) \\ &= x_{\mathrm{m}}(t)\cos 2\pi f_{\mathrm{c}}t + A\cos 2\pi f_{\mathrm{c}}t \\ &= [x_{\mathrm{m}}(t) + A]\cos 2\pi f_{\mathrm{c}}t \end{aligned} \tag{2.6}$$

where: B and f_{m} are amplitude and frequency, in Hz, of the message signal; A and f_{c} are amplitude and frequency of the carrier. The envelope of the modulated signal $y(t)$ is identical to that of $x_{\mathrm{m}}(t)$ but peaks at $A + B$ due to addition of the carrier sinusoid.

The form of amplitude modulation described by Eqn. 2.6 is formally known as double side-band large carrier (DSB-LC) [88], and is the most basic form of AM. This form is primarily used in communications because it allows envelope detection at the receiver using simple rectification and analogue low-pass filtering. Without this constraint it is more convenient to use

the double side-band suppressed carrier (DSB-SC) alternative; there is no addition of the carrier term and Eqn. 2.6 simplifies to:

$$y_{\mathrm{sc}}(t) = x_{\mathrm{m}}(t)x_{\mathrm{c}}(t) \tag{2.7}$$

Both forms of modulation translate the frequency spectrum of $x_{\mathrm{m}}(t)$ symmetrically about the carrier frequency f_{c}.
A key parameter in amplitude modulation is the modulation factor, m; this describes the depth of modulation and is found from the ratio of peak amplitude of the modulating signal to that of the carrier:

$$m = \frac{B}{A} \tag{2.8}$$

where: A is the carrier signal amplitude and B is the peak message signal amplitude [89, 90]. Section 2.2.2 describes two key requirements that ensure successful amplitude modulation.

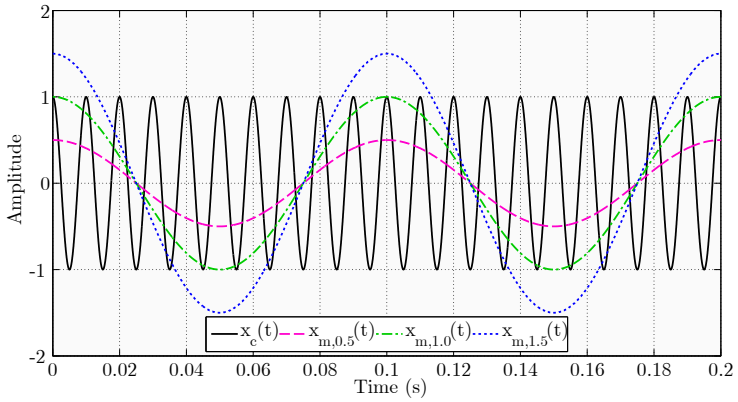
2.2.2 Requirements for Successful Amplitude Modulation

For successful transmission and recovery of an AM signal, it is generally assumed that:

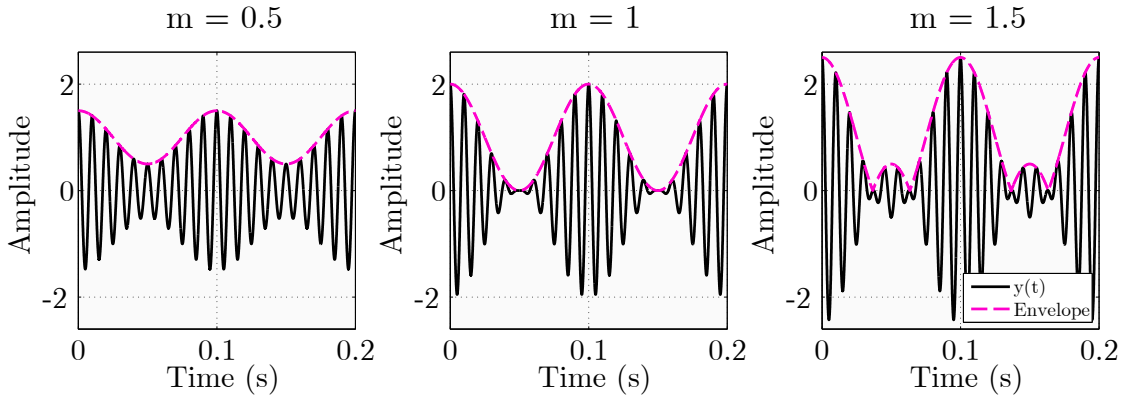
1. Message signal amplitude does not exceed that of the carrier: $A \geq B$.
2. Carrier frequency is much higher than that of the message signal: $f_{\mathrm{c}} \gg f_{\mathrm{m}}$.

Implications of these requirements are summarised in sections 2.2.2.1 and 2.2.2.2.

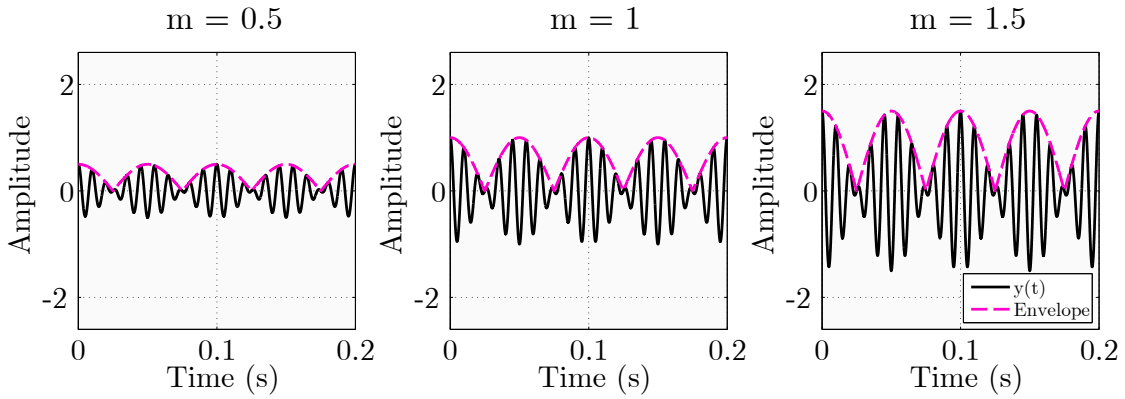
2.2.2.1 Requirement 1: Amplitude The simplest and most efficient case for AM is where $A = B$, i.e. the amplitudes of message and carrier signals are equal. This leads to $m = 1$, and the carrier said to be 100 % modulated [91]; this is the maximum modulation depth possible that will still allow accurate envelope detection of the resulting waveform. If Requirement 1 is violated in DSB-LC AM, i.e. $A < B$ and $m > 1$, the output waveform $y(t)$ is said to be over-modulated and its envelope is a distorted version of the original message signal equivalent. This distortion does not occur in suppressed-carrier AM, and a usable modulation envelope can still be recovered. Figure 2 demonstrates the implications of Requirement 1 for both types of AM in three cases: $m < 1$ (under-modulation), $m = 1$, and $m > 1$ (over-modulation).



(a) Carrier $x_{\mathrm{c}}(t)$ with amplitude $A = 1$ and three message waveforms, $x_{\mathrm{m}B}(t)$, having identical frequencies but different amplitudes: B =0.5A, 1.0A, and 1.5A



(b) Modulated waveforms after large-carrier AM; plot titles show modulation factor, m. Envelope of the modulated waveform is proportional to that of the message signal until Requirement 1 is violated i.e. $m = B/A$ exceeds 1



(c) Modulated waveforms after suppressed-carrier AM. Envelope of modulated waveform is not directly proportional to that of the message signal but the waveshape is not distorted even when $m > 1$

Figure 2: AM envelopes for three modulation factors: $m < 1$, $m = 1$, $m > 1$

It can been seen in Figure 2 that Requirement 1 must be met if large-carrier AM is used, but is not critical if applying the suppressed-carrier variant.

2.2.2.2 Requirement 2: Frequency The examples in Fig. 2 used signals where $f_\mathrm{c} = 10\, f_\mathrm{m}$, i.e. $f_\mathrm{c} \gg f_\mathrm{m}$. This is an important requirement because amplitude modulating a waveform in the time domain changes its frequency spectrum. In the simplest case of single-frequency modulation, the modulated signal $y(t)$ contains three components, at frequencies equal to:

$$[f_\mathrm{c} - f_\mathrm{m}], \quad f_\mathrm{c}, \quad [f_\mathrm{c} + f_\mathrm{m}].$$

Symmetrical sum and difference components are also produced either side of $-f_c$.

In typical communications applications, the message signal contains multiple components. This signal, $x_\mathrm{sig}(t)$, is described as a baseband message with components from f_L up to f_U, translated in frequency through amplitude modulation of the much higher-frequency carrier sinusoid $x_\mathrm{c}(t)$. The effect is illustrated in Figure 3, showing a fixed band of sinusoids after modulation with three carrier conditions: $f_c \gg f_U$, $f_c > f_U$, and $f_c \ll f_U$. Note that the latter condition is not used in communications applications, so the implications are not shown in textbooks when describing amplitude modulation. The DSB-LC variant of AM has been used to illustrate more clearly how the bands are translated about the carrier frequency; the same effect would be seen in the suppressed-carrier case, but the carrier component is absent and relative magnitude of the modulated bands increases because all power is contained within the sidebands, not shared with the carrier.

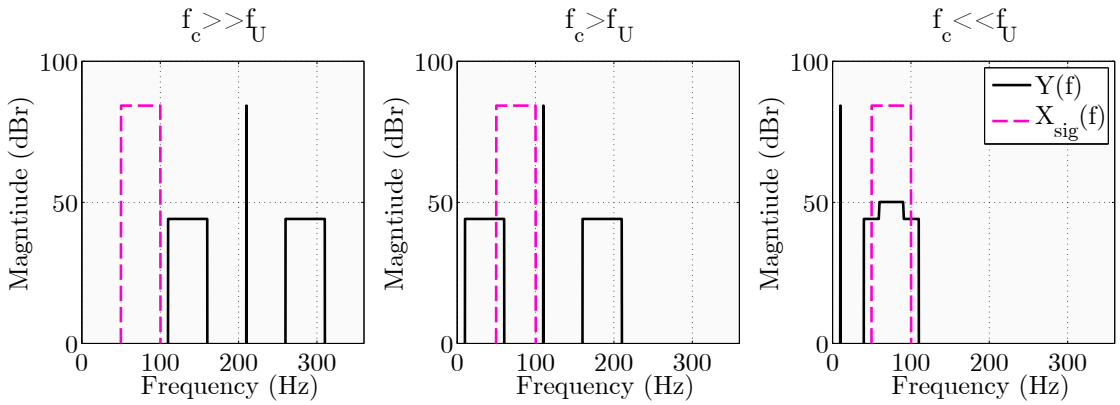


Figure 3: Amplitude modulation of a band-limited signal and carrier sinusoid $x_c(t)$ that is much higher than, just higher than, and much lower than the band. Plots show the spectra of modulated signals $y(t)$ when $f_\mathrm{c} = 2.1 f_U = 210\,\mathrm{Hz}$, $f_c = 1.1 f_U = 110\,\mathrm{Hz}$, and $f_c = 0.1 f_U = 10\,\mathrm{Hz}$. The signal $x_\mathrm{sig}(t)$ contains frequencies from $f_L = 50\,\mathrm{Hz}$ to $f_U = 100\,\mathrm{Hz}$. Note that $X_\mathrm{sig}(f) = \mathscr{F}\{x_\mathrm{sig}(t)\}$ and $Y(f) = \mathscr{F}\{y(t)\}$

The typical requirement in communications applications is that $f_c \geq 2f_U$; this ensures that the modulated signal sidebands are sufficiently separated from the original baseband message to be reliably recovered through demodulation [88, 89]. However, it can be seen from Fig. 3 that the upper and lower sidebands, USB and LSB respectively, are discrete either side of the carrier provided that f_c is above the signal band, $f_c > f_U$. When f_c is lowered, the spectrum of the

modulated signal also moves closer to 0 Hz. In the case where f_c is very small relative to the band being modulated, $f_c \ll f_U$, it appears that only one sideband is produced, located above f_c with bandwidth 40 Hz to 110 Hz. In fact, this is due to partial overlapping of the USB, ranging from $[f_c + f_L]$ to $[f_c + f_U]$, with the aliased band $[-f_c + f_L]$ to $[-f_c + f_U]$; a doubling in magnitude can be seen where these bands coincide. The LSB has been translated below 0 Hz and is no longer visible. Therefore, the overall effect when using a very low f_c is an apparent spreading of the signal band; the modulation changes the original bandwidth, β, ranging between f_L and f_U to an 'effective' bandwidth, β_e, of $[f_L - f_c]$ to $[f_U + f_c]$. The extent of this apparent spreading is proportional to the bandwidth for a fixed value of f_c. For example, if $f_c = 1$ Hz, a 10 Hz-wide band would be seen to have spread by 20 %; this reduces to 2 % for a band that is 100 Hz wide.

Implications of the effect described here are referred to in sections 2.3 and 2.4 when considering methods for calculating the MTF. Note that the term 'carrier frequency' has been used in this section, in line with communications theory. In further discussion, it is referred to as the modulation frequency, the terminology more commonly encountered in the context of the MTF.

2.3 Methods for MTF Calculation

In section 2.1 it was established that the fundamental principle of MTF assessment is measuring the extent to which a signal's envelope depth is reduced as it passes through a test system. It was seen that the MTF may be computed in different ways, and that the development of a successful MTF-based method requires selection of parameters that are meaningful for the application in question, ideally being selected through consideration of the type of signal that the system transmits and the frequency range over which it operates. This section presents the methods considered for MTF computation.

Four techniques for computing the MTF were considered. All are based on an impulse response measurement of the loudspeaker under test, but calculate the MTF in different ways. As was described in section 2.1.2, analysis is typically performed for a number of different frequency bands; band-limiting either the test system or the test signal is a valid approach in an MTF-based method. Two method groups were formed on this basis, referred to here as: band-limited impulse response (BLIR) and band-limited input (BLIP). Methods 1 and 2 band-limit the test system; Methods 3 and 4 band-limit the test signal. The methods are summarised in sections 2.3.1 to 2.3.4.

2.3.1 Method 1: Schroeder Expression (BLIR)

This method uses the equation described in section 2.1.1 to compute the MTF from a measurement of the test system's impulse response.

For a system (loudspeaker) with impulse response $h(t)$, the modulation index is calculated using:

$$m(f_m) = \frac{\int_0^\infty h^2(t) \mathrm{e}^{-\mathrm{j}2\pi f_\mathrm{m} t} \mathrm{d}t}{\int_0^\infty h^2(t) \mathrm{d}t} \tag{2.9}$$

where: $m(f_m)$ is the modulation index, f_m is the modulation frequency, and t is the time index.

The value of the modulation index is determined by two factors affecting the test system: the presence of noise, and the decay of energy over time. The second of these becomes the only factor in the absence of noise. Note that for the application being considered here, this refers to the decay of resonances within a loudspeaker, but in typical acoustic MTF applications that are concerned with room measurement, it refers to reverberation time. In either case, it is assumed that the impulse response used in the Schroeder equation is a measurement of the system under test, which may or may not include the room response, depending on the application. Inspection of Eqn. 2.9 shows that the MTF calculation is actually based on the squared impulse response. Schroeder had already shown that this could be integrated to obtain the system's energy decay with time [92]. This principle was later used in deriving the expression for computing the MTF shown here, which divides the Fourier Transform of the squared impulse response by the total system energy. This definition of the MTF can therefore be viewed as a normalised measure of energy transfer, reflecting the energy decay of the system at modulation frequency f_m as a proportion of the total system energy [80]. This corresponds to the direct measurement method described in section 2.1.1, which calculates the MTF as a normalised measure of intensity transmission, i.e. a time-averaged rate of energy transfer through the system. Equation 2.9 therefore correctly calculates the MTF, but only if it is valid to assume that the test system, and its measured impulse response, is noise-free. This is equivalent to mean intensities $\overline{I}_i$ and $\overline{I}_o$ being equal in the direct measurement method described by Eqns. 2.1 and 2.2. The additive effect of noise on the output is absent in this case, so reduction in modulation depth at the system output is therefore assumed to be due only to the effect of energy decay in the system, and the result $m(f_m)$ depends only on the ratio of the modulation envelope intensities.

The fact that the indirect computation only produces equivalent results to direct measurement under certain conditions means that the Schroeder implementation is not suitable for many practical applications, such as measuring the MTF of listening spaces with background noise or where the system response varies with time. However, the method was considered in this study as it was assumed that a loudspeaker's response was time-invariant and measured anechoically. Use of the Schroeder expression had also been demonstrated in the preceding studies of loudspeaker MTF assessment at low frequencies [64, 93].

Calculation of Eqn. 2.9 is performed for multiple frequency bands and using different modulation frequencies. As described in section 2.1.1, this expression returns the complex MTF, but the phase transfer component is not required for the application of interest here; therefore, only the modulus is used, as demonstrated in the steps below:

- Test system $h(t)$ is filtered in the frequency domain to produce the band-limited impulse response $h_{\mathrm{BL}}(t)$:

$$h_{\mathrm{BL}}(t) = \mathscr{F}^{-1}\left\{H(f) \cdot T(f)\right\} \tag{2.10}$$

where: $H(f)$ is the complex frequency response function (FRF) of the test system, and $T(f)$ is a filter with bandwidth β covering the range of frequencies from f_L to f_U. Details relating to band specifications are given in section 2.5.

- The reference and modulated signals, $v_{\mathrm{i}}(t)$ and $v_{\mathrm{o}}(t)$, are calculated for modulation frequency f_m:

$$v_{\mathrm{i}}(t) = h_{\mathrm{BL}}^2(t) \tag{2.11}$$

$$\phi(t) = \mathrm{e}^{-\mathrm{j}2\pi f_m t} \tag{2.12}$$

$$v_{\mathrm{o}}(t) = v_{\mathrm{i}}(t)\phi(t) \tag{2.13}$$

- Modulation index m is calculated from:

$$m = \left| \frac{\sum_{k=1}^{N} v_{\mathrm{o}}(t_k)}{\sum_{k=1}^{N} v_{\mathrm{i}}(t_k)} \right| \tag{2.14}$$

Output $v_{\mathrm{o}}(t)$ is complex and has magnitude identical to $v_{\mathrm{i}}(t)$; results are added for all values of t, where N is the length of the impulse response in samples. Modulation index m is obtained by taking the absolute values after summation.

2.3.2 Method 2: Formal Simulation (BLIR)

Like Method 1, Method 2 band-limits the test system. It simulates passing an amplitude-modulated noise signal through a band-limited system. From the review of modulation theory discussed in section 2.2.2, there appeared to be no requirement for using large-carrier AM. The simpler suppressed-carrier version was therefore applied; consequently, amplitude of the modulating and signal functions was not critical. The process is summarised below:

- Broadband noise $x_{\mathrm{n}}(t)$ is amplitude modulated by sinusoidal function $x_{\mathrm{c}}(t)$ having frequency f_m:

$$x_{\mathrm{nm}}(t) = x_{\mathrm{n}}(t)x_{\mathrm{c}}(t) \tag{2.15}$$

- The band-limited test system, $h_{\mathrm{BL}}(t)$, is obtained in the same way as for Method 1, using Eqn. 2.10.

- Convolution simulates transmission of the modulated test signal through the band-limited test system:

$$y_o(t) = x_{\mathrm{nm}}(t) * h_{\mathrm{BL}}(t) \tag{2.16}$$

- Modulation depth of input and output envelopes, $v_i(t)$ and $v_o(t)$, are used to calculate the modulation factors:

$$v_i(t) = |x_{\mathrm{nm}}(t)| \tag{2.17}$$

$$m_i = \frac{v_{i\ \max} - v_{i\ \min}}{v_{i\ \max}} \tag{2.18}$$

$$v_o(t) = |y_o(t)| \tag{2.19}$$

$$m_o = \frac{v_{o\ \max} - v_{o\ \min}}{v_{o\ \max}} \tag{2.20}$$

where: $v_{i\,\max}$ and $v_{i\,\min}$ are the maxima and minima of the input envelope, $v_{o\,\max}$ and $v_{o\,\min}$ are the equivalent maxima and minima of the output envelope, m_i is modulation depth of the input signal, and m_o is modulation depth of the output.

- Modulation index m is calculated from:

$$m = \frac{m_o}{m_i} \tag{2.21}$$

A reduction in modulation depth of signal $x_{\mathrm{nm}}(t)$ after passing through the test system in a given frequency band gives a reduction in m_o relative to m_i. If the modulation depth is unaltered by transmission through the system, $m_o = m_i$ and $m = 1$.

2.3.3 Method 3: Filter-Modulate (BLIP)

Method 3 avoided filtering of the system by band-limiting the modulated input signal instead. This process is therefore similar to the direct-measurement method described in section 2.1.2, as implemented when the Speech Transmission Index was developed. The advantage of simulating transmission through the test system, rather than measuring the transmitted and received time histories, is the ease and relative speed of computation. A potential limitation of this method is that, as with the Schroeder technique, the evaluation is inherently linear. In this study, loudspeaker measurements were derived using the dual-channel technique (described further in section 3.3.1); simulating transmission through the test system based on such a measurement means that nonlinearities or time-variance of the system are not considered. This was not considered to be a major problem in this study; it was assumed that, for practical application of the method, measurements of professional monitoring loudspeakers would be measured within their linear operating range and in a noise-free (anechoic) environment. The process is summarised below:

- Broadband noise $x_n(t)$ is band-limited in the frequency domain:

$$x_{\mathrm{nBL}}(t) = \mathscr{F}^{-1}\left\{X_n(f) \cdot T(f)\right\} \tag{2.22}$$

where: $T(f)$ is a filter with bandwidth covering the desired range of frequencies. Subscript nBL denotes that band-limiting was performed before modulation.

- The noise band is amplitude-modulated by sinusoidal function $x_{\mathrm{c}}(t)$ with frequency f_m:

$$y_i(t) = x_{\mathrm{nBL}}(t) x_c(t) \tag{2.23}$$

- Convolution simulates transmission of the modulated signal through the test system $h(t)$:

$$y_o(t) = y_i(t) * h(t) \tag{2.24}$$

- Modulation index is calculated from the intensity envelopes of the input and output signals:

$$m = \frac{m_o}{m_i} = \frac{\sum_{k=1}^{L} |y_o(t_k)|}{\sum_{k=1}^{L} |y_i(t_k)|} \tag{2.25}$$

The input and output envelopes are summed for all values of t, where L is the number of samples in one period of the modulating function. A reduction in modulation depth of signal $y_i(t)$ after passing through the test system gives a reduction in m_o relative to m_i. If the modulation depth is unaltered by transmission through system, $m_o = m_i$ and $m = 1$.

2.3.4 Method 4: Modulate-Filter (BLIP)

From the review of modulation theory discussed in section 2.2.2.2, it was demonstrated that violation of Requirement 2 cannot be avoided if modulating with a frequency lower than f_L, the lower limit of the test signal band. The apparent spreading of the band due to overlapping aliased modulation components means that the test system is excited outside the band of interest. Therefore, the Method 3 process does not truly reflect the desired analysis condition; the system is being evaluated over the band defined by an 'effective bandwidth', β_e, rather than the intended range of β. Method 4 was developed to investigate the effect of band-limiting the noise after, rather than before, modulation, thereby avoiding spreading of the signal outside the desired frequency range. This differs from Method 3 only in the point at which filtering is performed:

- Broadband noise $x_{\mathrm{n}}(t)$ is amplitude modulated by sinusoidal function $x_{\mathrm{c}}(t)$ with frequency f_m:

$$x_{\mathrm{nm}}(t) = x_{\mathrm{n}}(t)x_{\mathrm{c}}(t) \tag{2.26}$$

- The modulated signal is band-limited in the frequency domain :

$$y_i(t) = x_{\mathrm{nmBL}}(t) = \mathscr{F}^{-1}\left\{X_{\mathrm{nm}}(f)T(f)\right\} \tag{2.27}$$

  Subscript nmBL denotes that band-limiting was performed after modulation.

The modulation index is then calculated in the same way as for Method 3.

2.4 Method Selection

From the review in section 2.1.2, it was known that different computation strategies had advantages and limitations within an MTF-based method; for example, the Schroeder method has been favoured because it allows fast computation, but band-limiting of the test system will affect results. It was concluded that any of the processes described in section 2.1.2 could be considered for use in the method being developed, but their advantages and limitations in relation to the intended application would need to be investigated. This section describes how a method was selected for further development and experimental work.

An additional point is included here to explain the way in which the selection process has been presented in this chapter. The process described in the following sections is intended to show and justify the selection in a clear and formal way. In practice, initial comparisons of the computation methods with different test systems resulted in two of them being considered for a large part of the study: Method 1 (Schroeder, BLIR), and Method 3 (BLIP, filter–modulate). The final decision between these two alternatives was heavily influenced by objective evaluation of the experimental loudspeakers (presented in chapter 4), and comparison with the subjective data (presented in chapter 7). Further results and justification for the final method are presented

in section 8.4, where the algorithm performance is compared against a number of other objective measures that may be used for evaluating loudspeaker reproduction accuracy at low frequencies; one of these alternatives is an algorithm that uses Method 1 for MTF computation.

2.4.1 Test Parameters and Systems

Three key parameters that affected the MTF algorithm output were identified; these were subject to further investigation to try and optimise performance after a suitable computation method had been chosen. The three parameters and values selected for initial evaluation are listed below:

Frequency bands Four bands with width $\beta = 20\,\text{Hz}$ were used for the analysis, with centre frequencies $f_{\text{cf}} = [30,\ 50,\ 70,\ 500]$, in Hz. Upper- and lower-frequency limits were defined as: $f_L = f_{\text{cf}} - \frac{\beta}{2}$ and $f_U = f_{\text{cf}} + \frac{\beta}{2}$.

Modulation frequencies Four modulation frequencies were applied in each band, defined as a proportion of bandwidth, β: $f_{\text{m}} = [0.1\beta,\ 0.25\beta,\ 0.5\beta,\ 1\beta]$, in Hz.

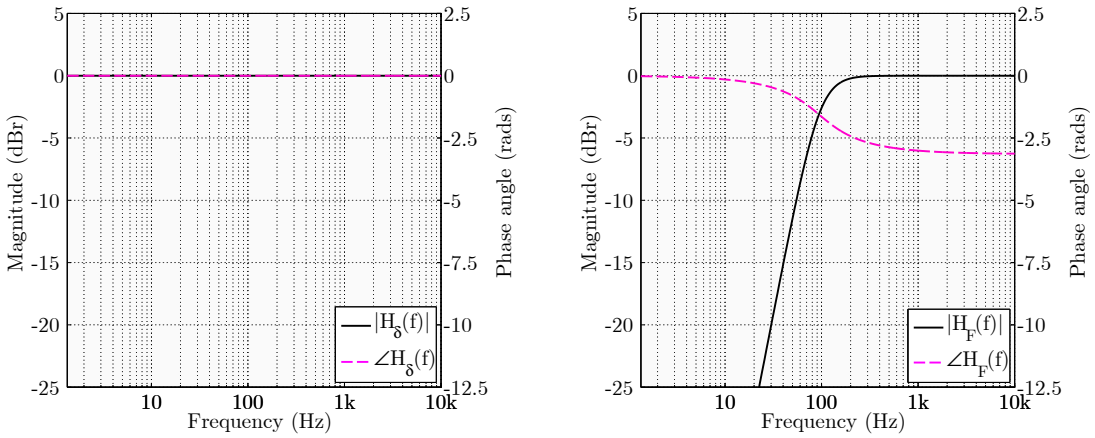
Modulating function The function described in the original STI development [72]was used:

$$x(t) = 0.5(1 + \cos(2\pi f_m t)) \tag{2.28}$$

The four MTF methods were compared through application to a number of test systems. Two systems are used here to demonstrate and discuss differences between the methods, but the findings are consistent with others tried. System 1 was an impulse, approximating a perfect reproduction system:

$$h_\delta(t) \quad = \quad \begin{cases} 1 & t = 0 \\ 0 & \text{otherwise} \end{cases}$$

The second was a Butterworth high pass filter, $h_{\text{F}}(t)$, approximating a loudspeaker with a cut-off frequency of 95 Hz and 2^{nd}-order roll off (12 dB/oct) below this point. Simulation of loudspeaker alignments in this way is described in more detail in section 3.2. The steady-state complex FRFs are shown in Figure 4.



(a) System 1 **(b)** System 2

Figure 4: Test systems: Complex frequency responses

The use of these different systems allowed comparison of the MTF methods in two key aspects. Theoretically, a perfect transmission system should allow an input signal to pass through without any modification; hence, there would be no reduction in modulation depth of the input signal, and System 1 should exhibit a perfect MTF, i.e. $m = 1$ in all frequency bands for all modulation frequencies. Use of the second system approximating behaviour of a loudspeaker at low frequencies would show whether the algorithms responded to this type of response; it was expected that higher m values would be returned in the highest bands where the response of $h_F(t)$ approached that of $h_\delta(t)$.

2.4.2 Evaluation of Test Systems

The four MTF methods were applied to the test systems. Each method produced a matrix of modulation indexes with $n_f n_m$elements, where n_f is the number of bands, and n_m is the number of modulation frequencies. The results of the analysis are shown in Table 1. Only the mean matrix values, $\overline{M}$, are shown here for brevity; the full tables are given in Appendix A.

Method	$\overline{M}$ System 1: $h_\delta(t)$	$\overline{M}$ System 2: $h_F(t)$
1 (BLIR)	0.56	0.54
2 (BLIR)	0.58	0.56
3 (BLIP)	1.00	0.47
4 (BLIP)	1.00	0.46

Table 1: Comparison of results from four MTF methods applied to two systems. Overall mean scores, $\overline{M}$, calculated from the 4-by-4 element matrix of m values and shown to 2 d.p.

The results in Table 1 show that Methods 1 and 2 produced very similar results but with some fluctuation in individual values of m. This had been expected because, as with Methods 3 and 4, the use of random noise caused some variability in the computation of modulation index from input and output signal envelopes. The most effective method for stable envelope convergence was found to be averaging both across multiple cycles of the noise, 2.5 s total duration, and repeating with different noise iterations. Fifty noise iterations were used for the comparisons presented here, though ten iterations were found to give a reasonable compromise between computation time and accuracy. More details about this issue and the variation limits stated here are given in section 2.10.

This comparison of MTF methods produced a number of conclusions:

i) Method 1 was considered to be an effective substitute for Method 2, the formal equivalent using modulated noise.

ii) Method 1 produced identical results every time almost instantly. Methods 2-4 are computationally intensive, and therefore slow, due to the large number of averages required to form a smooth envelope from which each modulation index is calculated. If insufficient repetitions are performed, the algorithm does not produce exactly identical results for the same system on repeated evaluation.

iii) Results for Methods 1 and 2 (BLIR) agreed closely. Results for Method 3 and 4 (BLIP) agreed closely. The two pairs of results did not agree with each other. Therefore, the decision to filter either the system or the input signal is a critical distinction that has a much larger effect on the results than minor differences within a given type of method.

iv) Methods 1 and 2 appeared to show results that would limit usefulness for the present study; the methods did not return a perfect MTF of $\overline{M} = 1$ and did not respond adequately to low frequency attenuation in the simulated loudspeaker system.

v) Methods 3 and 4 demonstrated the behaviour that was expected and required for the application in question, returning $\overline{M} = 1$ for a perfect system and an appreciably lower score for the approximated loudspeaker LF response. The reduction in $\overline{M}$ was considered appropriate given the limited low-frequency extension of the model.

Conclusion ii) required further consideration. The method developed in this study was intended for evaluation of professional-use loudspeakers, designed to be used in very sensitive listening conditions by the most critical listeners; it was therefore decided that variation greater than 1 % in mean MTF score could not be tolerated. It was necessary to find an averaging strategy that would always produce consistent values for $\overline{M}$ to at least 2 decimal places. Following an analysis of error distribution, it was found that 100 iterations produced a fully consistent result for m to 2 d.p. and reduced the error variation in maximum theoretical score to below 0.5 % (change in mean score less than 0.005) for modulation frequencies up to 10 Hz: maximum mean-score standard deviation was $\sigma_{\overline{M}} = 0.002$, based on a distribution of 30 examples. Fluctuations in results of this magnitude were considered to be acceptable.

Following this comparison of the different approaches, Method 2 was discounted for further use as it is computationally intensive and produced results sufficiently similar to Method 1. Results suggested that either Method 3 or 4 would be the most suitable for the intended

application, but further investigation was required to understand the observed behaviour before one of the remaining methods was selected.

2.4.3 Method Selection Through Further Investigation

This section summarises how a method for MTF computation was selected by looking at the effects of each process in further detail.

2.4.3.1 Method 1 Error Method 1 had demonstrated a reduction in m with increasing modulation frequency for a fixed bandwidth, with an apparent error for $\overline{M}$ being in the order of 50 % when evaluating a distortionless transmission system. Experimentation with other values of band centre frequency, f_{cf}, showed that the results were independent of f_{cf} relative to f_{m}. For an infinitely wide band, i.e. no band-limiting of the test system, Method 1 returned values of $m = 1$ for any value of f_{m} i.e. produced the 'correct' result. Reducing the bandwidth from this maximum value produced a decrease in m. The reduction in observed m appeared proportional to the ratio $\frac{f_m}{\beta}$. Note that System 1 was used for this investigation as an a-priori correct MTF score could justifiably be assumed, unlike System 2, or any system other than 'perfect'.

Evidence was found in the literature to support the observed effect that a reduction in m with increasing modulation frequency was due to filtering of the test system [78, 94]. Linkwitz [94] alluded to the necessity for some sort of post-processing normalisation, suggesting a comparison with equivalent results from an ideal system. In relation to the Speech Transmission Index, Rife [78] simply stated that only the averages in each band and the subsequent overall score were significant; the individual modulation indexes, the discrete MTF matrix elements, were not important, and errors due to band filtering would be reduced during the averaging procedure. Fazenda *et al.* [95], applying the MTF to evaluation of room responses at low frequencies, noted the dominating effect of the filtering process on results and provided a solution: calibration by the response of a delta function.

It was suspected that a method for correcting band-limiting errors had not been developed because the STI focuses on a much higher region of the audio spectrum than considered in this project. The use of octave-wide bands in the typical speech range means that the test bandwidth is large compared to the modulation frequencies. The standardised bands and modulation frequencies used in the STI method produce $\frac{f_m}{\beta}$ ratios ranging from 0.0001 to 0.1404. In an application at low frequencies, also assuming octave-wide bands and the same modulation frequencies, these ratios range between 0.1404 and 0.5590. Therefore, for the frequencies in speech assessment, errors due to band-limiting the test system range from observable to negligible; at low frequencies, these errors are always large enough to be a critical problem. Avoiding the error would be impossible in a method that requires band-limiting, but it could be reduced to a known value, e.g. 5 %, by specifying a maximum allowable $\frac{f_m}{\beta}$ ratio. This was undesirable, as key analysis parameters, bandwidth and modulation frequency, would have to be determined by limitations in the method.

As an alternative to fixing the $\frac{f_m}{\beta}$ ratio, a correction technique was attempted which used normalisation by the perfect-system MTF matrix. This is the same form of correction used by Fazenda *et al.* [95]. If the proposition that a perfect system should return a perfect MTF is correct, any deviations from a score of 1 must be due to the band-limiting errors. The perfect-system scores were therefore used as a set of correction values to amend the results from

any other system analysed using the same parameters. The perfect-system normalisation was applied to the Method 1 results as:

$$m_{N(i,j)} = \frac{m_{F(i,j)}}{m_{\delta(i,j)}} \tag{2.29}$$

where: m_N is the normalised modulation index, m_F is the modulation index before normalisation, and m_δ is the equivalent perfect-system modulation index, for all values of modulation frequency i and frequency band j.

Normalised Method 1 results were compared with those from Methods 3 and 4. The scores were expected to be more similar than without the normalisation, but not necessarily identical. Table 2 shows the full MTF matrix for test system $h_F(t)$, analysed using Method 1 with perfect-system normalisation. The mean modulation index in each band is given by $\overline{m}_{\text{cf}}$.

		f_{m} (Hz)				
		2	**5**	**10**	**20**	$\overline{m}_{\mathbf{cf}}$
f_{cf}(Hz)	**30**	0.96	0.91	0.81	0.64	0.83
	50	0.99	0.97	0.93	0.85	0.94
	70	1.00	0.99	0.98	0.93	0.97
	500	1.00	1.00	1.00	1.00	1.00

Table 2: Method 1 results for system $h_F(t)$ after perfect-system normalisation; $\overline{M} = 0.94$

Using the same parameters as the original analysis, the normalisation produced MTF behaviour similar to that seen in Methods 3 and 4. It returned $\overline{M} = 1$ for test system $h_\delta(t)$ (not shown here). For system $h_F(t)$, it seemed that normalisation had reduced the sensitivity to parameter f_m, but increased variation in $\overline{m}_{\text{cf}}$ as a function of f_{cf} i.e. showed greater variation in mean modulation index across frequency bands; the range increased from $\overline{m}_\Delta = 0.06$, to $\overline{m}_\Delta = 0.17$ after normalisation. However, the method still seemed too insensitive to attenuation of low frequencies. With reference to Figure 4b, the output of system $h_F(t)$ in the lowest band, covering 20 to 40 Hz, is negligible compared to the passband; the mean modulation index does not reflect this, returning a high score of $\overline{m}_{30} = 0.83$. The equivalent scores for Methods 3 and 4 were $\overline{m}_{30} = 0.11$ and $\overline{m}_{30} = 0.12$ respectively. This lack of sensitivity appeared to be the same effect that is recognised as a potential limitation of the STI method; the current standard [65] (section 4.5.8) states that if the frequency response of the transmission channel under test is not reasonably flat, the STI will produce scores that are misleadingly high compared to the true perceived intelligibility of speech. This requirement would always be violated for the application of interest in this study, sometimes severely so; even the response of a well-aligned and very extended loudspeaker will never be flat throughout the entire low-frequency region. As discussed in section 1.1.3.3, output level is not the only aspect of accurate bass reproduction, but it is important. It was believed that the method developed here should reflect this parameter, both in overall attenuation at low frequencies, and by showing variations in localised regions, i.e. reflecting the relative level and uniformity of output across different MTF bands. Therefore, it

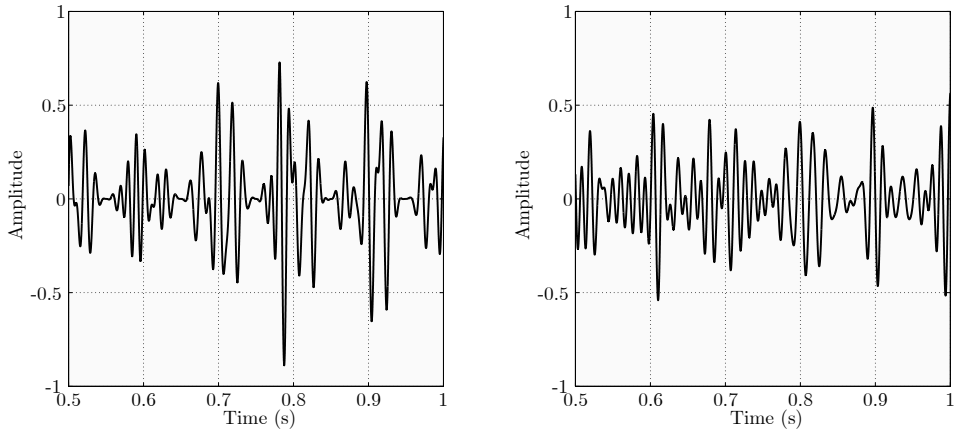
was concluded that even if applying normalisation to the Method 1 scores, either Method 3 or 4 would be more suitable for evaluating loudspeaker behaviour at low frequencies.

2.4.3.2 Method 3 vs 4 Investigation so far had indicated that Methods 3 and 4 produced very similar results; this implied that it did not matter whether the input noise signal was band-limited before or after being modulated. Some further analysis of the two methods was performed to better understand the techniques and their potential impact on the results.

A final decision between Methods 3 and 4 was reached after comparing their respective input and output signals. Figure 5 illustrates the difference between the methods by presenting results for a single band and one modulation frequency:

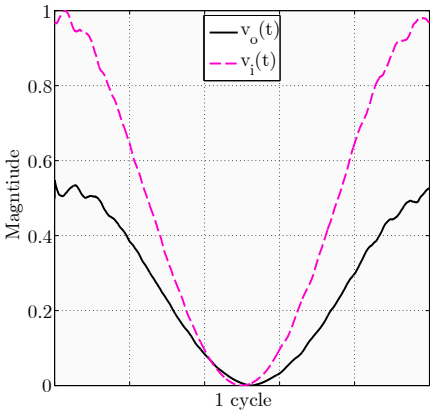
$$\beta = 50\,\text{Hz}; \quad f_L = 50\,\text{Hz}, \quad f_U = 100\,\text{Hz}; \quad f_m = 10\,\text{Hz}.$$

A section of the system input waveforms are shown, along with the final modulation envelopes from which m was calculated.

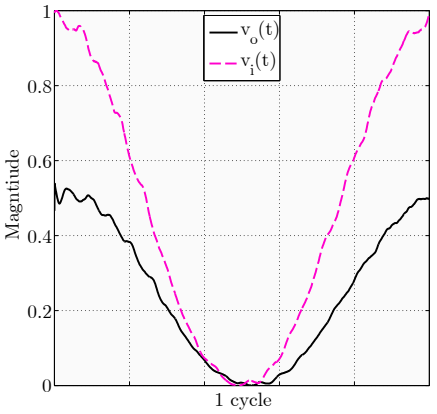


(a) M3 test signal, $y_i(t)$

(b) M4 test signal, $y_i(t)$



(c) M3 input and output envelopes



(d) M4 input and output envelopes

Figure 5: Methods 3 and 4: Inspecting input and output for a single band: $\beta = 50\,\text{Hz}$, $f_m = 10\,\text{Hz}$, $f_{\text{cf}} = 75\,\text{Hz}$. Figures 5a and 5b show 0.5 s of 2.5 s total duration. Figures 5c and 5d show results based on 100 noise iterations. The input and output envelopes have been normalised by a common factor to show how they compare more clearly

It can be seen in Figs. 5a and 5b that the modulation profile of the test signal is less defined in Method 4 than Method 3. This was suspected to be due to rectangular filtering in the frequency domain; the effects reduced when windowing was applied, but could not be removed completely. The two remaining methods for MTF computation were therefore summarised as:

Method 3: Modulates noise after band-limiting;
Maintains modulation profile of $y_i(t)$ but spreads β.

Method 4: Band-limits noise after modulating;
Changes modulation profile of $y_i(t)$ but maintains β.

Although the output envelopes in Fig. 5 appear to be very similar for both methods, it was concluded that an algorithm based on detecting modulation depth should preserve this feature in the test signal. Therefore, Method 3 was chosen for further development.

A modification was developed to compensate for the band-spreading effect inherent in Method 3. If successful, this would remove a known limitation in the method: excitation outside the nominal analysis bandwidth. In section 2.2.2.2, it was shown that the effective bandwidth after modulation, β_e, can be calculated from the modulation parameters:

$$\beta = f_U - f_L \tag{2.30}$$

$$\beta_e = (f_U + f_m) - (f_L - f_m) \tag{2.31}$$

where: f_L and f_U are the lower and upper frequency limits of the band, and f_m is the modulation frequency.

Adjusting the band limits before modulation would therefore return an effective bandwidth after modulation that covered the intended frequency range:

$$f_{\mathrm{La}} = f_L + f_m \tag{2.32}$$

$$f_{\mathrm{Ua}} = f_U - f_m \tag{2.33}$$

$$\beta_e = (f_{\mathrm{Ua}} + f_m) - (f_{\mathrm{La}} - f_m) = f_U - f_L = \beta \tag{2.34}$$

where: f_{La} and f_{La} are the adjusted lower and upper limits respectively of the signal band.

This method did effectively maintain the desired bandwidth, but could only be used while $f_m < 0.5\beta$; if this requirement is not met, f_{La} becomes a higher frequency than f_{Ua}, and the subsequent modulation produces unusable results. This limitation would impose restrictions on development of the MTF algorithm parameters, as either the modulation frequencies would need to be kept low or the bands made suitably wide. Method 3 was therefore used with the knowledge that it produced some excitation in the test system outside the band of interest.

With a computation method selected, parameters of the MTF algorithm were investigated in more detail. These are discussed in sections 2.5 to 2.7.

2.5 Optimisation I: Frequency Bands

As for selection of a computation method, some investigation was needed to choose an arrangement of frequency bands. The correct selection would be the one considered most appropriate for the intended application of the algorithm being developed. This section describes how the frequency range to be covered was defined, then how a suitable set of bands to cover this range was selected.

2.5.1 Defining the Range

As discussed in section 1.1.2, the range of frequencies covered by the term 'bass' is not explicitly defined. The limits specified for this application had to be sufficient to cover the region in which a professional monitoring loudspeaker could be expected to exhibit all characteristics of its low-frequency alignment. The range should be below the typical crossover region between woofer

and midrange drivers so that behaviour of other parts of the loudspeaker system would not dominate results in the MTF.

It would have been appropriate to define the frequency coverage in this study according to the lower limit of human hearing and the upper limit of woofer output in professional mix monitors, but there is no universally accepted value for either of these parameters [96]. Given the intended application of the method being developed, it was appropriate to consider the content of musical signals. Based on observations by Fielder and Benjamin [12] and Colloms [9], limits of 16 to 160 Hz seemed suitable when considering bass reproduction in music. The exact limits were not regarded as critical; it was concluded that the number and width of bands could be adjusted to any values so long as they ensured coverage of the decade 16-160 Hz, without exceeding an upper limit of 200 Hz. An acceptable lower limit was not defined, but no band would have an upper limit below 20 Hz; analysis of a frequency region entirely below the commonly accepted limits of audibility would be unlikely to have any perceptual relevance.

2.5.2 Defining the Band Parameters

A number of factors were considered when defining the frequency bands:

- Quantity
- Spacing of centre frequencies
- Width
- Overlap
- Linear or logarithmic definition of limits

Any number of bands could be chosen to cover the specified frequency range, but the parameters listed above were not independent. For example, a fixed number of bands could not be defined without consideration of their width; few bands would be required to cover the specified frequency range if they were very wide. Quantity was therefore determined by the other parameters.

Bandwidth and spacing were considered jointly after reaching a decision about the amount of acceptable overlap. As described in section 2.4.3.2, it would be prudent to use bands that minimised this factor. Band interaction in the STI application was addressed by Steeneken and Houtgast [81], where it was believed that overlapping bands were introducing errors in the overall result through redundancy and mutual dependence, i.e. the result from a given band was affected by the one below it because they contained a proportion of the same information for at least part of the band. To minimise this effect in the current application, contiguous bands might be used, defined to give continuous but non-overlapping coverage of the required frequency range. The chosen parameters for band width and spacing would need to cover all points in the specified bass region i.e. bands either would overlap or be immediately adjacent.

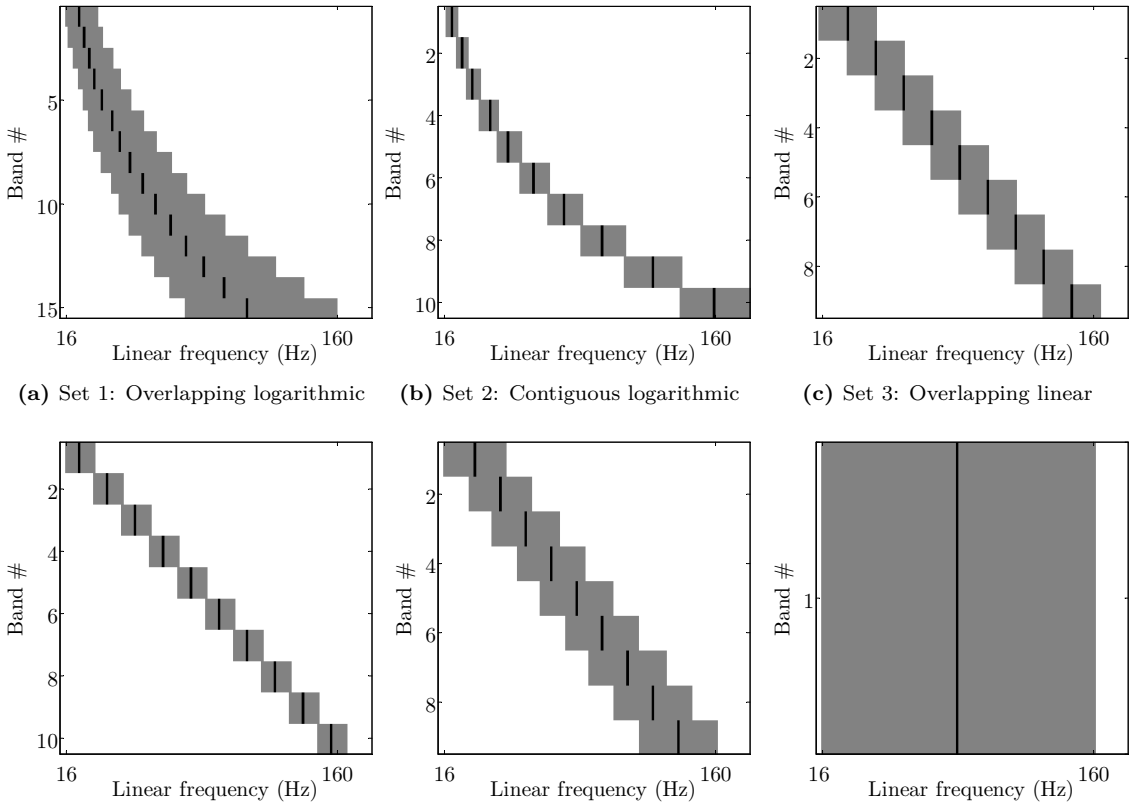
Logarithmic bands were considered, but there was no requirement to define the band width or spacing in this way. As the algorithm was focussed on low frequencies, octave-wide bands covered a large proportion of the overall range being investigated; conversely, bands defined in fractions of an octave covered a very narrow range at the lowest frequencies. The chosen arrangement had to ensure that even the lowest bands encompassed a range of frequencies that was seen to make a useful contribution to the MTF results.

2.5.3 Selecting a Band Arrangement

Six potential arrangements were developed to look for fundamental differences in the MTF results, given the various parameter options listed in section 2.5.2. Table 3 lists the band parameters; each set is illustrated in Figure 6, showing the width and relative spacing of each band. Note that Set 5 uses the equivalent rectangular bandwidth (ERB) for the specified centre frequencies. This is a model for the width of auditory filters in the human hearing system [97]; limits for this band arrangement were calculated using the VOICEBOX toolbox in MATLAB [98].

Set	Description	β definition	f_{cf} spacing	Coverage
1	Overlapping logarithmic	1 oct.	$1/6^{\mathrm{th}}$ oct.	16.1 - 158.8 Hz
2	Contiguous logarithmic	$1/3^{\mathrm{rd}}$ oct.	$1/3^{\mathrm{rd}}$ oct.	17.5 - 177.6 Hz
3	Overlapping linear	30 Hz	15.0 Hz	14.8 - 162.8 Hz
4	Contiguous linear	15 Hz	15.0 Hz	16.1 - 164.2 Hz.
5	ERB with linear spacing	ERB	13.5 Hz	16.5 - 160.8 Hz.
6	Single band	-	-	16.0 - 160.0 Hz.

Table 3: Summary of test band set parameters

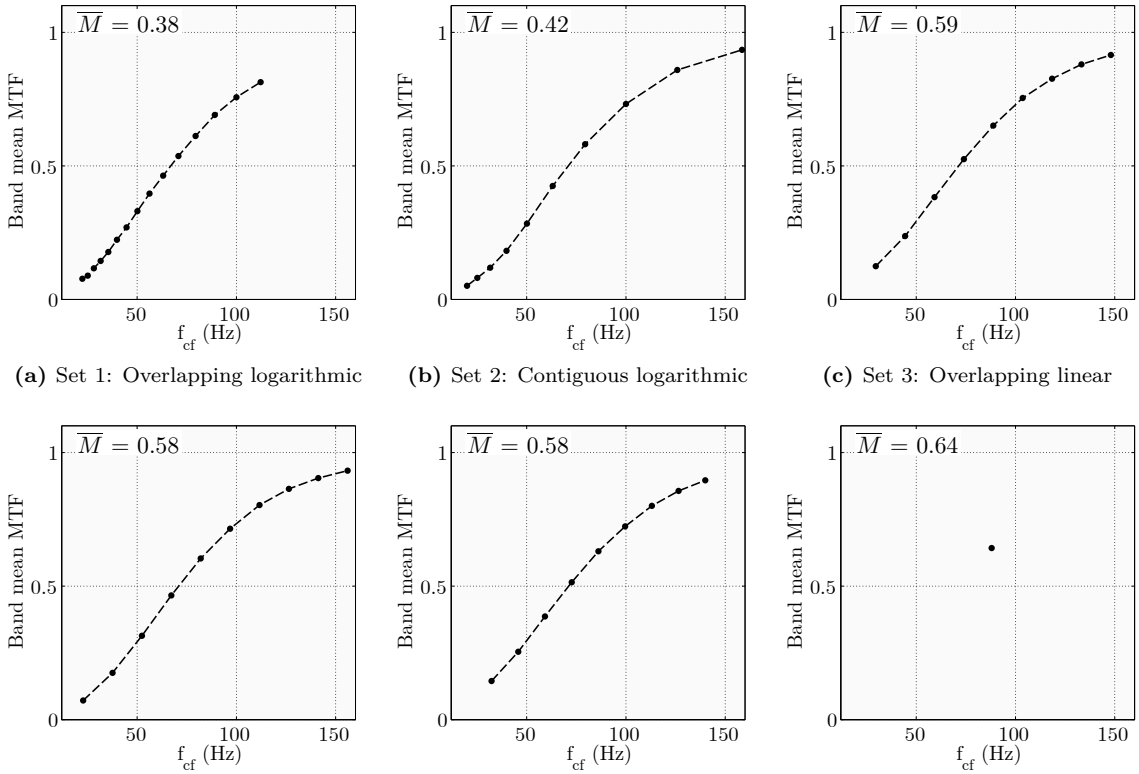


(a) Set 1: Overlapping logarithmic **(b)** Set 2: Contiguous logarithmic **(c)** Set 3: Overlapping linear

(d) Set 4: Contiguous linear **(e)** Set 5: ERB with linear spacing **(f)** Set 6: Single band

Figure 6: Six frequency band arrangements; black lines show the position of the band centre frequency

Each set of frequency bands was evaluated using the selected MTF method for computation. The same test systems and modulation frequencies described in section 2.4 were used. Results for System 1, $h_\delta(t)$, are not shown because all band sets produced $\overline{M} = 1$. Figure 7 shows the results from analysis of test system 2, $h_F(t)$; plots show the mean modulation index, $\overline{m}$, in each band, averaged across all modulation frequencies.



(a) Set 1: Overlapping logarithmic **(b)** Set 2: Contiguous logarithmic **(c)** Set 3: Overlapping linear

(d) Set 4: Contiguous linear **(e)** Set 5: ERB with linear spacing **(f)** Set 6: Single band

Figure 7: Comparison of MTF results for System 2, $h_{\mathrm{F}}(t)$, using six different band arrangements. Values for $\overline{M}$ are given to 2 d.p.

The following conclusions were made after comparing the MTF results for each band arrangement:

- Plots for the logarithmic bands, Sets 1 and 2, showed a 'clustering' towards the lower end of the overall frequency range due to the specification of logarithmic centre frequency spacing. This seemed to bias the overall mean score, reducing its magnitude compared to test sets where band centre frequencies were not logarithmically spaced. The contiguous band case, Set 2, seemed to be affected less by this, although the extended frequency coverage in this set also contributed to the increased mean score compared to the overlapping band version.

- The linear bands showed similar results, but the contiguous-band case, Set 4, produced a slightly greater range of scores overall for the individual bands. This was primarily seen in the lowest band where a score closer to zero was produced, as would be expected in this region of the test system's response. The plot for the overlapping band case, Set 3, showed a slightly flatter or compressed shape, suggesting that the use of overlapping and marginally wider bands in this set led to a loss of detail compared to the non-overlapping equivalent.

- The ERB bands, Set 5, showed behaviour very similar to that of the linear sets. The shape of the $\overline{m}$ distribution was most similar to Set 3; this might have been predicted, given that the band widths and centre frequency spacing of those two sets were very similar.

- Analysis with Set 6, a single band covering the specified frequency range, produced a greater mean score than any of the other sets. It seemed that this simplification was not reflecting degraded performance of the loudspeaker response at the lower frequencies, though the overall value for $\overline{M}$ (equal to $\overline{m}$ in this case) was closer to that of the linear bands than the logarithmic equivalents. It was concluded that this minimal set was of little use for comparing multiple loudspeakers, as the use of a single band prevented comparison of performance in different parts of the overall low frequency region.

Following this analysis, Set 4 was considered to be the most suitable arrangement for use in the MTF algorithm. The band width and spacing provided good coverage of the specified frequency range, appeared to show more detail in mean-band plots than the other sets, and returned an overall value for $\overline{M}$ that seemed appropriate for the alignment of the tested system.

2.6 Optimisation II: Modulation Frequencies

The use of different modulation frequencies in the MTF algorithm would allow assessment of whether a given monitor was able to accurately reproduce the envelope of a musical signal. As discussed in section 1.1.3.1, phase distortion in a loudspeaker's low-frequency alignment can change the envelope of a signal as it is reproduced. The individual frequency components may be reproduced with the same magnitude, but the relationship between them is altered. This degrades the impact of the musical presentation and makes it hard for mix engineers to create the correct balance between instruments.

The modulation frequencies used in preceding work by Holland *et al.* [64] were similar to those used in the STI for assessment of speech intelligibility. They appeared to be effective but had not been based on a formal investigation of the modulation components found in musical content at low frequencies. This section describes how musical signals were analysed to develop an appropriate set of modulation frequencies for use in the MTF algorithm.

2.6.1 Spectrum Calculation

If it was known in advance what musical content the loudspeaker under test would be reproducing, an exact set of modulation frequencies could be used in the algorithm. The MTF method being developed was intended for general application. Therefore, the aim was to find a generic set of modulation frequencies, suitable for analysing any mix monitor, regardless of which type of music an engineer usually works with. The objective was not to find an averaged spectrum for the frequency content of music, but rather the typical rate at which fluctuations in the envelope occur. This required detection of the envelope from a wide range of musical signals.

The envelope is commonly detected in communication applications using the Hilbert transform. For function $h(t)$, the Hilbert transform is represented by $h_{\mathrm{H}}(t) = \mathscr{H}\{h(t)\}$ and is found from the analytic signal $h_{\mathrm{A}}(t)$ [99]:

$$h_{\mathrm{A}}(t) = h(t) + \mathrm{j}h_{\mathrm{H}}(t) \tag{2.35}$$

The magnitude of the analytic signal returns the envelope of the original signal:

$$|h_{\mathrm{A}}(t)| = \sqrt{h^2(t) + h_{\mathrm{H}}^2(t)} \tag{2.36}$$

A method for detection of musical modulation frequencies was described by Polack *et al.* [74]; an alternative technique was developed, sharing some similarities with that work, which was simple to execute and extremely accurate. The process is summarised below:

- Extract an audio waveform, $x_m(t)$ from one channel of a CD-quality .wav file; low-pass filtered above 200 Hz to produce the file for processing, $x_{\mathrm{F}}(t)$. Attenuation of all other frequency content ensures that only modulation components due to the low frequencies are present.

- Find the envelope using the Hilbert transform:

$$\tilde{x}_{\mathrm{F}}(t) = |\mathscr{H}\{x_{\mathrm{F}}(t)\}| \tag{2.37}$$

  Figure 8 shows part of an example waveform and its envelope.

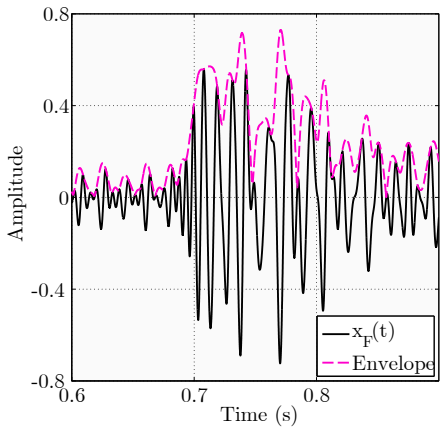


Figure 8: Musical extract envelope example. Plot shows 0.30 s of each signal; $x_F(t)$ is the musical extract after low-pass filtering below 200 Hz

- Calculate the modulation spectrum:

$$\tilde{X}_{\mathrm{F}}(t) = \mathscr{F}\{\tilde{x}_{\mathrm{F}}(t)\} \tag{2.38}$$

  As the amplitude envelope of the signal in the time domain tracks changes in the music, the Fourier transform of this shows the corresponding modulation frequencies. Figure 9 shows an example envelope spectrum. Discrete frequency components have been plotted with 0.1 Hz resolution up to 100 Hz; it can be seen that some components are more prominent than others. Three such components have been marked with asterisks in the plot, located at 0.5, 1.3, and 4.0 Hz.

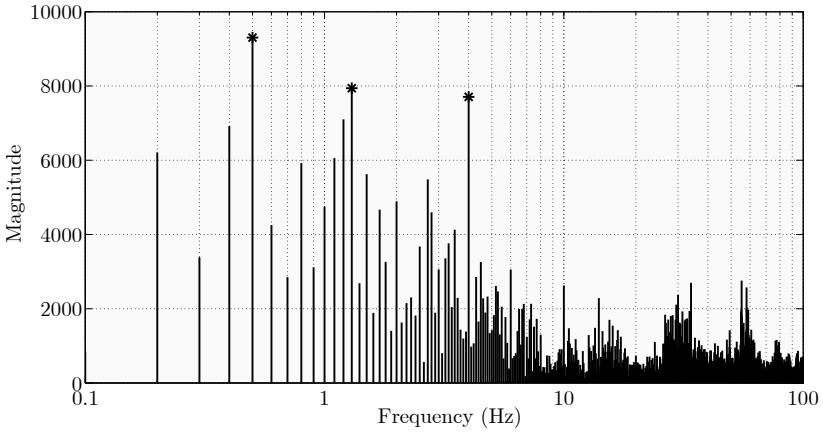


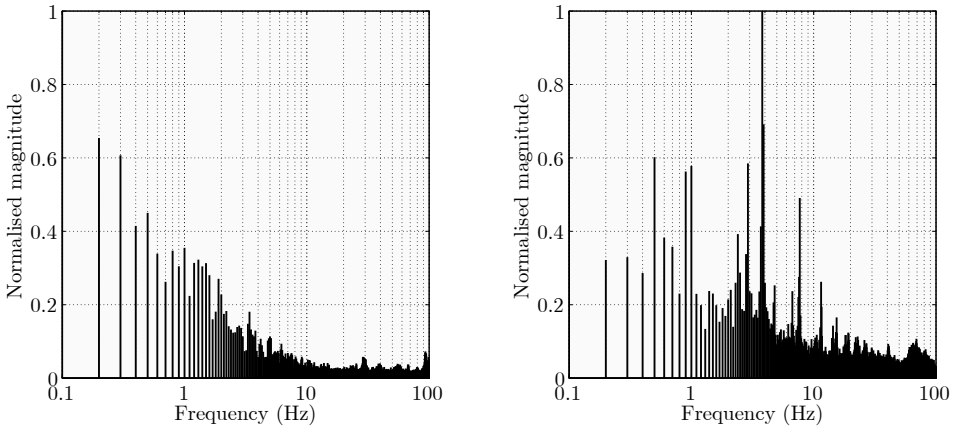
Figure 9: Example modulation spectrum calculated from a single musical envelope segment. Asterisks mark the most prominent modulation components

An ensemble summation process was used to create the final spectrum of modulation components for each musical extract. This was achieved by dividing the filtered extract $x_{\mathrm{F}}(t)$ into L successive non-overlapping segments, where segment length N was selected to give a frequency resolution of 0.1 Hz in the subsequent Hilbert and Fourier transforms. All extracts were either 2 or 3 minutes in duration, leading to $L = 12$ or $L = 18$ respectively:

$$\left\langle \tilde{X}_{\mathrm{F},1:L}(f) \right\rangle = \sum_{k=1}^{k=L} \left| \tilde{X}_{\mathrm{F},k}(f) \right| \tag{2.39}$$

where: $\widetilde{X}_{F,k}(f)$ is the envelope spectrum of extract $x_F(t)$ in segment k.

Figure 10 shows the envelope spectrum from two different extracts, illustrating how the results were seen to vary across different source material.



(a) Extract 1: Envelope spectrum, 2 minute extract; $\left\langle \tilde{X}_{\mathrm{F},1:12}(f) \right\rangle$

(b) Extract 2: Envelope spectrum, 3 minute extract; $\left\langle \tilde{X}_{\mathrm{F},1:18}(f) \right\rangle$

Figure 10: Envelope spectra from two different musical extracts

It can be seen in Fig. 10 that these two extracts produce distinctly different modulation profiles. This is due not to the different number of segments used for computation, but to the different temporal characteristics of the music they represent. It should be noted that the use of different length extracts here was only performed due to the later addition of extra tracks; these were either too short (full song less than three minutes), or contained passages where the temporal or spectral content differed significantly from the rest of the track. For consistency with the rest of the test set, the analysis could have been repeated whilst making the duration of all extracts equal to the length of the shortest sample.

2.6.2 Selected Modulation Frequencies

The aim of this investigation was to develop a 'generic' modulation frequency spectrum that was based on a range of musical styles. Summation of envelope spectra was used as a way to average results from many different musical examples; assuming a sufficient number of extracts from a range of genres was used, the cumulative result would form a reliable estimate of the modulation frequencies occurring most commonly in typical examples of recorded music. It was expected that the 'noise' of non-common modulation frequencies would be dominated by peaks formed from summation of components that occurred repeatedly across many excerpts of music (only the relative magnitude of frequency components was of interest here, so a summation was sufficient).

Envelope spectra from 168 different musical extracts were summed to produce the final spectrum of musical modulation frequencies. The aim in generating this spectrum was to inspect the range of modulation frequencies present in typical programme material, then select ones that gave useful algorithm results. Note that the term 'useful' here meant selecting values that revealed differences between loudspeakers and were representative of typical programme material. Any values picked from the final envelope spectrum were considered to satisfy the latter criterion as the results were based on a range of genres, taken from commercially available recordings. Choosing values considered to be most revealing was not straightforward.

The source material used to generate the final spectrum was broadly categorised into genres; Figure 11 shows the approximate distribution of musical styles used. It can be seen that rock and pop contained the most extracts but, intentionally, no single category dominated the selection. A full listing of tracks used and their assigned genres is given in Appendix B.

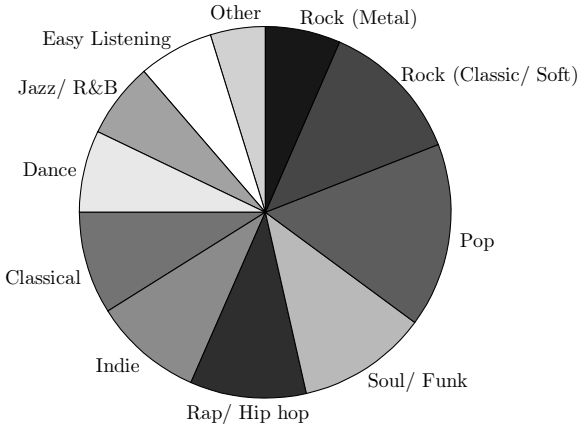


Figure 11: Distribution of genres used to calculate musical modulation frequency spectrum

The final modulation frequency spectrum calculated from this material is shown in Figure 12.

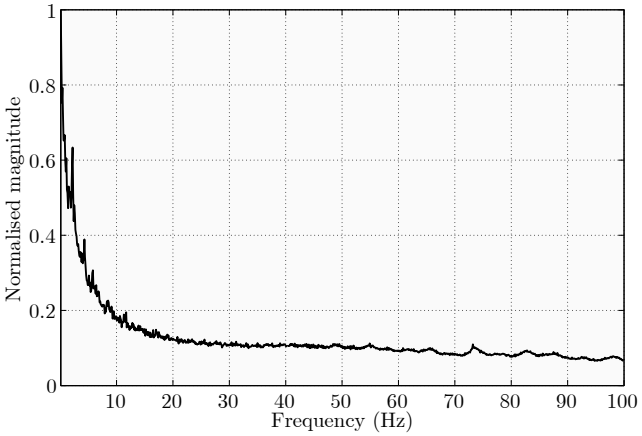


Figure 12: Musical modulation frequency spectrum, $Y_{\mathrm{mf}}(f)$; based on signal content below 200 Hz of 168 different extracts. Frequency resolution is 0.1 Hz.

Selecting a small number of revealing values was desirable to make computation faster and allow clearer presentation of the visual results (discussed in section 2.9), but choosing a limited number of values to include in the algorithm was not easy given the result shown in Fig. 12. The final envelope spectrum showed a near-continuous distribution of values, indicating that very few modulation frequencies are common to different types of music. This 'generic' envelope spectrum showed an exponential decay with few dominant modulation frequencies up to approximately 20 Hz, with a low-level, even distribution above this point. Small peaks, dominant modulation components, sometimes appeared to stand out from the surrounding region. One such peak can be seen at 2.2 Hz in Fig. 12, corresponding to a temporal fluctuation in a signal's envelope of period 0.45 s. The prominence of this peak implies that envelope fluctuations at this rate are common to many styles of music. Musical tempo is often described in terms of beats per minute (BPM); a period of 0.45 s returns approximately 133 BPM, which is a typical mid-tempo speed for many types of music [100–102]. The sample of test extracts used here to identify musical modulation frequencies contained a large proportion of pop, rock, and dance music (as shown in Fig. 11). The strong peak around 2.2 Hz therefore agrees with existing knowledge about musical tempo and suggests that the method was, as intended, identifying dominant components of temporal structure present in the musical extracts. This is not surprising given that envelope analysis has been used elsewhere for automatic detection of musical tempo, although more sophisticated algorithms are now typically used for that application [102].

The method of calculating the envelope spectrum meant that the modulation frequencies occurring most often in the analysed extracts resulted in the highest normalised magnitude values in the final spectrum; as shown in Fig. 12, the general trend in the data was that of greater occurrence as frequency decreased. The initial decision was to select only the lowest modulation frequencies for computation because the envelope spectrum showed that these

occurred most often in the music analysed. However, there was concern about using only values spanning such a restricted range; the analysis showed that the distribution did not plateau until approximately 20 Hz, although the prominence of envelope fluctuations at this rate were in the order of only 10 % relative to the lowest values.

As illustrated in Figure 13, a separate analysis showed that more distinct envelope spectra were returned when the extracts were grouped by specific genres. The extracts used in this analysis correspond to the following item numbers as listed in Appendix B: *Dance:* 16–25; *Rock:* 103–112; *Rap:* 114–121, 125, 130.

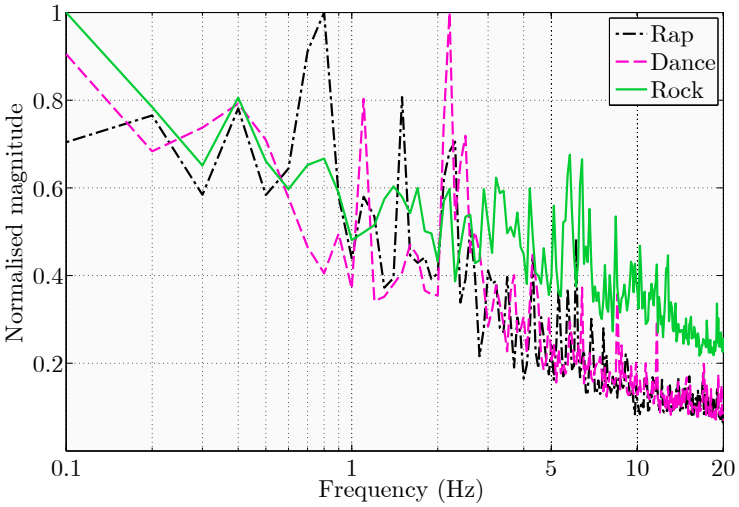


Figure 13: Genre-specific envelope spectra. Results are based on ten 2-minute extracts from each genre

The assumption was made that extracts from the same genre would be likely to possess similar spectral and temporal characteristics that would make them likely to have more similar envelopes than extracts of different musical styles. The distinct differences between the resulting envelope spectra appear to support this assumption. It can also be seen that two of the examples presented in Fig. 13 show some harmonic behaviour, with *Rap* having a fundamental of approximately 0.75 Hz with strong harmonics at 1.5 and 2.25 Hz, and *Dance* having prominent peaks at 1.1 and 2.2 Hz. The reason for this behaviour was not investigated, but it may be related to timbral and polyphonic complexities of the test signals; 'local periodicities' of this nature are commonly encountered in musical feature analysis, usually in relation to tempo extraction, and are most prominent in spectrally-complex music with a strong and very consistent beat [103].

It must be noted that this survey was conducted on only ten 2-minute extracts from each style and was not intended to be a definitive classification of envelope spectra for different genres; this would require analysis of many more extracts and formal assessment of waveform characteristics. However, the different profiles returned in this investigation demonstrated that selection of modulation frequencies for the algorithm should not be restricted to only the very lowest values. A decision was therefore made to select values across a wider range of the distribution but below 20 Hz; the following points were chosen:

$$f_m = [0.8,\ 1.1,\ 2.2,\ 4.3,\ 5.8,\ 8.5,\ 11.7]\,\text{Hz}.$$

Figure 14 shows the spectrum $Y_{\text{mf}}(f)$ again, plotted to allow closer inspection of the range below 20 Hz; solid vertical lines mark the chosen modulation frequencies.

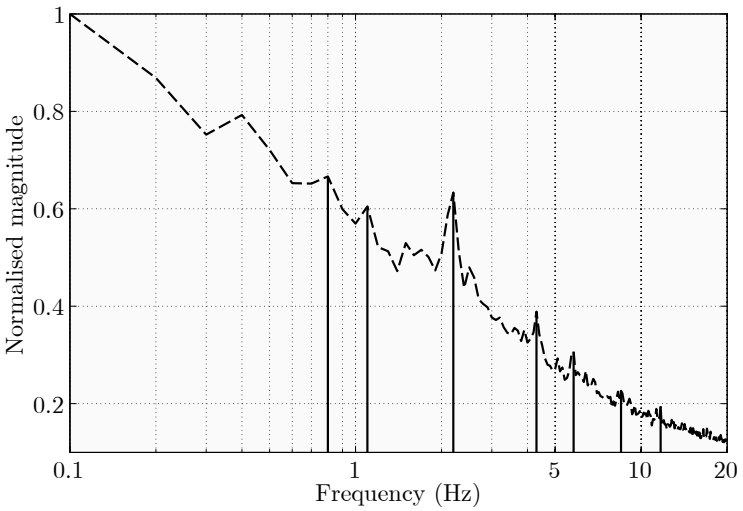


Figure 14: Generic musical envelope spectrum up to 20 Hz. The y-axis begins at 0.1 to show the approximate value where the distribution plateaus. The final spectrum $Y_{\text{mf}}(f)$ is shown by the dashed line; chosen values are marked by vertical solid lines

It can be seen that the selected frequencies were chosen at points corresponding to peaks relative to the surrounding region in the distribution. It was not obvious how to choose the number of peaks to include, but with reference to Fig. 14, it can be seen that the number of selected points throughout the defined range of the distribution reduces with increasing frequency:

Range	Number of selected points
$1 \leqslant f_m < 5\,\text{Hz}$	3
$5 \leqslant f_m < 10\,\text{Hz}$	2
$f_m \geqslant 10\,\text{Hz}$	1

This was an attempt to reflect the trend observed in the data without restricting the chosen values to a very limited range at the lowest frequencies. The exception was the range below 1 Hz, where only a single value was selected. To define a lower limit for f_m, algorithm results were compared for 25 real mix monitors[§]. The MTF calculation was performed using the chosen method, as described in section 2.3.3. The analysis used the bands defined in section 2.5, but

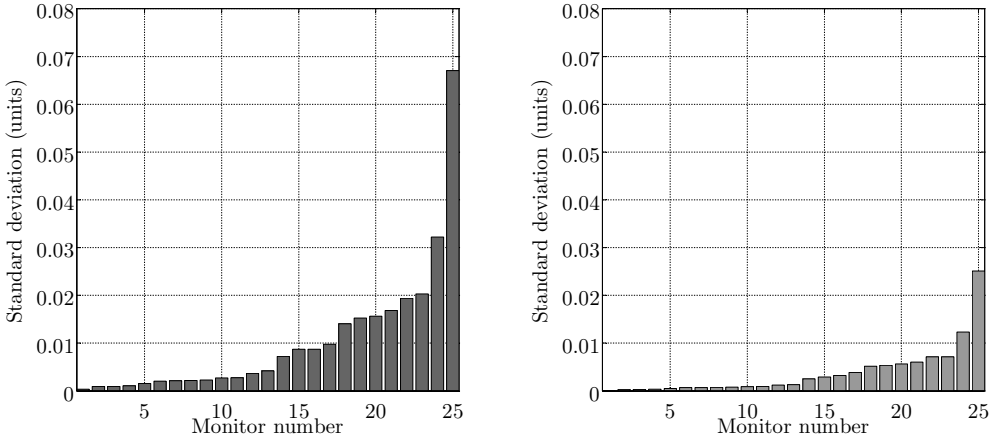
[§]Note that existing measurements were used for this analysis and the validation presented in section 2.9; further details about these measurements are given in section 3.2.2.

only four values for f_m: 0.1, 0.2, 0.4, and 0.8 Hz; with reference to Fig. 14, these were chosen as they appeared to be the location of peaks in the envelope spectrum below 1 Hz. The aim of analysis was to understand whether more of the lowest frequencies should be included, or whether the chosen lower limit of 0.8 Hz, corresponding to a temporal fluctuation with period 1.25 s, was sufficient. The mean score generated with each of the modulation frequencies was calculated as:

$$\overline{m}_v = \frac{\sum_{q=1}^{n} m_{q,v}}{n} \tag{2.40}$$

where: $m_{q,v}$ is the individual modulation index returned for band q and modulation frequency v, and n is the total number of bands; in this case, $n = 10$. The mean result, $\overline{m}_v$, therefore took four values for each of the tested loudspeakers: $\overline{m}_{0.1}$, $\overline{m}_{0.2}$, $\overline{m}_{0.4}$, and $\overline{m}_{0.8}$.

The variation in these mean values was compared for the selection of measured monitors; the standard deviation, σ, across the four values of $\overline{m}_v$ for each loudspeaker was calculated and the results for all of the monitors are plotted in Figure 15a, sorted by increasing σ to aid comparison. These values show how much the band-averaged modulation indexes differ as a result of using the different (single) modulation frequencies. It was proposed that if the variation was consistently low, it would indicate that inclusion of the additional very low modulation frequencies was not especially useful in discriminating between different systems– if the mean score was the same regardless of the modulation frequency used, any of them could be used in place of the other because they returned similar information about the system. It can be seen from Fig. 15a that the standard deviation was below 0.07 for all monitors, and below 0.04 for 24 out of 25. Considering the maximum theoretical range of $0 \leqslant m \leqslant 1$, the variation due to use of these lowest modulation frequencies was up to approximately 7 % of the total score. Monitor 25 showed a variation across the four modulation frequencies that was more than double any of the others. Further investigation showed that the results for 0.8 Hz were causing most of this deviation; Figure 15b shows the standard deviation for the sample of monitors after removing the results for 0.8 Hz.



(a) Results when $f_m = 0.1$, 0.2, 0.4, and 0.8 Hz.

(b) Results when $f_m = 0.1$, 0.2, and 0.4 Hz.

Figure 15: Standard deviation, σ, of mean score, $\overline{m}_v$, across modulation frequencies below 1 Hz. The bars show σ for results from 25 real measured monitors. Each value of $\overline{m}_v$ was calculated from results in ten frequency bands, covering the range 16–160 Hz.

After excluding the results for 0.8 Hz, the maximum σ within the sample of monitors reduced to 0.025. It was therefore concluded that inclusion of the lowest modulation frequencies increased computation time without providing a substantial increase in discrimination between different systems. The lowest modulation frequency was therefore kept at 0.8 Hz.

2.7 Optimisation III: Modulating Function

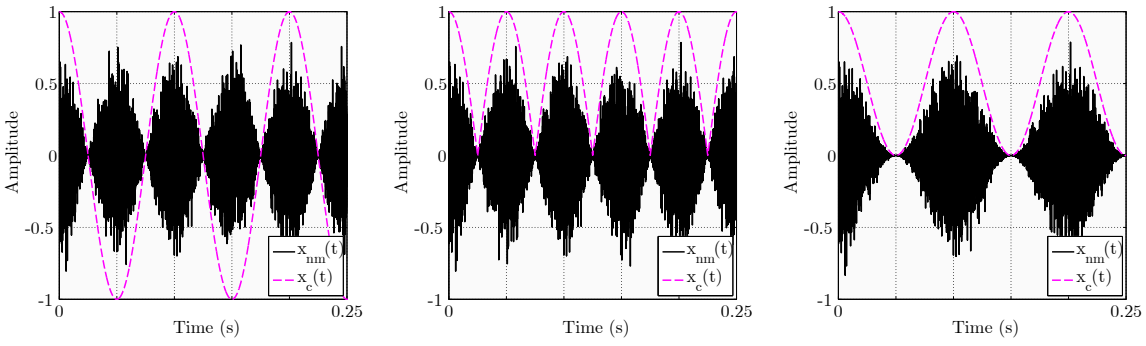
This section describes investigation of the function used to modulate the test signal. With reference to the discussion in section 2.1.1, it was necessary to choose a function that produced a distinct intensity envelope in an amplitude-modulated noise signal, with fluctuations at the rate corresponding to f_m. As discussed in section 2.2.2.1, the use of suppressed-carrier AM meant that amplitude of the modulating function relative to the original noise signal was not critical. Three modulating functions were compared to see how sensitive the algorithm was to this parameter, and whether there might be any advantage in using a function other than the one used up to this point:

i) Cosine: $x_{\mathrm{c}}(t) = \cos(2\pi f_{\mathrm{m}} t)$

ii) Absolute-value cosine: $x_c(t) = |\cos(2\pi f_{\mathrm{m}} t)|$

iii) Shifted-cosine: $x_{\mathrm{c}}(t) = 0.5\,[1 + \cos(2\pi f_{\mathrm{m}} t)]$

The functions and the results after modulation of a noise signal, $x_{\mathrm{n}}(t)$, are illustrated in Figure 16, where $x_{\mathrm{nm}}(t) = x_{\mathrm{n}}(t)x_{\mathrm{c}}(t)$.



(a) Modulating function 1 **(b)** Modulating function 2 **(c)** Modulating function 3

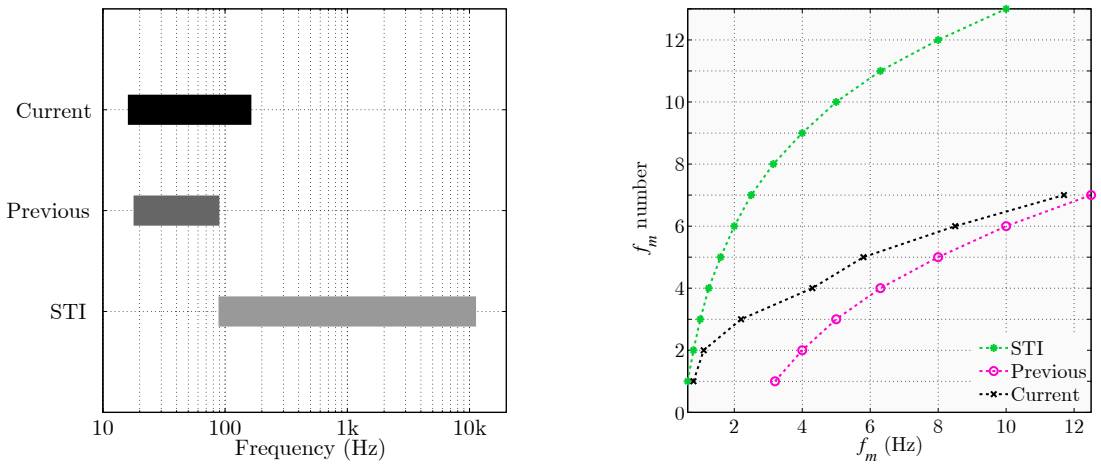
Figure 16: Modulating functions, $x_c(t)$, and modulated noise waveforms, $x_{\mathrm{nm}}(t)$; $f_{\mathrm{m}} = 10$ Hz

Inspection of the modulated noise signals showed that functions i) and ii) produced fluctuations at twice the wanted rate, so $0.5f_{\mathrm{m}}$ was used in their MTF analysis. The band-mean MTF scores generated from each modulating function were compared; the greatest deviation in $\overline{m}$ was 0.01, with no difference in the overall mean scores to 2 d.p.

No function appeared to cause a significant change in results, so function iii) was retained for further use as it produced a modulated noise input signal with required intensity envelope without any adjustment of the modulation frequency.

2.8 Comparing Parameters With Previous Work

Figure 17 shows how the chosen analysis range and modulation frequencies compare to those used in the preceding work by Holland *et al.* [64]. The values used in STI are also shown here for comparison, demonstrating how the parameters for an MTF-based application focussed on reproduction of music at low frequencies differ from those of a method evaluating speech. Figure 17a compares the extent of the frequency bands (developed in section 2.5); it can be seen that the current algorithm covers a wider analysis range than the previous method. This plot also highlights the issue discussed in section 2.4.3.1, where the impact of system band-limiting in the STI and current application differed considerably due to the much narrower bandwidths in absolute terms of the latter. Fig. 17b compares the modulation frequencies (developed in section 2.6); it can be seen that the current method uses the same number of modulation frequencies as applied in previous work, but their extent is more similar to those used in STI.



(a) Extent of analysis range

(b) Quantity and value of modulation frequencies

Figure 17: Comparing analysis range and modulation frequencies from current, previous, and STI MTF-based methods

2.9 Algorithm Output

A primary objective for the algorithm was to develop a method that would allow results to be compared across multiple loudspeakers, presented in a format that was intuitive and simple to understand. The algorithm therefore presented results in three different ways:

i) Mean MTF score; the unweighted average across all matrix elements:

$$\overline{M} = \frac{\sum_{k=1}^{n} m_k}{n} \tag{2.41}$$

where: m_k is the k^{th} modulation index, and n is the total number of MTF matrix elements.

ii) Band-mean scores, showing the average modulation index across all modulation frequencies within a band:

$$\overline{m}_q = \frac{\sum_{k=1}^{p} m_k}{p} \tag{2.42}$$

where: $\overline{m}_q$ is the mean modulation index in the q^{th} frequency band, and p is the total number of modulation frequencies.
These values would be plotted to allow visual comparison, as presented in Figure 7. Newell and Holland [35] used this form of presentation, referring to them as 'low frequency quality plots'.

iii) MTF intensity image. This method was developed to allow easy visual inspection of the entire MTF matrix; it is summarised in subsection 2.9.1.

2.9.1 Visual Representation

It was not initially known whether fine detail in the MTF matrix would be useful, or whether the mean scores $\overline{M}$ and $\overline{m}$ were sufficient to characterise the behaviour of a loudspeaker's low-frequency alignment and provide sufficient discrimination when comparing multiple systems. It was found that direct comparison of the MTF matrices became cumbersome when the number of elements increased beyond a few combinations of modulation frequency and band. A method was developed to display larger matrices in visual form. Three requirements for this method were defined:

i) Content should be presented in grayscale to allow printing without colour, such as on data sheets.

ii) Format should be intuitive, not requiring specialist knowledge to interpret.

iii) Presentation must allow direct comparison of multiple loudspeakers without requiring any conversion or compensation.

The MTF matrix is well suited to the intensity image format for two reasons. Firstly, a given algorithm, having a fixed number of modulation frequencies and bands, will have a fixed number of elements; this permits a layout that can be presented in a consistent format. Secondly, the theoretical range $0 \leqslant m \leqslant 1$ for the individual matrix elements presents a clear basis for development of a suitable colour scheme.

The MATLAB `imagesc` command was used to produce the MTF intensity images. A linear grayscale colormap was generated which scaled the individual modulation index values across a fixed range, allowing direct comparison of different systems. In this format, $m = 1$ is white, and $m = 0$ is black, with values falling between these extremes represented by grey; therefore, a higher MTF score is portrayed as a whiter image. It was found that a 20-row colormap matrix was the best compromise between accuracy and visual discrimination of individual values, particularly when printing; this provides mapping of individual m-scores to approximately the nearest 0.05. The format is illustrated in section 2.11.

2.10 Averaging for Consistency of Results

It was described in section 2.4.2 that using noise as a test signal led to variability in results that required averaging to reach a consistent mean score. This section gives more detail on the

problem and the solution that was implemented in this method.

Results are calculated from the input and output intensity envelopes. Each combination of frequency band and modulation frequency returns a single modulation index score, m. Smoother envelopes produce more consistent results; smooth envelopes are obtained through an ensemble averaging method. A random noise signal was generated with a duration equal to twice the period of the lowest modulation frequency, 0.8 Hz: $2T_{\min} = 2.5\,\text{s}$. A sampling frequency of $f_s = 44.1\,\text{kHz}$ was assumed when generating the noise, making the duration of each extract equal to 55125 samples:

$$N_{\text{nz}} = \frac{1}{f_{m\,\min}} f_s \tag{2.43}$$

where: N_{nz} is the duration of each noise extract in samples, $f_{m\,\min}$ is the lowest modulation frequency, equal to 0.8 Hz, and f_s is the sampling frequency.

To compute results, segments of length equal to one period of each modulation frequency were extracted and averaged. These segments therefore ranged in length between:

$$L_{\max} = \frac{N_{\text{nz}}}{2} \tag{2.44}$$

$$L_{\min} = \frac{1}{f_{m\,\max}} f_s = \frac{f_s}{11.7} \tag{2.45}$$

where: $L_{\min}$ and $L_{\max}$ are the shortest and longest segment lengths, determined by the period of the modulating function; the highest modulation frequency, $f_{m\,\max}$, therefore returned the shortest envelope segment length, at $L = 3768$ samples. This method allowed envelope averaging across each noise signal, or iteration. To increase the number of averages per noise extract, as many segments as possible within the fixed duration were extracted and averaged. Figure 18 illustrates this process.

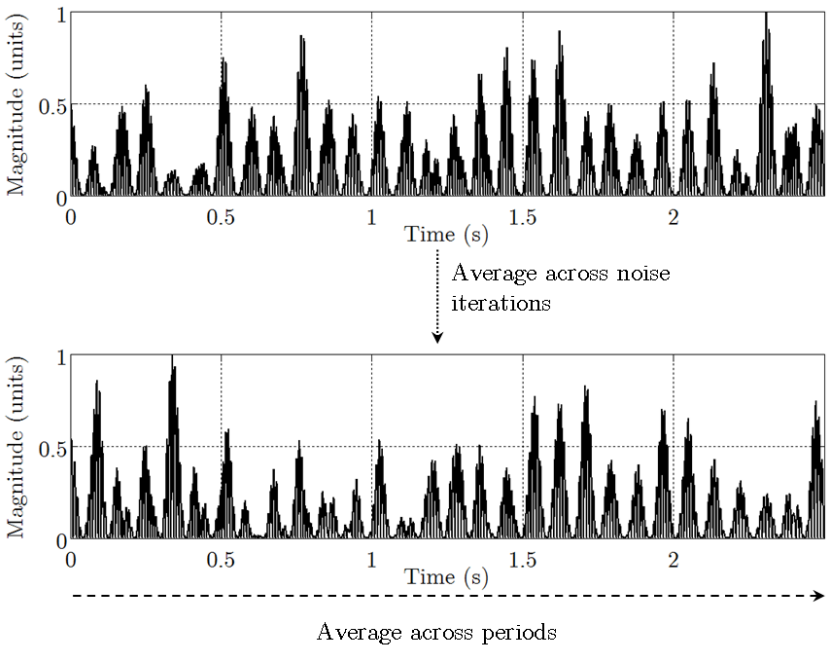


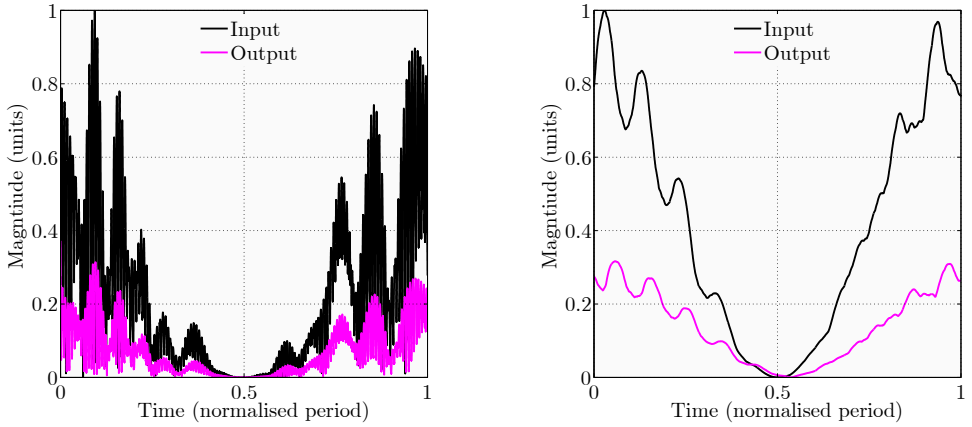
Figure 18: Illustrating the averaging process. Input and output intensity envelopes were repeatedly averaged for different instances of modulated noise. Averaging was then performed across multiple periods of the modulation frequency

The problem with this method is that the lower modulation frequencies, having a longer period, are subject to less averaging:

$$f_{m,\min} = 0.8\,\text{Hz};\ T_{\min} = 1.25\,\text{s};\ \rightarrow 2 \text{ averages};$$

$$f_{m,\max} = 11.7\,\text{Hz};\ T_{\max} = 0.09\,\text{s};\ \rightarrow 29 \text{ averages};$$

Figure 19 shows results calculated from a single instance of noise. Results are based on test system 2, calculated in a single band (band 3, 45 to 60 Hz). Results are shown for the lowest and highest modulation frequencies to illustrate the range of performance.

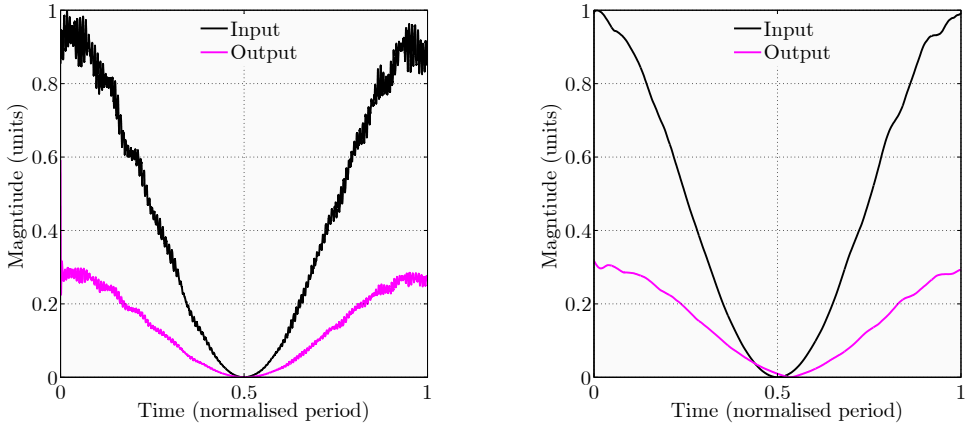


(a) $f_m = 0.8\,\text{Hz}$: envelope averages from 2 periods of f_m

(b) $f_m = 11.7\,\text{Hz}$: envelope averages from 29 periods of f_m

Figure 19: Input and output envelopes formed from a single noise iteration

Averaging across many instances of noise was chosen as the primary method of averaging for computational reasons; storing arrays long enough to average across tens of periods of $f_{m,\text{min}}$ for 70 combinations of f_{cf} and f_m for both input and output required a large amount of memory. Figure 20 shows results of the same system in Fig. 19, but calculated from 100 noise averages. It can be seen that much smoother envelopes, from which the result m is calculated, are formed when averaging in this way.



(a) $f_m = 0.8\,\text{Hz}$: envelope averages from 2 periods of f_m

(b) $f_m = 11.7\,\text{Hz}$: envelope averages from 29 periods of f_m

Figure 20: Input and output envelopes formed from 100 noise iterations

The number of noise instances required to reduce erratic behaviour in the input and envelopes to an acceptable level was investigated. The limit defined for consistency was a variation in $\overline{M}$ of no more than 1 % of the maximum theoretical score ($\overline{M} = 1.00$), i.e. the final mean matrix result

must not vary by more than 0.01 across repeated evaluations of the same test system. In order to evaluate this, it was necessary to repeatedly perform analysis with a fixed test system. A given loudspeaker should produce identical MTF results every time, assuming that the algorithm is run on the same measured data, but it had been observed that this was not the case due to the use of a noisy test signal; repeating analysis with a fixed test system therefore allowed evaluation of how much the mean MTF score varied due to this issue, and how the variation could be expected to reduce if more averaging is performed. For this analysis, results from a real measured monitor were repeatedly compared and the mean-score variation analysed. Results were computed 30 times, each using a different number of noise averages; the envelope averaging already described was also performed and not adjusted. Five different numbers of noise averages were compared: n_{it} =10, 100, 500, 1000, 1500. The standard deviation in mean score, $\sigma_{\overline{M}}$, in each case is compared in Figure 21.

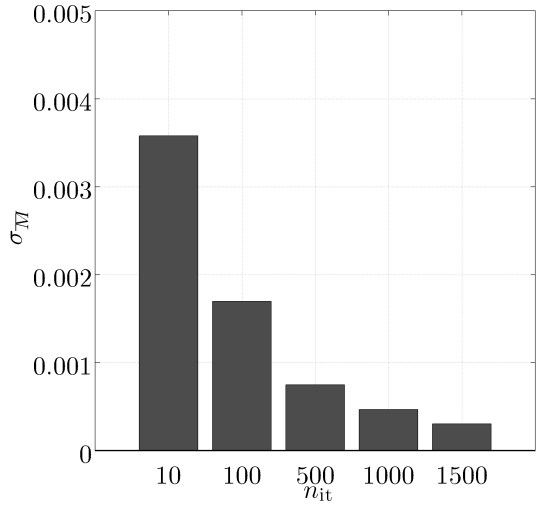


Figure 21: Five results for $\sigma_{\overline{M}}$, the standard deviation in mean MTF matrix score ($\overline{M}$) across 30 evaluations of the same loudspeaker. The bars show the value of $\sigma_{\overline{M}}$ observed when increasing the number of noise averages, n_{it}, used to calculate MTF results.

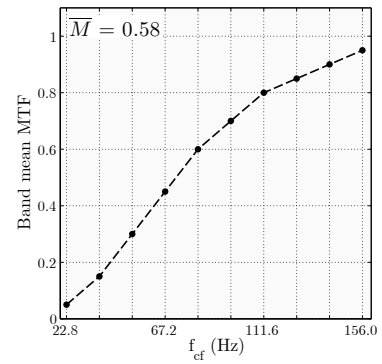
It is seen that increasing the number of noise averages always reduced variation in the mean score, but the largest reduction was achieved when n_{it} increased from 10 to 100: $\sigma_{\overline{M}} = 0.004$ and $\sigma_{\overline{M}} = 0.002$ respectively. In all cases, the variation was within the specified limit of 0.01. Based on this data, it was concluded that 100 noise averages would be used a compromise between stability of results and computation time; this was therefore implemented in the final algorithm, but it was noted that as few as 10 averages could be used for a quick estimate of results.

It has been shown that the main issue that motivated this averaging method was a limitation of computer processing speed and memory; it is therefore considered to be an aspect of the algorithm that may be improved and refined in the future. It should be noted that this limitation is primarily due to the way in which the calculation was executed here; the implementation relied on storage of long arrays in a matrix before calculating final results, i.e. all data required to compute every combination of modulation frequency and frequency band was generated for input

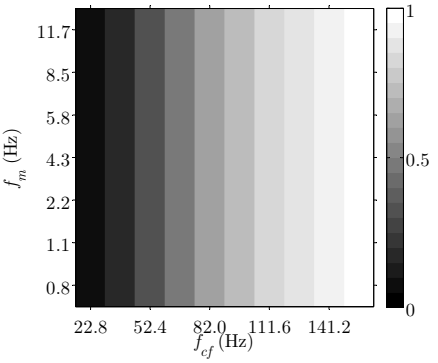
and output envelopes before computing the final modulation index values. This approach made the process conceptually simpler to follow but is computationally very inefficient. Computing the final MTF matrix scores in an element-wise process instead, deleting intermediate data required for calculation after each iteration, would reduce the reliance on memory; the additional overhead of using longer arrays (time histories form which the modulation envelopes are derived) would then no longer be a significant issue. Regardless of the procedure used to perform the envelope averaging, it is suggested that the consistency analysis needs further investigation to confirm that the number of averages is sufficient; this must be performed with a large set of real monitor measurements having a range of different alignments, ideally covering the extremes of performance that might be expected in real mix monitors.

2.11 Validation with Test Systems

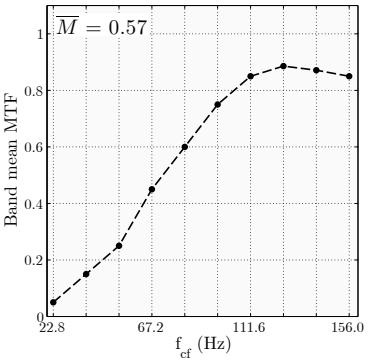
The algorithm was tested on a group of real measured loudspeaker responses, plus the test systems $h_\delta(t)$ and $h_F(t)$. As expected, the perfect system $h_\delta(t)$ returned a perfect MTF i.e. $\overline{M} = 1.00$. Measured monitor 2 shown here was the system used for the stability analysis presented in section 2.10. Results for the other systems are presented in Figure 22.



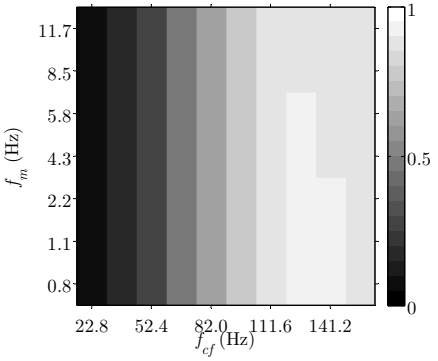
(a) Test system 2, $h_F(t)$



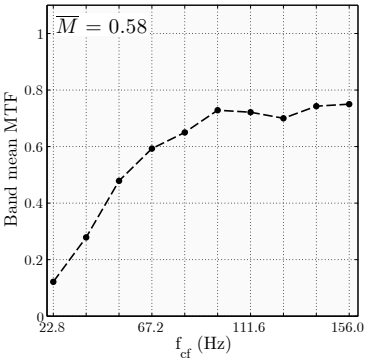
(b) Test system 2, $h_F(t)$



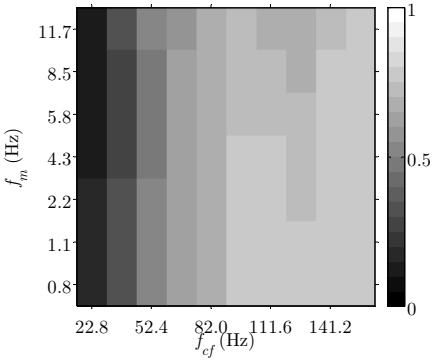
(c) Measured monitor 1



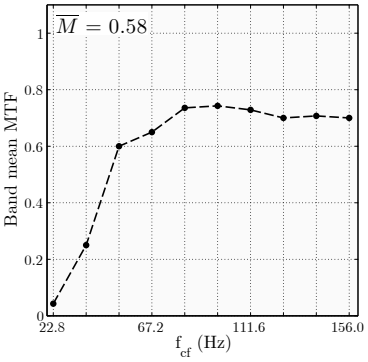
(d) Measured monitor 1



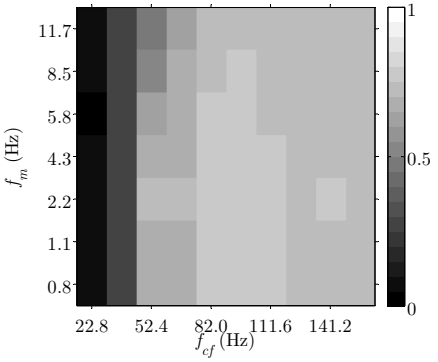
(e) Measured monitor 2



(f) Measured monitor 2



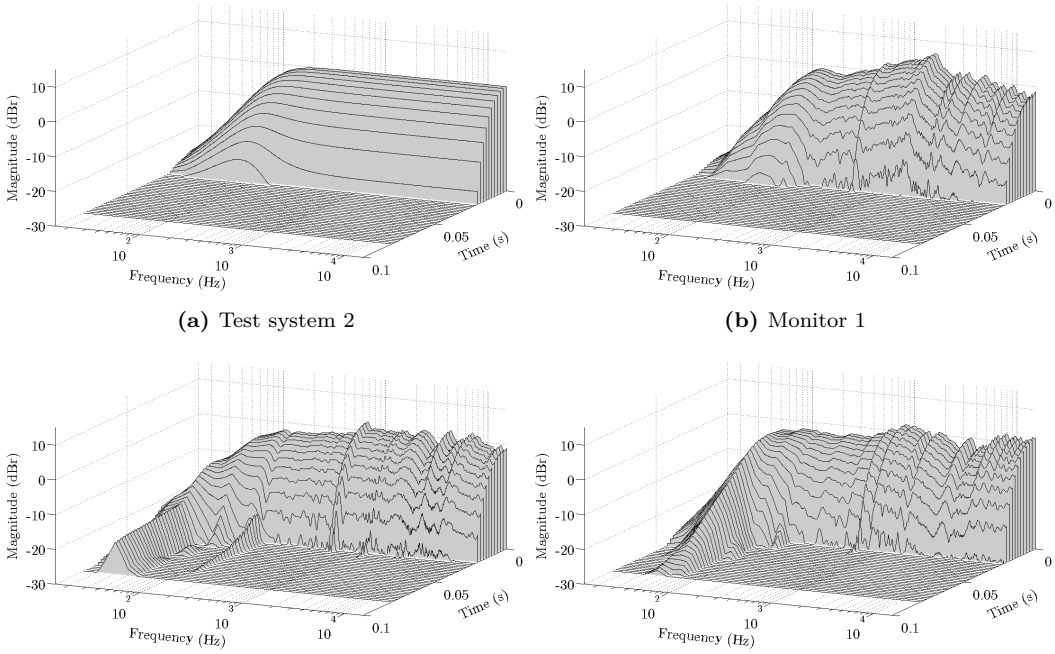
(g) Measured monitor 3



(h) Measured monitor 3

Figure 22: Validation of the MTF algorithm: Results from four different loudspeakers

To decide whether these results were appropriate, the systems were compared using more established objective measures. Figure 23 shows the waterfall plots for these systems. The plots for measured monitors 1 to 3 are consistent with those presented by Newell *et al.* [34] for the same systems[¶]; however it is difficult to directly compare waterfall plots if the processing parameters, and in this case, the viewing angle, are not identical (this issue is described in more detail in section 8.4.1).



(a) Test system 2

(b) Monitor 1

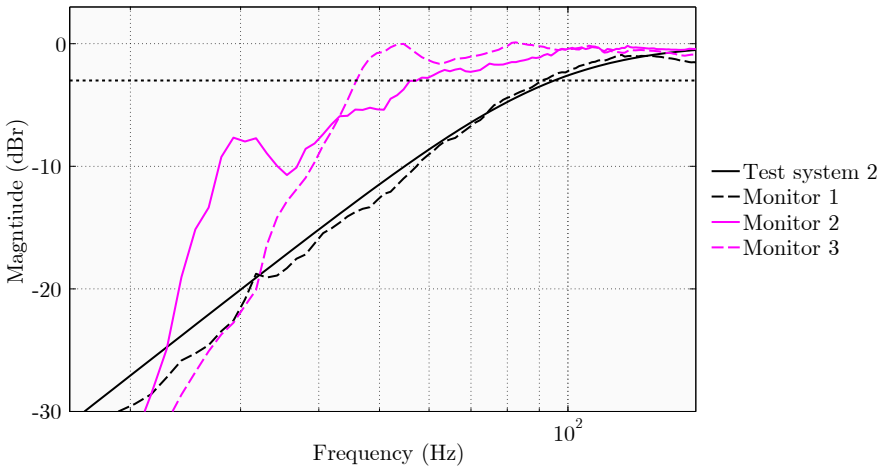
(c) Monitor 2

(d) Monitor 3

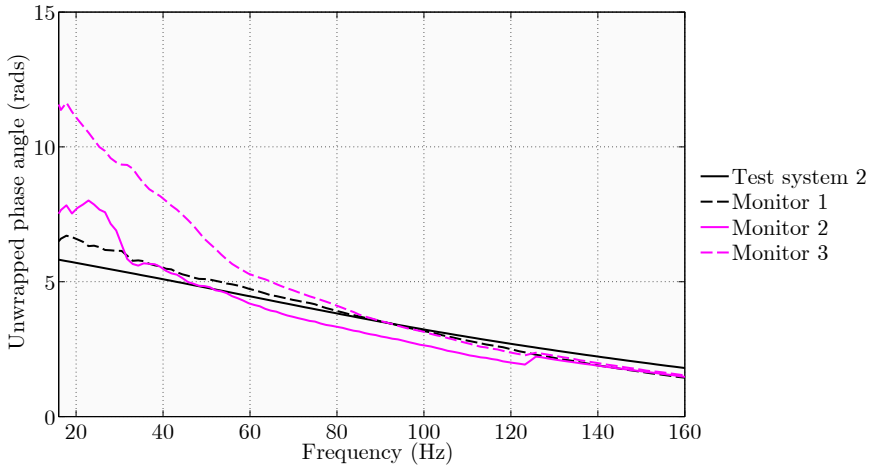
Figure 23: Waterfall plots for validation systems

Inspection of the waterfall plots was interesting as it was seen that Monitors 3 and 4 have slowly-decaying low-frequency resonances in the time domain. Initial comparison with the MTF results showed that these systems produced a kind of 'rippled' behaviour in the intensity images, indicating that they have a different type of alignment to test system 2 and Monitor 1; however, it appeared that the mean scores could not be correct: they were approximately equal for all of these loudspeakers. To investigate further, the steady-state frequency responses for these validation systems were inspected. The results are shown in Figure 24.

[¶]Monitor 2: #(36) Yamaha NS10M; Monitor 2: #(1) Acoustic Energy AE2; Monitor 3: #(2) ADAM S2A.



(a) Magnitude. The dotted horizontal line marks –3 dB



(b) Unwrapped phase responses

Figure 24: Frequency response magnitude and phase for validation systems. In both plots the frequency axis has been restricted to the range covered by the MTF algorithm

It was discussed in section 1.1.3.1 that deviations from a linear phase characteristic distort the temporal envelope of a signal, because some components are delayed more than others as they pass through the system. Monitor 3 showed the greatest amount of phase distortion at low frequencies, so based only on this, it would be expected get a lower mean MTF score. However, the magnitude response shows that this loudspeaker, like Monitor 2, has greater output at low frequencies; Monitor 3 has a full octave of extra extension (f_c approx. 45 Hz, compared to 90 Hz for test system 2 and Monitor 1); this octave is crucial for bass reproduction as it contains fundamental frequencies of rhythm section instruments. Therefore, weighing up the importance of temporal fidelity against the requirement to reproduce the frequency content of low-frequency instruments, the near-equivalent mean scores returned by the algorithm no longer seemed erroneous. Instead, was concluded that the mean score was providing a useful way to quantitatively compare a group of monitors that have very different low-frequency alignments. In

addition, it appeared that the intensity image format was a very helpful source of further information in discriminating and comparing these different behaviours.

The following conclusions about the final MTF algorithm were drawn from this analysis:

- It produces $m_{qk} = \overline{M} = 1.00$ for a perfect system, showing a completely white intensity image and flat band-mean plot.
- It shows differences between low-frequency alignments in real measured monitors and simulated responses.
- It responds to changes in a loudspeaker's output level across bands and how it responds to temporal fluctuations in the input signal.
- Inspection of the intensity image alone indicates smoothness of the response; smoother responses exhibit clear vertical banding but fluctuations in the steady-state response produce a rippled or mottled effect.
- Similar overall mean scores may be produced for loudspeakers that behave differently at low frequencies. All four systems shown in Fig. 22 return approximately the same value for $\overline{M}$, but the band-mean and intensity images show clear variations in their alignment. It is not known for these loudspeakers whether the differences shown in the visual MTF results are perceptually relevant i.e.whether they sound different when reproducing music at low frequencies, but the findings show that it is important to present the MTF results in a visual as well as a numerical format.

Based on these findings, the algorithm appeared to be suitable for the intended application and was considered suitable for further experimental work.

2.12 Summary of Algorithm Development

Earlier studies had indicated that the Modulation Transfer Function might be used to evaluate the reproduction accuracy of loudspeakers at low frequencies. The method required further investigation to develop a suitable algorithm and most effective choice of parameters.

A review of communications theory identified relevant features of amplitude modulation, the basis of the MTF technique, and implications of applying the method at low frequencies were explored. With reference to this review, four methods for calculating the MTF were proposed. These could be classified according to a fundamental difference in approach: the first pair of methods used a band-limited test system; the second pair instead used a band-limited test signal. Each of these methods was applied to two tests systems, designed to allow assessment in two key aspects: i) inherent error, shown by deviation from 1.00 in the MTF results for a simulated perfect system, and ii) responsiveness to changes in low-frequency alignment, using a high-pass filter to approximate a loudspeaker's roll-off at low frequencies. Each class of method produced fundamentally different results; further investigation of these results allowed final selection of a method for MTF computation. Although this method was known to cause some out-of-band excitation in the test system, it was considered to be the most suitable approach when trying to overcome the problems of MTF application at low frequencies.

Following the selection of a suitable MTF method, three key parameters were investigated in more detail. After defining the range of low frequencies to be covered, six frequency band arrangements were developed. One of these was chosen for use in the algorithm after comparing the MTF results from each set. The selected arrangement had band widths and centre frequency spacing that appeared to be the most revealing of behaviour in a loudspeaker's alignment, without introducing excessive overlap or bias towards one part of the low-frequency spectrum. A set of suitable modulation frequencies was selected after analysis of the envelopes found in musical content at low frequencies. Individual extracts showed distinct modulation profiles, but based on results from 168 tracks taken from a wide range of styles and genres, the final envelope spectrum formed an almost continuous distribution with dominant content below approximately 10 Hz. The final parameter to be investigated was the modulating function; the algorithm output appeared to be relatively insensitive to this feature, but a function was selected that preserved temporal fluctuations in the signal envelope at the required rate and gave clear separation between modulation cycles.

After development of a suitable algorithm, output of the results was considered. A combination of numerical and graphical presentation was used; together, these allowed simple comparison of low-frequency behaviour across multiple loudspeakers in an intuitive and consistent format that was suitable for inclusion on product data sheets. Parameters for the optimised MTF algorithm as developed throughout this chapter are summarised below:

Method: Bands of noise are amplitude-modulated before being passed through the full-bandwidth test system. Modulation index, m, is calculated from comparison of output to input modulation depth. Fluctuations in the results are reduced to an acceptable limit by using averaged envelopes, formed from 100 iterations of modulated noise, each 2.5 s in duration.

Frequency bands: Calculation is performed in ten contiguous linear bands, covering the range 16.1 to 164.2 Hz.

Modulation frequencies: Within each band, m is calculated for an input signal with increasing rates of temporal fluctuation; a set of seven modulation frequencies was selected from the averaged low-frequency musical envelope spectrum:
$f_m = [0.8,\ 1.1,\ 2.2,\ 4.3,\ 5.8,\ 8.5,\ 11.7]$ Hz.

Modulating method and function: The DSB-SC form of AM is used for modulation with the function $x_\mathrm{c}(t) = 0.5\,[1 + \cos(2\pi f_\mathrm{m} t)]$.

Output: MTF matrix mean scores, $\overline{M}$, are presented along with a plot of band-mean scores showing $\overline{m}$, the average across all modulation frequencies within a given band, as a function of band centre frequency. Detail in the MTF matrix is represented as an intensity image, showing resolution for m to 0.05.

Preliminary validation of this algorithm was performed using a simple model of a loudspeaker's low-frequency alignment and three measured responses of real professional mix monitors. Results indicated that the method was effective in revealing changes in output level and ability to accurately reproduce temporal fluctuations in the input signal. It was observed that the unweighted mean score could be similar for systems that showed distinctly different behaviour in

the visual results. Therefore, both aspects of the algorithm output were considered to be useful. The $\overline{M}$ score summarised overall MTF matrix behaviour and allowed simple direct comparison across different loudspeakers; the plots allowed more detailed inspection of a loudspeaker's alignment throughout the low-frequency range of analysis. From this validation it was concluded that the final algorithm was suitable for use in further experimental work. Development of loudspeaker models for further assessment of the method is discussed in chapter 3.

3 Creating Virtual Loudspeakers

Chapter 2 described how the MTF algorithm was developed and optimised to make it suitable for the intended application. This chapter shows the stages involved in generating appropriate models that could be assessed using the algorithm. The chosen loudspeaker modelling strategy is discussed in section 3.1; section 3.2 describes how the different loudspeaker models were designed, with section 3.3 explaining how they were realised, both for inspection in the digital domain and for acoustic playback of signals.

3.1 Choosing a Virtual Approach

The main experimental work in this project was carried out with what will be termed 'virtual loudspeakers'. These were created by modifying the response of a single real loudspeaker using digital signal processing (DSP). The low-frequency behaviour of the real loudspeaker was altered to simulate a range of different bass alignments whilst keeping the mid- and high- frequency regions the same. As a key motivation of the project was to evaluate the MTF algorithm both objectively and subjectively, it was necessary to reproduce a range of responses that measured and sounded like a group of similar-sized mix monitors differing only in their low-frequency performance. A detailed justification of the listening test strategy is given in chapter 5, but certain aspects are especially pertinent to the virtual loudspeaker approach; these are described in the following subsection, followed by a brief review of using DSP to modify a loudspeaker's performance.

3.1.1 Advantages of Simulation

Headphone reproduction may be used when investigating low-frequency effects in loudspeakers [104–107]; this is highly advantageous because subjective experiments can be performed relatively quickly and easily, requiring only a quiet room for reproduction. However, it was decided at the start of the project that headphones would not be used for simulation. The intended practical application for the MTF algorithm in this project was clearly defined, so there was a desire for the reproduction conditions to be as realistic as possible; although experimental conditions would not exactly resemble those experienced by an engineer when creating a mix, presentation using a loudspeaker was thought to be a minimum requirement. This would increase confidence in the conclusions about usefulness of the MTF algorithm for its intended application; experiments using headphones would be one step removed from reality, not least due to absence of the 'vibro-tactile sensation', the physical impact on the body of bass reproduction at high levels which is in itself a valid part of the subjective impression [9, 62, 96, 105]. It was believed that this might have some influence on the listener's evaluations, but no attempt was made to quantify this effect prior to testing.

The decision to use loudspeaker reproduction did bring some difficulties. In order to achieve accurate simulations, it was necessary to conduct measurements and listening tests inside a large anechoic chamber. The space used for experimentation was not anechoic or free from modes down to the lowest frequencies of interest in this study (details in sections 3.3.1.1 and 3.3.1.4). Secondly, accurate simulation could only be achieved at one specific point in space; the transfer function between source and receiver would be unique, and any subsequent equalisation would

only be strictly correct for that exact arrangement. There was no attempt to control listener head movements during experimentation; although this might be seen as a source of experimental error, if findings appeared to be robust to this effect, it would further support applicability of the method to practical mixing situations. It should be noted that such effects may have been considered critical if the study was focussed on high, rather than low, frequencies [107].

Despite the difficulties introduced by loudspeaker reproduction, it was deemed worthwhile to pursue this approach. The necessity for anechoic conditions was not considered to be especially detrimental with respect to realism of the experimental presentations; as described in chapter 1, mixing engineers prefer to work in conditions where they can evaluate the reproduced sound without interfering effects of the listening environment. It is, of course, acknowledged that they do not work in anechoic chambers, but the added influence of room effects to subjective impression at low frequencies was beyond the scope of this project. Contribution of this factor was therefore considered to be an extraneous experimental variable that should be controlled as carefully as possible.

As the decision to use loudspeaker reproduction has been justified, the question of why models were simulated is now addressed. Even if there had been a large stock of studio monitors to hand for testing, a virtual approach would still have been preferred because it offers significant advantages in the execution of listening tests. These aspects of experimental design have all been investigated within an audio context, and shown to be significant influencing factors when trying to obtain information about listeners' perception of the sound of a loudspeaker:

Placement Even small placement differences can cause significant changes in perception of a loudspeaker [108, 109]. Though this variable is most important when performing tests in a room, where reflections, reverberation, and modes will be strongly present, placing multiple loudspeakers next to each other and using some device to switch between them during listening is not a desirable strategy.

Changeover time To avoid placement errors, one location could be used, and multiple units swapped over during the experiment. This method might be acceptable for tests with extended listening periods, such as in monadic evaluation where a loudspeaker is evaluated in isolation, without direct comparison to another model. When fast, direct-comparison judgements are to be made, the switching time between different sources must be almost instantaneous because acoustic memory is known to be very short [107]; this is especially critical when the audible differences are very subtle. There have been studies which used advanced motorised mechanical systems to swap loudspeakers over to the same position accurately and quickly, but such systems are inaccessible to those outside of prestigious multinational audio companies with teams of expert listeners who are required to perform listening tests on an almost daily basis [110].

Visual bias Visual bias in listening tests has been formally studied [111, 112]. This is where listening test ratings for preference or quality are affected by the visual appearance of a piece of audio equipment, perhaps due to its size, apparent construction quality, or sometimes because it is recognised to be from a known manufacturer about which the participant holds an opinion. This is sometimes overcome by using an acoustically-transparent curtain to hide the test equipment from listeners. This requirement is avoided in the virtual loudspeaker approach; there was only ever one

loudspeaker used during testing, and participants were not aware that they were, in effect, comparing different loudspeakers.

Variable control From an experimental point of view, having control over the most significant experimental variable is the primary advantage of using virtual rather than real loudspeakers in this investigation. With unlimited time and resources, a setup could have been created which would switch and disguise the loudspeakers with sufficient speed and accuracy, but they would all still have different mid- and high-frequency characteristics. By using virtual loudspeakers, the fundamental experimental variable, i.e. response at low frequencies, was modified whilst leaving the rest of the frequency spectrum unaffected. The experiments were therefore equivalent to listening to groups of loudspeakers that differed only in their low-frequency behaviour; it was therefore concluded with some confidence that perceived differences between the models were necessarily due to differences in their bass performance.

In addition to controlling these sources of bias, using virtual loudspeakers also removed a number of other factors that may have influenced subjective impression. Real monitors, even those of a similar size, have an array of factors that influence the response measured at the listening position; these include cabinet panel resonances, edge diffraction, driver alignment, cone geometry, chassis structure, and the way it is fixed into the cabinet [43, 113]. Simulation of these factors was not accounted for in the virtual loudspeaker models. Whilst failure to simulate all of these aspects detracts from the realism of the presentation, they are also independent of the alignment differences that are observed in loudspeakers due to their fundamental design strategy, i.e. sealed, ported, with or without protection filtering. It was only considered worthwhile pursuing the effects of secondary design factors if the MTF algorithm was shown to be useful in resolving the differences due to these primary features. Therefore, it was considered unnecessary and undesirable to try and create very complicated loudspeaker models that would add extraneous variables to the objective and subjective evaluations.

3.1.2 Response Equalisation Using DSP

In the simplest terms, inverse filtering, or deconvolution, can be thought of as applying a filter whose response is the exact opposite to that of the system in question, e.g. a loudspeaker; the corresponding peaks and dips negate each other and a flat response is produced. The concept has been understood for many decades, but implementing it with analogue filters was somewhat difficult [16]. Early work with digital computing produced great advances in the field, but limited memory and processing speed meant that good results were still difficult to achieve as relatively short filters had to be used. Even low-specification modern computers have sufficient processing capacity to enable DSP at the level required for this project without any major difficulty.

Designing an inverse filter can be complicated if the processing has to be performed in real-time, or if multiple channels need to be equalised to perform crosstalk-cancellation [114, 115]. It is also difficult if needing to compensate for reverberation due to conducting listening experiments with a loudspeaker in a room [116]. These were not required for this project. All signal processing of experimental stimuli was performed prior to testing, so there was no need for any real-time filtering; there was therefore no need to compromise on accuracy of the results due to the necessity for reduced filter lengths. In addition, the use of a single loudspeaker for

reproduction reduced the inversion to a single-channel problem; the decision to reproduce music in mono rather than stereo was primarily chosen for other reasons, as explained in chapter 5, but it had the considerable advantage of simplifying the necessary signal processing.

The theory behind digital filtering can be found in many standard textbooks, but practical advice regarding deconvolution for loudspeakers is less commonly provided. However, Kirkeby *et al.* [117] provide an especially clear and concise summary of the single-channel deconvolution problem, including some advice on practical implementation. The method is defined as:

$$G(\omega) = H(\omega)F(\omega) \tag{3.1}$$

where: $H(\omega)$ is the system (loudspeaker) frequency response function, and $F(\omega)$ is the response of the equalising filter; $G(\omega)$ is the transfer function of the equalised system. For a flat frequency response, $G(\omega) = 1$ and thus, the equalising filter is the direct inverse of the original system function:

$$F(\omega) = \frac{1}{H(\omega)} \tag{3.2}$$

If $H(\omega)$ and $F(\omega)$ are complex in frequency, $G(\omega)$ will be perfectly equalised in both magnitude and phase at all frequencies specified by ω.

In practice, imposing a flat frequency response on a loudspeaker across all frequencies is not realisable due to the inherent attenuation at the extremes of the audio spectrum. Attempting to exactly equalise deep notches in a response is especially problematic; the reproducing system is forced to provide high gain at these frequencies. This brings the risk of amplifier saturation and therefore gross distortion. There is additional risk of distortion, and possibly damage, if trying to equalise a loudspeaker to reproduce frequencies at high levels far beyond its capabilities. Drive units are typically considered linear only with a specified displacement range; forcing them to move beyond this region produces non-linear distortion. This distortion is unpleasant and strongly audible if the motor suspension reaches its maximum physical excursion limit; it will also be present, to a lesser extent, if excessive displacement of the voice coil outside the magnet assembly produces a non-linear driving force on the diaphragm [118, 119]. Therefore, if attempting to equalise a loudspeaker in his way, it must be large, designed to reproduce low frequencies at high SPLs, and driven by a high-power amplifier.

As a final matter regarding deconvolution in audio applications, it must be acknowledged that apparent correction of a response may lead to a degradation in perceived quality; this is a particular problem if the system to be equalised is non-minimum phase, such as room reverberation and loudspeaker crossovers [25, 26, 107, 120, 121]. These were not believed to be a problem in this study, as the experiments were conducted in anechoic conditions and the experimental loudspeaker was only equalised below the frequency of its first crossover (discussed further in section 3.3); therefore only the output from the low-frequency drive unit was being equalised, and this is assumed to be minimum phase, as described in section 1.1.3.1. However, all stimuli were auditioned under experimental conditions before formal assessment to ensure that there were no obvious audible artefacts which might affect the subjective judgements.

3.2 Woofer Modelling

As virtual loudspeakers were the basis of all experimental work, it was imperative that realistic models could be developed and simulated. Section 3.2.1 describes the modelling strategy and why this approach was used, and 3.2.2 shows how it was validated.

3.2.1 Lumped-Parameter Loudspeaker Model

The low-frequency response of loudspeakers may be modelled as high-pass filters. Some studies have exploited this to investigate the audibility of different alignments in a 'parameterised' way, carefully controlling aspects of a simulated response such as roll-off slope and passband ripple [105, 113]. For the present study, it was decided that a less artificial strategy was appropriate, as if the overall low-frequency response of a real monitor was being combined with the mid and high frequencies of another. The loudspeaker model used in this project was based on a set of equations that compute the complex on-axis frequency response function for a loudspeaker, derived from a small set of electroacoustic input parameters. Influential work by Small in the 1970s showed how these easily measured electrical and acoustical parameters could be used to accurately model the low-frequency behaviour of loudspeakers. It was demonstrated that knowing certain parameters of a given drive unit would allow accurate prediction of the overall system response when it was used in different types of cabinet [30]. The discussion in section 1.1 might give the impression that 'sealed good, reflex bad'. This is not the case, as even sealed-cabinet systems can be 'misaligned'; however, the extra complexity of ported systems makes it harder to get the balance of parameters right [122]. Small's work therefore removed the need for the wasteful trial-and-error approach to loudspeaker design that was unavoidable at the time [123]. The most relevant papers in the series for this project are the ones that characterise the behaviour of sealed [18] and ported [19] loudspeakers. Each of these papers gives an historical background to the subsequent models, and characterises the behaviour for each type of design; this includes showing the impact on transient behaviour for a range of different alignments, i.e. the specific combination of individual system parameters that together determine a loudspeaker's overall response shape at low frequencies.

The lumped parameter model used in this project is not sufficient for modelling a full-range (multiple-driver) loudspeaker response; however, it is valid for modelling single-unit responses at low frequencies if factors such as location of the port relative to the drive unit, and structural cabinet resonances are ignored [6, 124, 125]. It was therefore well suited to this study, where the behaviour controlled by the low-frequency drive unit and general cabinet design strategy was of primary interest. A full derivation of the pressure response using this method is presented in Appendix C.

An algorithm was developed which kept certain parameters fixed within the model, allowing other key variables to be adjusted according to the design. In a physical sense, this was equivalent to being constrained by cabinet volume, but being able to change the drive unit, within reason - elements such as the spider and surround were not directly adjustable, but were accounted for to some extent by the ability to adjust overall mechanical Q factor. There was also the option to add a port, making it a bass-reflex cabinet, as well as a 2^{nd}-order protection filter if desired. Cabinet volume was not a modifiable input parameter because in a practical sense, keeping a constant cabinet volume would mean comparing models that would likely be used in a

similar way in real monitoring applications. The input arguments that were modified to create the experimental models are as follows:

$\mathbf{A_d}$ (m^3) Diaphragm surface area.

$\mathbf{V_{as}}$ (m^3) Volume of air equivalent to compliance of the drive unit.

$\mathbf{f_d}$ (Hz) Drive unit (free air) resonance frequency.

$\mathbf{Q_{ms}}$ (–) Mechanical quality factor.

Bl (Tm) Product of magnetic flux density in driver air gap and length of voice coil wire inside the field.

$\mathbf{R_e}$ (Ω) Voice coil electrical resistance.

$\mathbf{f_p}$ (Hz) Port resonance frequency.

The algorithm was eventually developed into a graphical user interface (GUI) to enable fast modification and visualisation of different models. Figure 25 shows an example screenshot from the user interface. As well as being able to quickly view the impulse and complex frequency response of a model, other aspects such as simulated driver displacement could be inspected, and a port or protection filter could instantly be added or removed using a checkbox. The model results could be exported and saved for future use and further processing. The added feature of being able to open the GUI with default parameters already loaded meant that previously saved models could immediately be recalled and plotted. This interface made it quick and easy to compare different designs and develop a set of experimental models, especially when adjusting several parameters at once.

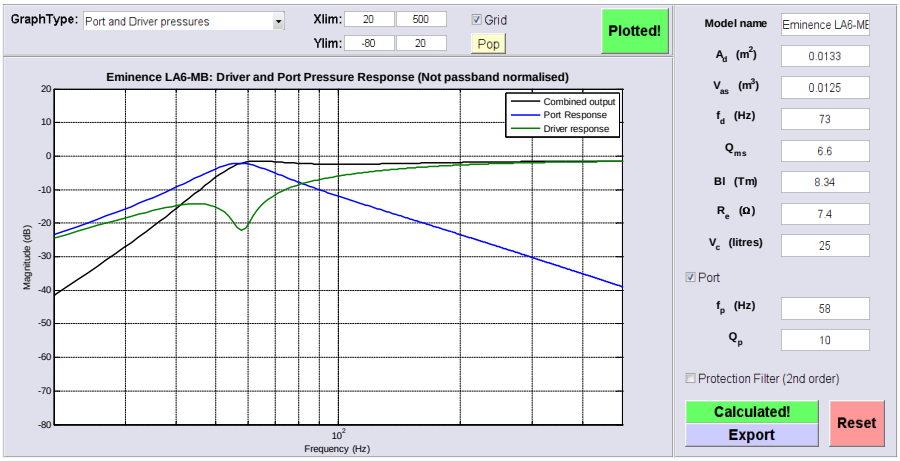
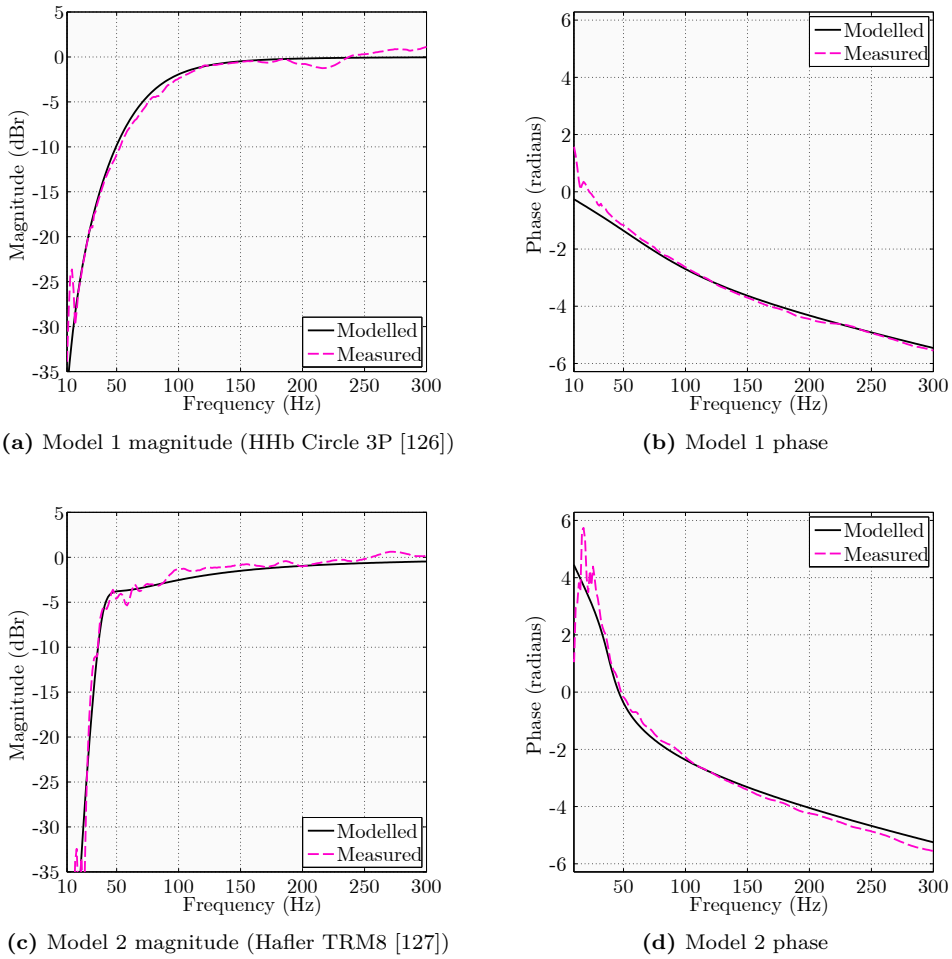


Figure 25: Example screenshot from the *WooferMaker* GUI

3.2.2 Model Validation

Outputs from the model were initially validated against high-pass filters to check that sealed and ported designs would give the correct low-frequency roll-off. However, this was only useful when

the model was forced into approximating specific filter transfer functions, and was considered to be an insufficiently realistic approach. To further validate the model, the output was compared against measured data. Figure 26 shows two examples of models created using the *WooferMaker* GUI, along with the responses of two real studio monitors, measured by Dr Keith Holland at the ISVR for published reviews in *Studio Sound* magazine[‖]. Results are shown in the frequency domain as the models simulate behaviour at low frequencies only; it is therefore easier to directly compare their responses with those of real multi-way loudspeakers by inspecting the magnitude and phase, rather than trying to assess similarity from comparison of simulated and real impulse responses which look different due to the absence of mid- and high-frequency behaviour in the models. Identical linear x-axis scales have been used in the plots to allow closer inspection of the low frequencies. The upper limit of the plots is 300 Hz, slightly less than an octave above the upper limit of the MTF algorithm, as defined in section 2.5.



(a) Model 1 magnitude (HHb Circle 3P [126])

(b) Model 1 phase

(c) Model 2 magnitude (Hafler TRM8 [127])

(d) Model 2 phase

Figure 26: Comparison of two woofer models against measurements of similar studio monitors

The models shown in Fig. 26 were adjusted to normalise the phase shift due to time-of-flight,

[‖]Back issues may be found at *http://www.americanradiohistory.com/Studio-Sound.htm*

as if the simulated loudspeaker had been measured at the same distance from the microphone as the real one. As phase shift due to acoustic transmission over a fixed path in air is a linear function of frequency, this is a pure delay [39]. This was applied to the complex frequency response by first identifying how many samples were needed to align the impulse response peaks in the time domain, then applying a corresponding phase shift in the frequency domain:

$$H_{\mathrm{ps}}(f_k) = H(f_k) \cdot e^{-j2\pi f(k)\left(\frac{N_s}{f_s}\right)} \tag{3.3}$$

where: $H(k)$ and $H_{\mathrm{ps}}(f)$ are the original and phase-shifted responses, $f(k)$ is frequency, f_{s} is the sampling frequency, and N_{s} is a constant, the equivalent sample delay in the time domain.

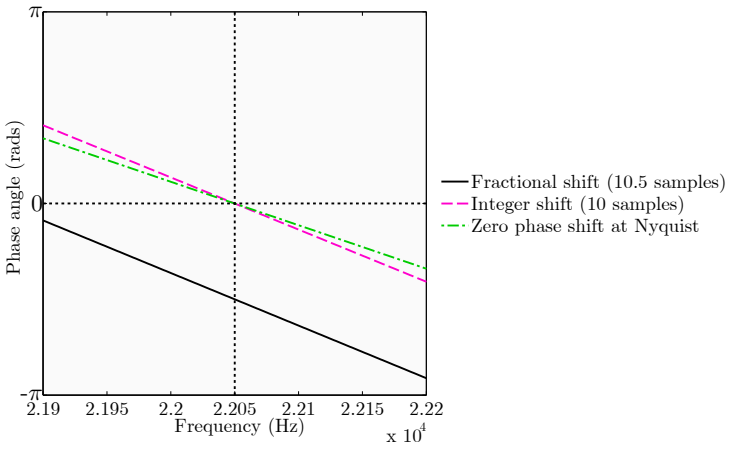
The k subscript in Eqn. 3.3 denotes that these are discrete functions of frequency; this convention can be assumed throughout further descriptions relating to experimental work. The resolution used for processing throughout the project is described further in section 3.3.1. Note that N_s was typically, but not necessarily, an integer; selecting a non-integer value is equivalent to shifting the impulse response by an increment of time less than the sampling period:

$$T = \frac{1}{f_s} \tag{3.4}$$

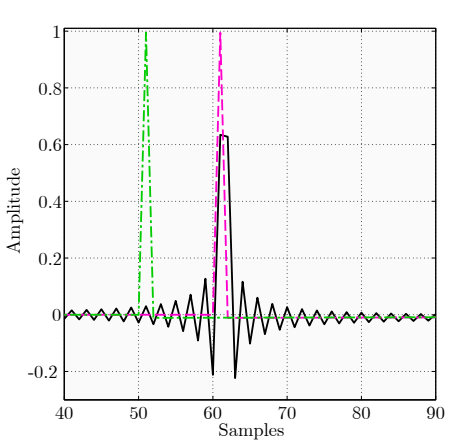
$$\Delta \begin{cases} = T & N_s = 1 \\ = kT & N_s = k \\ < T & N_s < 1 \end{cases} \tag{3.5}$$

where: T is the sampling period, and Δ is the equivalent time delay, both in seconds; k is an integer.

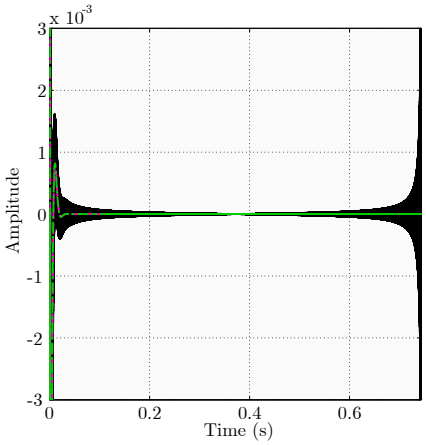
Fractional-sample phase shifts were not routinely applied, as a discrete function is not strictly defined between sampling periods. However, it was applied to the target loudspeaker responses to ensure that they had zero phase shift at the Nyquist frequency; functions not meeting this requirement show ringing in the impulse response when transformed to the time domain, as illustrated in Figure 27. The green dash-dot lines show one of the loudspeaker models that has been adjusted to have zero phase shift at the Nyquist frequency. In Fig. 27a it can be seen that the only effect on phase response of applying an integer-sample delay is an increase in the gradient (pink dashed line); in the time domain (Fig. 27b), the shape of the impulse response is identical to the original. Applying a non-integer sample delay (black solid line) produces a phase shift that does not pass through zero at the Nyquist frequency; in this example where the fractional delay is $\frac{1}{2}$ a sample, the corresponding phase shift at Nyquist is $-\frac{\pi}{2}$. In the time domain, the impulse response appears to have been delayed by the correct number of samples but is no longer identical to the original version, and shows sample-to-sample ringing; inspecting the full duration of the impulse responses (Fig. 27c) shows that this system is no longer causal.



(a) Phase response (wrapped); vertical dotted line marks the Nyquist frequency



(b) Impulse responses



(c) Impulse responses (full duration)

Figure 27: Effects of integer- and fractional-sample delays in frequency and time domain. The legend for all plots is the same

An additional convention is defined here, that of plotting the unwrapped phase. It is typical to show the principal phase, where the values vary through a maximum range of $\pm\pi$; in systems with large phase shifts, this portrayal makes plots difficult to interpret. When the gradient of the phase slope is of interest, such as when computing group delay, unwrapping is a necessary step as the function is non-differentiable at points where discontinuities occur [32]. By unwrapping the phase, multiples of 2π are added or subtracted; the original frequency response function is unchanged but the discontinuities are removed. This convention was consistently used throughout the project as it was found to be a very useful representation for inspecting the non-linear phase shifts in a loudspeaker's low-frequency response, especially when comparing across different models.

Note that the aim of comparing measured with simulated loudspeakers was not to exactly model the real units; this is not possible unless a loudspeaker manufacturer discloses their design parameters and states which drive units they use. However, most input arguments to the

loudspeaker model are typically provided on drive unit data sheets; part of the motivation for choosing this particular set of parameters was that they are generally easy to obtain, and therefore, an attempt could be made at simulating real components. The comparison illustrated in Fig. 26 was carried out to check whether, using freely available information on real drivers and appropriate corresponding cabinet sizes, the loudspeaker model was capable of creating responses similar to those that might be encountered in professional monitoring. The results demonstrated that this was indeed possible, with good agreement between the simulated and real loudspeakers in both magnitude and phase down to approximately 30 Hz; below this point, the measurement appears to have been affected by noise.

3.3 Virtual Reproduction

As described in section 3.1, the experimental models evaluated in this project were viewed as virtual loudspeakers: simulations of realistic low-frequency alignments that might be observed in real medium-sized mixing monitors. To perform this evaluation, two slightly different types of simulation were required; one was needed for MTF analysis, where all processing could be conducted in the digital domain, and another for listening to music, where analogue reproduction and acoustic transmission of a signal was required. Both types of virtual loudspeaker simulation required an initial accurate impulse response measurement of a single high-quality loudspeaker that would be used for all acoustic reproduction; this process is described in section 3.3.1. The signal processing procedures involved in creating both types of virtual loudspeaker simulation are then detailed in sections 3.3.3 and 3.3.4.

3.3.1 Measurement of the Experimental Loudspeaker

The physical loudspeaker selected for experimentation is from the Linear Spatial Reference (LSR) range made by JBL. The LSR32 is a high-quality passive studio reference monitor with a 50-litre cabinet and 12-inch low-frequency drive unit, capable of producing low frequencies at high sound pressure levels [128]. The manufacturer states that this loudspeaker has a –3 dB point at 54 Hz, and appears to have a 3$^{\text{rd}}$-order roll-off (18 dB/octave) rather than the conventional 4$^{\text{th}}$-order for most bass-reflex designs; this indicates that the cabinet is well damped internally, a technique which can be used in large cabinet loudspeakers to take advantage of the extended response obtained through porting, whilst minimising the time response effects that accompany a steep roll-off. The high sensitivity, low distortion, and high power handling capability made this an ideal loudspeaker for experimentation in this study [129].

3.3.1.1 Location Measurements were taken inside the large anechoic chamber at the ISVR, University of Southampton. Although this is a high-quality test chamber with an internal volume of 611 cubic metres, it is only nominally anechoic down to 80 Hz [130]. Preliminary assessment of the experimental loudspeaker in a listening room suggested that this environment would be unsuitable for testing. The response of the room was not measured, but it was evident from listening that the reverberation added by the room very obviously altered the impression of musical bass reproduction compared to listening in the anechoic chamber. The apparent increase in low-frequency level was so marked that it was a concern that any listener judgements of loudspeaker models in this space would be influenced by the properties of the room; it would

therefore be difficult to draw conclusions from the results and state with confidence that these judgements were due to differences between the loudspeaker models alone. Measurement in a smaller, semi-anechoic chamber confirmed that this was also an unsuitable space for assessment of low frequency reproduction; the experimental loudspeaker response measured in this chamber was so poorly controlled at low frequencies that it was not considered a good basis for stable equalisation. A comparison of measurements in the large and small chambers is presented in Appendix D. The large chamber was therefore concluded to be the best available location for measurement and listening as it provided maximum protection from external noise and minimum influence from room effects. Figure 28 illustrates the equipment layout during measurement. This setup was identical for the listening experiments except that the listener, rather than the microphone, was situated at the chair location. The microphone height during measurement was set to be level with the loudspeaker's tweeter; during listening tests, the chair was adjusted up or down as necessary to compensate for differing height of listeners such that their ears were also at tweeter level. This adjustment was approximate and not formally measured for each participant during testing. Measurement location was chosen with a number of factors in mind; the centre of the chamber was avoided as it was believed that the modes would be strongest here due to the regular dimensions of the space. An off-centre position was chosen, keeping as great a distance as possible from the wedges, experimental equipment, and edges of the grid floor, a safety precaution as this location was also used for listening.

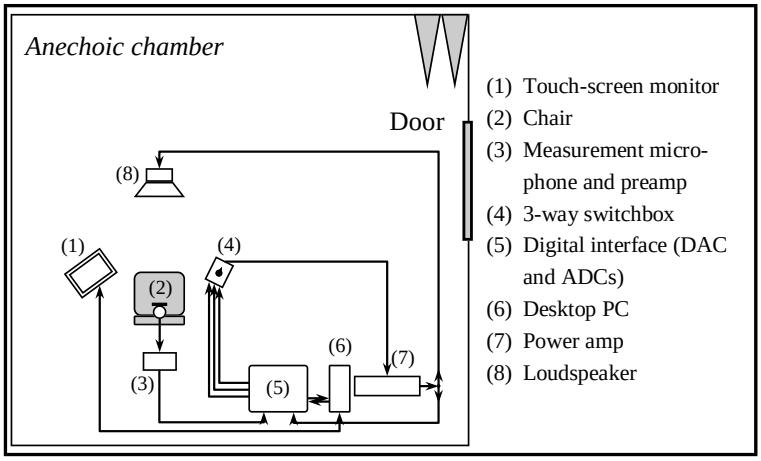


Figure 28: Plan view of the equipment layout for anechoic impulse response measurement of the physical experimental loudspeaker. Not to scale.

The front port on the loudspeaker was sealed for all measurement and reproduction using a chamfered wooden bung, fixed with adhesive tak to ensure a completely airtight seal; this was covered with black tape to make appearance of the modification less obvious compared to the rest of the cabinet. As a bass-reflex loudspeaker is a two degree of freedom (2DOF) resonant system, it was believed that this would be more difficult to successfully equalise than a simpler 1DOF system that would result if the cabinet was sealed. The steep roll-off of the 2DOF system below cut-off due to the port and driver being in antiphase meant that even larger levels of gain were needed to equalise the response in this region. Sealing the port also removed the possibility

of any unwanted audible effects of the port, such as 'chuffing'; this is due to turbulent flow as air moves in and out of the port, and is more likely at lower frequencies and higher SPLs [119, 124]. It should be noted that blocking the port has an effect on the alignment; the half-power point occurs at a higher frequency and the gradient of the roll-off reduces, but a loudspeaker optimised to operate as a bass-reflex model will not necessarily result in an optimised sealed-cabinet alignment simply by blocking the port. However, as demonstrated in section 3.3.1.4, blocking the port in this case did result in an acceptable response in that it was neither under- nor over-damped. The absence of resonances due to sealing the port was important given the nature of the study and the reliance on equalisation to reproduce the desired model responses.
Figure 29a shows the experimental loudspeaker used for measurement and playback after sealing the port; its position relative to the point of measurement (and listening) inside the anechoic chamber is shown in Figure 29b. Both images show the two loudspeaker stands and plywood base used to stabilise this large, heavy unit on the metal grids; rubber mounting feet were placed between the stands and loudspeaker cabinet to reduce any structure-borne vibrations.

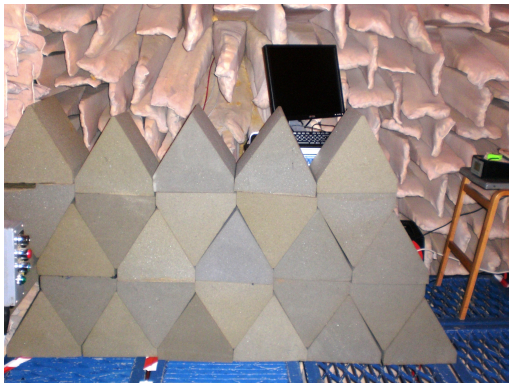
(a) Loudspeaker

(b) Measurement position

Figure 29: Experimental loudspeaker (29a), and location relative to the measurement microphone (29b).

3.3.1.2 Recording System A calibrated Bruel & Kjaer Falcon ½" pre-polarised condenser microphone was used for all measurements. The RME ADI-8 DS and ADI-648 were used for all recording and playback, both in measurement and in listening; these are professional-grade digital-to-analogue (DAC) and analogue-to-digital (ADC) converters and interface that allow multichannel audio conversion between the power amplifier and computer. The features include handling of up to 24-bit digital audio and 96 kHz sampling rates, high SNR and low jitter, ensuring accurate and low-noise acquisition of the acoustic signal. Figure 30 shows the digital recording system; the desktop computer, power amplifier and ADC/DAC rack can be seen in Figure 30a. The acoustic wedges shown in Figure 30b were used during all procedures to attenuate fan noise from the PC and made it inaudible from the measurement/listening position, recorded as being less than 30 dB(A) SPL. The power amplifier used to drive the loudspeaker was capable of delivering 250 W into an 8 Ω load in bridged mode; this mode was used for all experimental work. A full list of experimental equipment is provided in Appendix E.

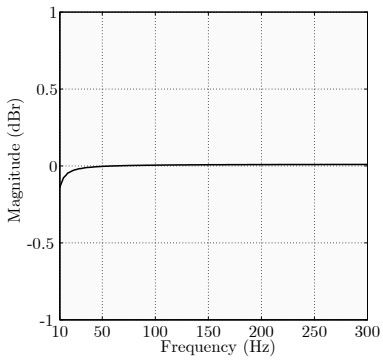
(a) DACs, PC and power amp



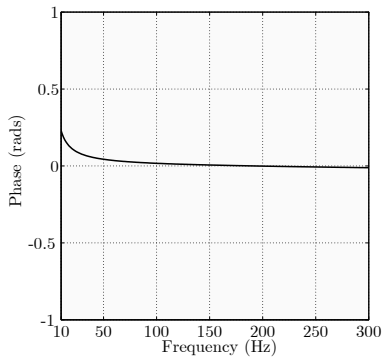
(b) Wedges placed to attenuate fan noise

Figure 30: Recording equipment

The transfer function of the measurement system was checked by passing a signal straight through from output to input, i.e. no acoustic transmission. This was to verify that there was no adverse behaviour at low frequencies that might affect the loudspeaker measurements or model simulations during playback. The results are shown in Figure 31 up to 300 Hz. It was concluded that the magnitude and phase response of the system at low frequencies did not compromise experimental work, because the observed deviations from a flat characteristic were only present at the lowest extreme of the frequency range of analysis, approximately 16 Hz, and any signal content this low was attenuated by the roll-off of the loudspeakers, both real and simulated. Some deviation from a 'perfect' response in this region was therefore considered acceptable. The slight increase in phase shift towards low frequencies was very gradual, with a value of 0.1 radians at 20 Hz; the corresponding magnitude deviation at this frequency was -0.04 dB. These values were small compared to deviations introduced by the loudspeaker models; at the same frequency, even the model considered to be an idealised loudspeaker response (listening test I reference) had magnitude and phase deviations of approximately -3 dB and 1.5 rads. Latency was commendable at 2.2 ms (separation between output and input channels = 98 samples).



(a) Magnitude



(b) Phase

Figure 31: Measurement system transfer function up to 300 Hz

3.3.1.3 Data Capture A dual-channel method was used to estimate the frequency response function (FRF) of the experimental loudspeaker. This method allows measurement of a system even in the presence of extraneous noise; if using a random input signal with a sufficient number of averages to reduce variance in the final spectral estimate, the FRF derived from this method represents the best linear fit to the true system response even if it exhibits non-linearities during excitation [131, 132]. The H_1 estimator was used as this is most robust to the presence of noise in the output signal, which was considered more likely in this application than noise at the input:

$$H_1(f) = \frac{G_{\mathrm{AB}}(f)}{G_{\mathrm{AA}}(f)} = \frac{A^{\star}(f)B(f)}{A^{\star}(f)A(f)} \tag{3.6}$$

where: $H_1(f)$ is the complex FRF, $G_{\mathrm{AB}}(f)$ is the cross-spectrum between $A(f)$ and $B(f)$, the spectra of loudspeaker and microphone inputs, $G_{\mathrm{AA}}(f)$ is the auto-spectrum of $A(f)$, and $\star$ denotes the complex conjugate. Note that the cross- and auto-spectra in Eqn. 3.6 were obtained from the average across all segments:

$$G_{\mathrm{AB}} = \frac{1}{k}\sum_{m=1}^{k} \tilde{G}_{\mathrm{AB}m}(f) \tag{3.7}$$

$$G_{\mathrm{AA}} = \frac{1}{k}\sum_{m=1}^{k} \tilde{G}_{\mathrm{AA}m}(f) \tag{3.8}$$

where: k is the total number of segments, $\tilde{G}_{\mathrm{AB}m}(f)$ is the cross-spectrum estimate for segment m, and $\tilde{G}_{\mathrm{AA}m}(f)$ is the auto-spectrum estimate for segment m.

The system input signal was 'tapped off' before reaching the loudspeaker; this is represented by the junction point at (7) in Figure 28; the system output was the signal received by the microphone at the listening position, shown by point (2) in the same figure. Pink noise was used as the excitation signal; this was chosen due to its constant power per octave, giving a better signal-to-noise ratio (SNR) at low frequencies compared to an equivalent white noise extract [133]. For processing, segments were extracted from the measurement recordings; these were .wav files of the loudspeaker input and microphone output recorded on separate channels. These two files were then used to compute the loudspeaker's FRF according to Eqn. 3.6**. For recording, the sampling rate was 44.1 kHz; for processing, $N = 2^{15} = 32768$-point FFTs were used. This allowed averaging over 456 segments of length N within the 170 s-duration recordings, with 50 % overlap. The resulting frequency resolution of the measurement was:

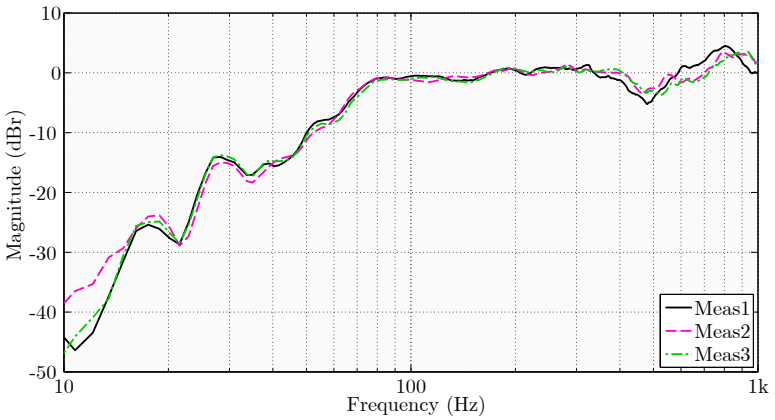
$$\Delta f = \frac{f_s}{N} = 1.346\,\mathrm{Hz}\ (3\,\mathrm{d.p.}) \tag{3.9}$$

where: f_s is the sampling rate, and N is the FFT length.
This resolution was considered sufficient to allow accurate inspection and adjustment of the frequency range of interest in this project, and was maintained for further processing.

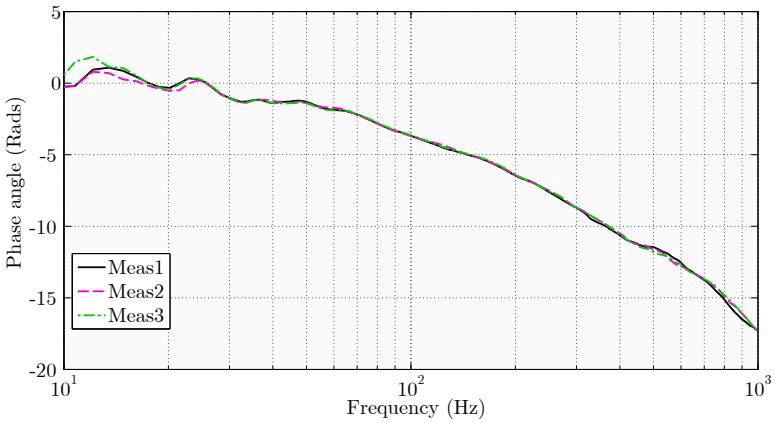
3.3.1.4 Results Figure 32 shows results up to 1k Hz from measurement of the JBL loudspeaker on three separate occasions. The response was remeasured before each set of

**MATLAB function based on an algorithm originally written by Dr. Keith Holland, ISVR

listening tests so that an inverse filter could be generated for each set of experiments when all equipment had been set up in the final listening position for that set of tests; therefore, any variations in the source-receiver transfer function due to slight differences in the acoustic transmission path between experiments could be compensated for. Note that slight absolute differences between the magnitude responses in Fig. 32a have been removed by normalising to the individual mean values in the range 80 Hz-400 Hz, making the three data sets easier to compare. These differences are due to not having exactly matched input and output gain settings in the dual-channel measurement; they do not affect the measured response characteristics, only their overall level. Also to aid comparison, the phase responses in Fig. 32b have been unwrapped and pure delay due to the time-of-flight between loudspeaker and microphone has been removed using the method described in section 3.2.2.



(a) Magnitudes; normalised to 0 dB using the mean level in the range 80 Hz-400 Hz



(b) Phase (unwrapped); pure delay of acoustic path between source and receiver has been removed

Figure 32: Comparison of three separate measurements of the experimental loudspeaker response, shown between 10 and 1k Hz

The similarity in both magnitude and phase responses from three independent measurement periods indicated that the data were reliable and stable, therefore forming a good basis for creating the inverse filters. Whilst the low frequencies appeared consistent, some differences were seen across the mid and high frequencies, for measurement 1 in particular; these were attributed

to reflections and minor changes in placement of test equipment across the measurements. These were not considered a problem, because these regions were not affected by subsequent equalisation, and placement of test equipment was fixed within a given set of experiments, with any unnecessary items removed from the chamber before testing. The peak around 30 Hz is believed to be a room mode due to the chamber not being strictly anechoic at very low frequencies. Pedersen [113] experienced the same problem at the same frequency in similar experimental conditions, but proceeded and did not report any adverse affects on the outcomes of the study.

3.3.2 Creating an Inverse Filter

Section 3.3.1 described how an accurate measurement of the experimental loudspeaker response was captured; this was fundamental to creating the virtual loudspeakers to be evaluated in listening tests. The next step was to use this measurement to create an inverse filter that modified the response only in the low frequencies. This process is summarised in the following steps; Figures 33 to 38 illustrate how the filter is developed over several stages.

1. The process begins with the experimental loudspeaker FRF, $H_{\mathrm{LL}}(f)$, as obtained from the measurement described in section 3.3.1. The very steep slope in the unwrapped phase response illustrated in Figure 33 is caused by the time-of-flight, the linear phase shift due to acoustic propagation between loudspeaker and microphone. Dotted horizontal lines in the magnitude plot mark –40 and 40 dB, and are referred to in points 4 and 6.

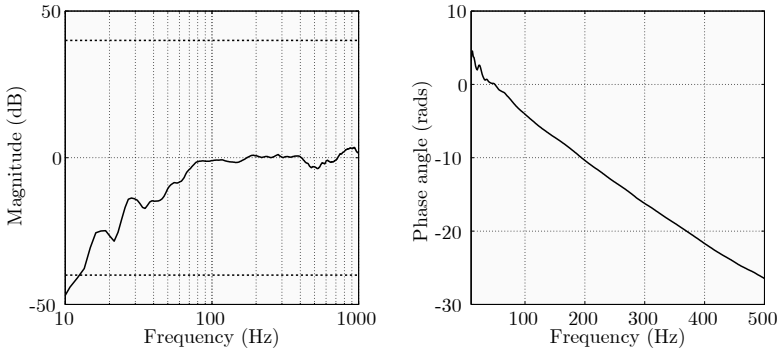


Figure 33: Inverse filter creation stage 1: Experimental loudspeaker FRF

2. The linear phase shift is removed using the method described by Eqn. 3.3. As demonstrated in Figure 34, this process leaves the magnitude unaffected, but removes the dominating phase shift due to acoustic propagation.

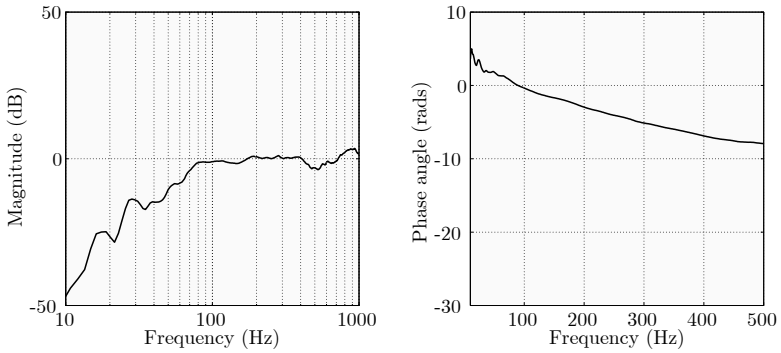


Figure 34: Inverse filter creation stage 2: Time-of-flight removed

This correction was required because the aim was to equalise part of the loudspeaker response only, not the acoustic path between source and receiver; the behaviour of the loudspeaker itself could not be properly inspected without first removing this delay [31,134]. Creating an inverse filter with the propagation delay still present would have simulated placement of the loudspeaker at the listening position in the subsequent reproduction.

3. A 'smoothing window' was created using a low-pass filter transfer function, $L_{\mathrm{sm}}(f)$

$$L_{\mathrm{sm}}(f) = \frac{1}{1 + \mathrm{j}\left(\frac{f^n}{f_c^n}\right)} \tag{3.10}$$

where: f is the frequency vector, f_c is the low-pass cut-off frequency, and n is the filter order; for all cases, $f_{\mathrm{c}} = 125\,\mathrm{Hz}$ and $n = 2$.
The values for f_c and n were chosen to give a gentle transition between the smoothed and unaffected regions of the response, i.e. the low frequencies being modified, and the natural mid- and high-frequency behaviour of the experimental loudspeaker. Given the nature of the study, it was not appropriate to choose a higher f_c with a steeper roll-off above this point.

4. The smoothing window was applied separately to the magnitude and phase of the 'de-delayed' loudspeaker response:

$$X_M(f) = 20 \log |H_{\mathrm{ps}}(f)| \tag{3.11}$$

$$X_{M\mathrm{sm}}(f) = X_M(f)\,|L_{\mathrm{sm}}(f)| \tag{3.12}$$

where: $X_M(f)$ is the logarithmic magnitude of response $H_{\mathrm{ps}}(f)$.

$$X_P(f) = \angle\,[H_{\mathrm{ps}}(f)] \tag{3.13}$$

$$X_{\mathrm{psm}}(f) = X_P(f)\,|L_{\mathrm{sm}}(f)| \tag{3.14}$$

where: $X_P(f)$ is the phase response of $H_{\mathrm{ps}}(f)$.

Note that the logarithmic magnitude is used in Eqn. 3.12 so that the response at higher frequencies appears to be smoothed rather than attenuated; multiplying the linear values by a function approaching zero produces the effect shown in Figure 35:

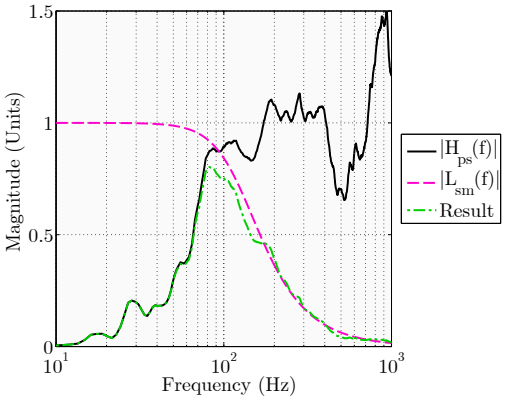
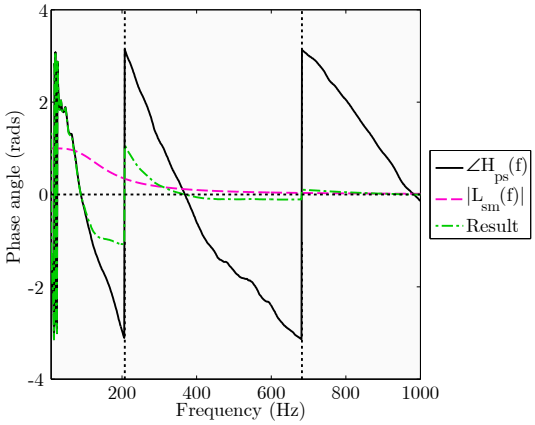
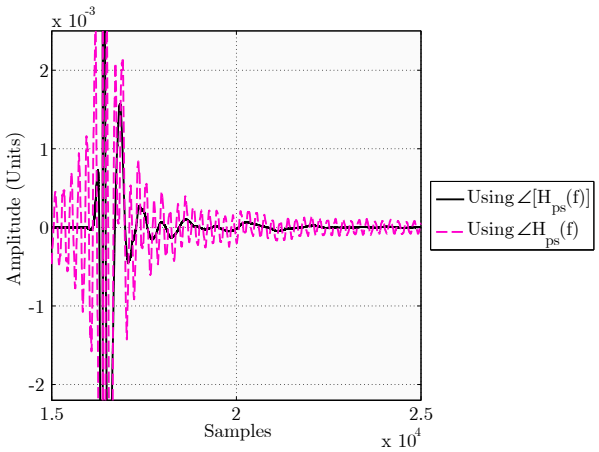


Figure 35: Effect of windowing with linear magnitude. The green dot-dash line is the product of $|H_{\mathrm{ps}}(f)|$ and $|L_{\mathrm{sm}}(f)|$

The square brackets in Eqn. 3.13 denote that the unwrapped phase was used for this calculation, adopting the notation of Oppenheim *et al.* [32]. Computation with the principal (wrapped) phase showed discontinuities in the resulting response; given the nature of the study, artefacts of this nature were avoided as they introduced ringing in the time domain. This is illustrated in Figure 36. It can be seen that discontinuities arising from smoothing the wrapped phase produce ringing in the time domain that is absent if the unwrapped version is used instead. (In this example a common delay was added to both IRs before plotting to align their peaks with $\frac{N}{2} = 16384$ samples.)



(a) Smoothing the wrapped phase leaves discontinuities in the result: green dot-dash line, the product of $\angle H_{\mathrm{ps}}(f)$ and $|L_{\mathrm{sm}}(f)|$. Their location is marked by vertical dotted lines for clarity



(b) Partial view of the impulse responses derived from smoothing the unwrapped phase (black solid line), and principal phase (magenta dashed line)

Figure 36: Effect of smoothing principal (wrapped) phase in frequency and time

5. The magnitude and phase were recombined to produce a complex FRF:

$$X_{\mathrm{msm}}(f) = 10^{\frac{X_{M\mathrm{sm}}(f)}{20}} \tag{3.15}$$

$$H_{\mathrm{sm}}(f) = \mathfrak{Re} + \mathrm{j}\mathfrak{Im} = X_{\mathrm{msm}}(f)\cos(X_{\mathrm{psm}}(f)) + \mathrm{j}X_{\mathrm{msm}}(f)\sin(X_{\mathrm{psm}}(f)) \tag{3.16}$$

where: $H_{\mathrm{sm}}(f)$ is the loudspeaker response, smoothed in mid and high frequencies only. Figure 37 shows how the low frequencies from $H_{\mathrm{ps}}(f)$ are preserved, but the mid and high sections appear to have been 'smoothed' by $L_{\mathrm{sm}}(f)$; the vertical dotted lines show f_c, the cut-off frequency of the smoothing filter. It can be seen how the selected values for f_c and n left the majority of the bass region unaffected whilst giving a smooth transition into the loudspeaker's mid- and high-frequency response in both magnitude and phase. This effect is produced because the absolute value of L_{sm} is equal to 1 at low frequencies and 0 at high frequencies, with the width and start point of the transition region determined by f_c and n.

The resulting magnitude and phase responses are equal to that of the measured loudspeaker below the transition region of L_{sm}, and equal to 0 above it.

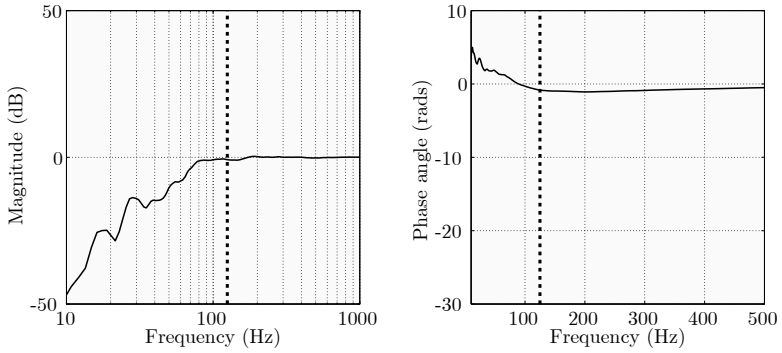


Figure 37: Inverse filter creation stage 3: Smoothed response

A separate process applied at this stage involved limitation of the magnitude response to a minimum value, preventing excessive gain when inverted:

$$H_{\mathrm{sm}}(f) = \begin{cases} 0.01\cos(X_{\mathrm{p}}(f)) + \mathrm{j}0.01\sin(X_{\mathrm{p}}(f)) & \text{if } |H_{\mathrm{sm}}(f)| < 0.01 \\ X_{\mathrm{m}}(f)\cos(X_{\mathrm{p}}(f)) + \mathrm{j}X_{\mathrm{m}}(f)\sin(X_{\mathrm{p}}(f)) & \text{if } |H_{\mathrm{sm}}(f)| \geq 0.01 \end{cases} \tag{3.17}$$

The lower dotted line in Fig. 33 shows the point below which the gain capping was applied. This value was chosen as a compromise between achieving an accurate equalisation and limiting the amount of gain required for the inversion. The effects of this process are demonstrated in step 4.

6. Finally, the smoothed response was inverted to create the bass-equalising filter, $T_{\mathrm{LF}}(f)$:

$$T_{\mathrm{LF}}(f) = \frac{1}{H_{\mathrm{sm}}(f)} \tag{3.18}$$

Figure 38 illustrates how the resulting filter $T_{\mathrm{LF}}(f)$ has a flat response (0 dB magnitude and 0 radians phase shift) in mid and high frequencies, meaning that it will not modify those parts of the experimental loudspeaker's response.

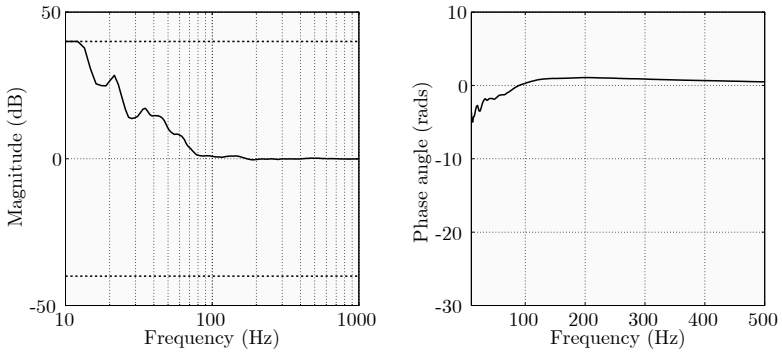


Figure 38: Inverse filter creation stage 4: Inverse filter with gain capping. Lower and upper dotted horizontal lines mark –40 and 40 dB. Gain of the inverse filter is seen to be limited to the maximum value of 40 dB for frequencies below 13 Hz.

Referring to Eqn. 3.17, the magnitude of the response being inverted was limited to a minimum value of 0.01; the maximum value of the inverse filter was therefore 40 dB. It can be seen in Fig. 33 that the magnitude of the measured system drops below the minimum allowable value, marked by the lower horizontal dotted line, at approximately 13 Hz. As a result, the magnitude of the inverse filter at these frequencies is capped at the maximum allowed value, seen as a 'flattening off' of the magnitude in Fig. 38. Therefore, the gain capping slightly reduced the accuracy of inversion, but only affected the deep notch in $H_{\mathrm{LL}}(f)$ around 10 Hz. It should also be noted that although Fig. 38 suggests that the amplifier would need to supply a large amount of gain at very low frequencies, the physical loudspeaker was never actually 'equalised flat' in practice; as described in section 3.3.4, the equalised response was always modified by the low-frequency roll-off of the loudspeaker models when reproducing signals. Therefore, gain capping was not considered detrimental to the equalisation and subsequent loudspeaker simulations during playback; the affected frequencies were at such a low register that even if musical content existed in this region, it was attenuated by the roll-off of the models. These frequencies were also below the lower limit of the algorithm analysis range.

3.3.3 Creating a Response for Analysis

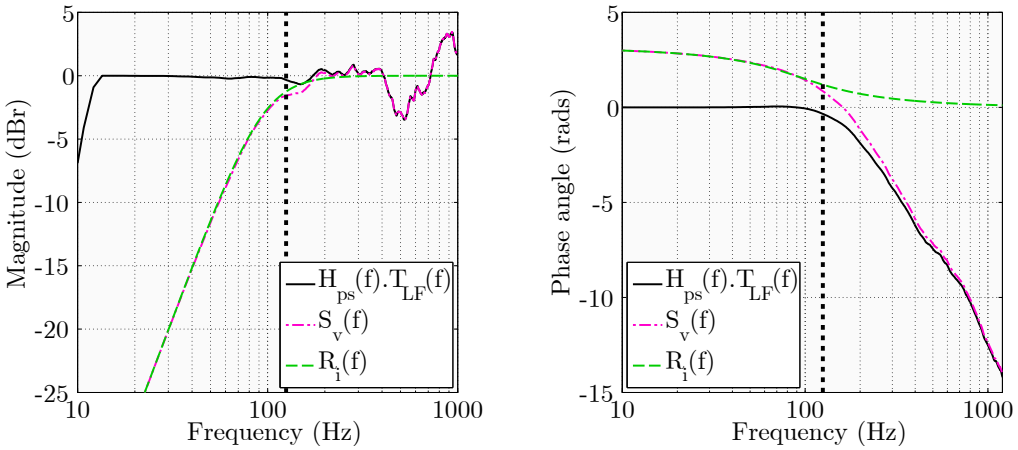
Using the experimental loudspeaker response, $H_{\mathrm{ps}}(f)$, and inverse filter, $T_{\mathrm{LF}}(f)$, the virtual loudspeakers were simulated using:

$$S_{\mathrm{v}}(f) = H_{\mathrm{ps}}(f)T_{\mathrm{LF}}(f)R_{\mathrm{i}}(f) \tag{3.19}$$

where: $S_v(f)$ is the virtual loudspeaker complex FRF, and $R_i(f)$ is the target low-frequency alignment.

The use of different target alignments, $R_i(f)$, meant that different virtual loudspeaker responses could be simulated whilst $H_{\mathrm{LL}}(f)$ and $T_{\mathrm{LF}}(f)$ remained constant. Figure 39 illustrates the process for a single target alignment; the vertical dotted lines show the cut-off frequency of the smoothing filter, f_{c} of $L_{\mathrm{sm}}(f)$ described in section 3.3.2. The plots illustrate how the virtual loudspeaker output, $S_{\mathrm{v}}(f)$, smoothly transitions between the target and bass-equalised responses

around this point; the low-frequency alignment of the target is combined with the unmodified mid and high frequencies of the loudspeaker.



(a) Magnitude **(b)** Phase

Figure 39: Virtual loudspeaker response simulation. Plots show the bass-equalised experimental loudspeaker response, the target $R_i(f)$, and the final virtual loudspeaker response, $S_v(f)$

3.3.4 Creating a Response for Listening

The simulations derived using Eqn. 3.19 were sufficient for MTF analysis as this objective evaluation did not require any acoustic playback of signals. For subjective evaluation, listeners assessed the virtual loudspeakers as they reproduced a variety of musical extracts. A different simulation process was required to achieve this reproduction. The initial steps for creating inverse filter $T_{\mathrm{LF}}(f)$ remained the same, but final implementation had to be performed in the time domain by filtering the musical extracts before playback to ensure that the final signal reaching the listening position produced the desired response. The steps are described below.

- The target alignment and inverse filter were used to create a 'partial target system' impulse response, $k_{\mathrm{i}}(t)$:

$$k_{\mathrm{i}}(t) = \mathcal{F}^{-1}\left\{R_{\mathrm{i}}(f)T_{\mathrm{LF}}(f)\right\} \tag{3.20}$$

- A modified musical extract, $\hat{m}_{\mathrm{n}}(t)$, was created though convolution with a musical extract $m_{\mathrm{n}}(t)$:

$$\hat{m}_{\mathrm{n}}(t) = k_{\mathrm{i}}(t) * m_{\mathrm{n}}(t) \tag{3.21}$$

- Playback of this signal through the experimental (real) loudspeaker is then equivalent to a convolution with the original system impulse response:

$$s_{\mathrm{p}}(t) = \hat{m}_{\mathrm{n}}(t) * h_{\mathrm{LL}}(t) \tag{3.22}$$

where: $s_p(t)$ is the musical waveform reaching the listening position. Note that:

$$h_{\mathrm{LL}}(t) = \mathcal{F}^{-1}\left\{H_{\mathrm{LL}}(f)\right\} \tag{3.23}$$

Therefore, $h_{\mathrm{LL}}(t)$ includes the linear phase shift due to acoustic propagation between loudspeaker and listener.
Using this process, the waveform reproduced at the listening position is equivalent to playback of musical extract $m_n(t)$ through a virtual loudspeaker having the response $S_v(f)$, as described by Eqn. 3.19. For each set of listening tests, i target responses and n musical extracts were processed in this way to allow reproduction of a selection of programme material through multiple virtual loudspeakers. Selection and preparation of the musical extracts is discussed in more detail in chapter 5.

3.4 Summary

Experimental reproduction was performed using a single loudspeaker inside a large anechoic chamber. Although this added some complications to the experimental procedure, it allowed a more realistic representation of the audible impression experienced in the intended MTF application, that of mix monitoring. A large, high-quality professional monitor was used throughout; its low-frequency response was modified through digital signal processing to simulate the reproduction of musical material through a range of loudspeakers, differing only in their bass alignment.

Target low-frequency alignments were generated using a lumped-parameter model. A function was developed based on this model; the input arguments to the function were chosen to be electroacoustic parameters that are typically shown on drive unit data sheets. The model therefore allowed control over primary factors in designing a loudspeaker: selection of a drive unit, and the decision whether to mount it in a sealed or ported cabinet, with the option of adding a protection filter. The function was built into a graphical user interface to allow easy adjustment of parameters and inspection of the results. Through comparison with measurements of real medium-sized professional mix monitors, it was verified that the model was sufficient to simulate the type of realistic loudspeaker responses required for this study.

The target alignments were imposed onto the experimental loudspeaker through a process of partial deconvolution of the complex frequency response; the virtual loudspeakers for analysis were formed by combining the target bass alignments with the real loudspeaker's natural mid- and high-frequency behaviour. This was achieved by creating an inverse filter that only equalised the experimental loudspeaker at low frequencies, whilst ensuring a smooth transition into the unequalised part of its response; this prevented any sudden discontinuities that would produce ringing in the time domain. This filter was based on an accurate measurement of the loudspeaker's complex frequency response; repeated measurements showed this to be stable and insensitive to minor positional changes throughout the frequency range of interest, indicating that the inversion of this system was reliable.

The virtual loudspeakers were simulated using two different procedures. For objective analysis, i.e. MTF evaluation, simulation in the digital domain was sufficient. For subjective evaluation, i.e. acoustic reproduction in listening tests, part of the processing had to be performed in the time domain; a given virtual loudspeaker could only be realised by reproducing a pre-processed musical extract through the experimental loudspeaker upon which the inverse filter was based. The virtual loudspeakers chosen for experimentation, and their assessment with the MTF algorithm, are presented in chapter 4.

4 Objective Evaluation of Loudspeaker Models

Chapter 3 described the strategy for designing and realising loudspeaker models for experimentation. This chapter presents the chosen groups of virtual loudspeakers, followed by their results from assessment with the MTF algorithm. A description of the model groups and the rationale for developing them is given in section 4.1. The experimental loudspeaker models are presented in sections 4.2 to 4.5; their algorithm results are presented after a description of the design parameters and plots showing the simulated and measured steady-state responses. Conclusions following the objective analysis are summarised in section 4.6.

4.1 Model Groups

It was relevant in this study to assess virtual loudspeakers that had subtle differences in low-frequency alignment. Assuming testing of a group of typical mid-sized professional mix monitors, they would be expected to have relatively similar limits for low-frequency extension but, as discussed in section 1.1.3, the manufacturers may have used different design strategies to achieve adequate levels of bass output. A test group of real monitors would therefore show a range of magnitude and phase behaviours; consequently, they would be expected to produce different algorithm results. The focus of experimental work was to simulate a group of systems representative of those in the target application, then assess how both the algorithm and the listeners judged differences between them.

4.1.1 Group Development

Three groups of models were assessed by listeners in this study. Group I was developed first to address the primary experimental question: the gathering of subjective data for comparison with algorithm results from the same virtual loudspeakers. Prior to testing, it was not known how difficult participants would find the listening task, or even whether the majority would be able to detect differences between all model pairs. Selection of alignments for the first listening test was performed with this in mind; it was intended that the chosen range of systems would allow all participants to respond to the experimental question in at least some of the trials. This was achieved by including some virtual loudspeakers that had very different limits for low-frequency extension; these systems clearly lacked audible bass output, and hence, it was expected that all listeners would be able to reach a decision quite easily in trials containing these models. If results did not demonstrate this to be the case, it would indicate a problem with the experimental reproduction of models or selection of participants, and further investigation would be required before basing any conclusions on the data or performing future tests.

Following the first set of listening tests which assessed Group I models, it was concluded that differences between the alignments were detectable. As a result of having gathered this data for addressing the main experimental question, it was decided that a second round of experiments was needed and was appropriate, given that the task appeared to be manageable for most participants. The Group II models were subsequently developed. As in Group I, these virtual loudspeakers had a range of different phase responses due to modelling a mixture of design strategies, but were overall more similar in low-frequency extension. The intention was that listeners were less likely to rely solely on the perceived level of bass to make their assessments of

reproduction accuracy than was the case in Group I. It was therefore expected that participants would find assessment of the Group II models more difficult, but the task was more representative of the range of differences that would typically be encountered when comparing real mix monitors.

From the experiment with Group II models it was concluded that, as expected, the comparison of systems lacking gross differences in bass level was a more difficult task for the majority of listeners. This raised the question of whether the differences due to phase response between speakers were influencing judgements at all, or whether assessments were based only on variations in magnitude response, i.e. relative levels of overall bass content in the reproduced music. As described in section 1.1.4, both aspects of a monitor's response are important for successful mix monitoring, and this provided the motivation for creating a new measure which offers more information than a magnitude response plot alone. The Group III models were developed to investigate this important aspect of interest to the study; the intention was to assess whether participants could detect audible differences in reproduced music when only a loudspeaker's phase response at low frequencies was altered. As discussed in sections 1.1.3.1 and 1.1.4.3, different monitor design strategies largely determine the phase response at low frequencies, and can therefore lead to different reproductions of the same source material; but there will also be accompanying variations in the steady-state magnitude response, and these are often considered to be the cause of perceived differences between loudspeakers reproducing otherwise identical signals. It has been suggested that phase distortion alone, especially in music, is not perceptible or may only be detected by the most critical listeners under the most sensitive conditions. The Group III systems lacked the corresponding magnitude response changes, and were therefore not representative of real monitor comparisons; however, if evaluation of these models showed that phase distortion alone was perceived by the non-expert participants in this study, it is reasonable to presume that the same effect would be even more obvious to the intended users: professionals who are highly critical and experienced at listening for subtle changes in bass reproduction accuracy, such as those which arise due to the use of different monitors for mixing. Provided that the algorithm reflected these subtle differences, there would then be evidence to justify its use in the target application. Creation of the Group III models therefore allowed a preliminary investigation of how the algorithm responded to changes in a loudspeaker's phase response alone, and how, if at all, these changes correspond to audible differences in musical bass reproduction. If differences were audible to the majority of listeners and were reflected in the algorithm results, it would support both the motivation for, and the use of, the new method.

4.1.2 Group Composition

All three groups contained a reference for comparison in listening tests. The need to include a reference for subjective assessment is described in more detail in section 5.1; it is sufficient here to state that the reference was intended to simulate a loudspeaker with more accurate low-frequency reproduction than any other model in its experimental group. This means that it had the most extended magnitude response (Groups I and II only), and least phase distortion throughout the bass region. Under the assumption that the reference was the most accurate reproducer of musical content, it was expected to show the best algorithm results in its group, having the highest mean score and intensity images that showed little or no variation with modulation

frequency. Failure in this respect would violate one of the key assumptions in the experimental method, and could therefore invalidate the subjective data (as explained in section 5.2.6).

The decision to include six models in Groups I and II was reached through calculation of the duration of each listening test; it was a compromise between gathering as much data as possible and limiting the time each participant would need to commit to the study. This was largely influenced by the pairwise comparison method chosen for experimentation. The issues relating to duration in listening tests are discussed further in sections 5.1 and 5.2.4; six models per group, allowing for up to two repeats by musical extract, were chosen as the best compromise. This was considered sufficient to gather enough data to address the experimental question without placing too great a demand on participants.

Group III contained only four models. The number of test systems in this group was reduced for two reasons: partly due to the limited time available for this experiment, but mainly because of its preliminary nature. It was unknown before testing whether most listeners would be able to detect any differences between these systems; the differences were expected to be generally more subtle than for the comparisons in Groups I and II. It was therefore considered unwise to subject participants to a lengthy experiment, making many comparisons in a session where they might quickly become frustrated due to a consistent lack of audible differences, potentially leading to a sense of 'failure' at their inability to complete the task. In addition, it was of interest to include a greater variety of source material for evaluation in this experiment (as explained in section 5.3.3.3); this increased the number of trials each participant had to complete, so the inclusion of four models was considered a sufficient range to be able to address the experimental question without unnecessarily extending listening session duration.

4.1.3 Design Strategy for Groups I and II

Groups I and II were constructed to represent a mix of design strategies, each modelling three sealed- and three ported-cabinet loudspeakers, one of which had a protection filter. As described in section 1.1.3.1, these high-pass filters are sometimes added to prevent excessive diaphragm excursion at low frequencies, but increase phase shift and the steepness of the roll-off in this region. It was therefore appropriate to evaluate this type of design to see how it affected audible judgements as well as algorithm results.

As stated in section 3.2.1, the intention was to simulate similar-sized monitors as these would likely be used and mounted in similar ways during real mixing; a group of mid-sized monitors are also likely to show a range of design strategies, featuring both sealed and bass-reflex cabinets, and use similar-sized drive units; therefore, the real drivers upon which the models were based were all between 6 and 8 inches nominal diameter.

It is usual in bass-reflex designs to tune the port to a similar frequency as the drive unit free-air resonance. This was used as the starting point for selecting f_p in the models, but the port tuning frequency was set very low in two models in Group I to see how this strategy affected results; as mentioned in section 1.1.1, this approach may be used in high-fidelity loudspeakers as it offers some of the benefits of using a vented cabinet whilst reducing the negative impact on transient response. The ported models in Group II were tuned to produce more typical vented-box alignments.

Following these considerations, Groups I and II featured a balanced range of alignments and fundamental design strategies typical of the real systems they were supposed to represent.

However, as mentioned in section 4.1.1, these groups were constructed with slightly different priorities. The main differences within Group I were in cut-off frequency (taken as the point where the pressure response was 3 dB down from the mean passband level). Audibility of differences for listeners was a concern in this first set of experimental models; the focus was on including large differences in bass extension, under the assumption that all participants should be able to detect this type of variation quite easily. The cut-off frequencies in Group I therefore covered a very wide range: 20–166 Hz at the extremes, otherwise varying between 40 and 79 Hz. Variations in phase response between the models were present but generally occurred towards the lower end of the frequency range; appreciable differences between four of the systems in Group I only existed below 60 Hz. The priority for Group II was to make differences between the models more realistic, i.e. more typical of the variations in alignment that could be expected when comparing similarly sized professional mix monitors. The listening task was expected to be more difficult as cut-off frequencies within the group were more similar: 40–94 Hz at the extremes, otherwise ranging between 53 and 65 Hz. The chosen alignments produced differences in phase response across the group throughout the full analysis range of the algorithm; these were greatest between approximately 40 and 100 Hz, with the differences between sealed and ported models being overall more pronounced than in Group I. Thus, the differences in response between these models occurred in a frequency range where real monitors of this type are likely to differ most, and where music contains the fundamental content of bass and drums which form the rhythm section. Of special interest in Group II were the model pairs that featured different alignments but similar cut-off frequencies. In particular, the comparison of models D and G in this group were designed to present a 'classic' sealed- vs ported-cabinet design trade-off. They were made to have almost identical cut-off frequencies that are realistic extension limits for monitors of their size, in a critical region for rhythm section reproduction. Comparing just the extension of these models suggests that they are equivalent reproducers of bass content. It was of interest to include this type of comparison in the experiment because distinguishing between two monitors that have nominally the same bass extension is where applying the MTF algorithm is especially useful; the cut-off frequency is typically stated on a product data sheet in the absence of other useful information that describes the alignment and its impact on dynamic signals. It was expected that the algorithm would show clear differences between the models in such pairs, but it was not known how participants in the experiment would respond to this type of comparison, because the audible impression is not as straightforward as one reproducing overall more bass content than the other.

4.2 Group I Models

Table 4 shows the key design features of virtual loudspeakers in Group I, including the approximate f_c (-3 dB frequency); these factors help to determine the low-frequency alignment of each model. The full electroacoustic parameters used as inputs to the *WooferMaker* program are given in Appendix F.

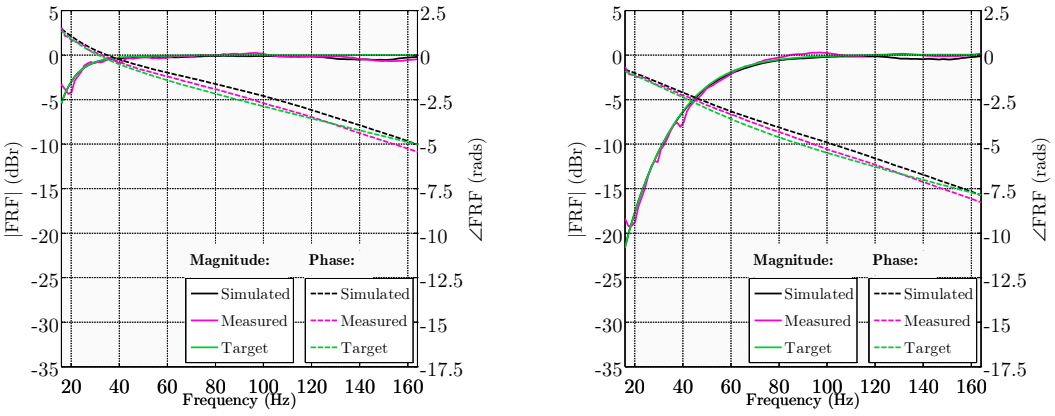
Label	Cabinet	Protection filter?	$\approx f_c$(Hz)
R	Sealed	N	20
C	Sealed	N	53
D	Sealed	N	162
E	Ported	N	79
F	Ported	Y	41
G	Ported	N	40

Table 4: Key design features for Group I virtual loudspeakers

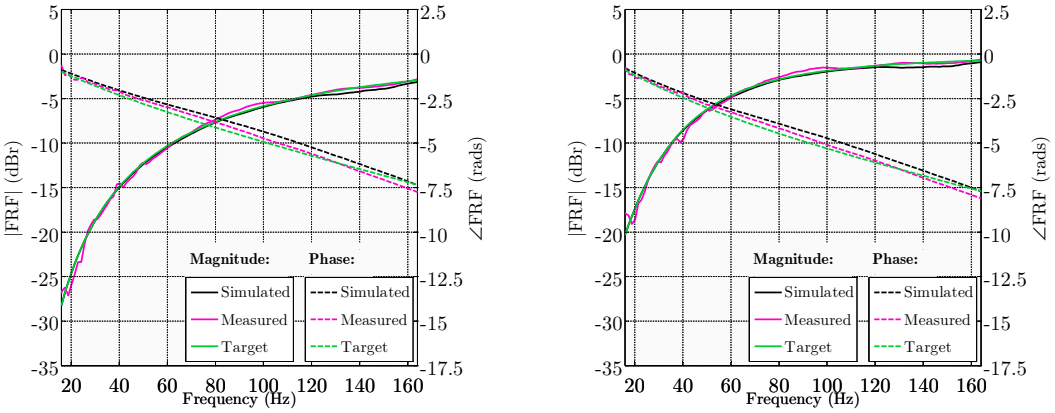
This model group contained a variety of LF alignments, including common design strategies and therefore showing a distribution in both magnitude and phase behaviour. System labels were arbitrarily assigned except for R; this was the reference against which other models were compared in subjective testing. System R was regarded as an idealised monitoring system, having a very extended low-frequency response and gentle phase characteristic. Model D was regarded as a somewhat extreme example in this group, having a very high cut-off frequency. It was not known how obvious the audible differences between the models would be for the majority of test participants; if they could not distinguish D from the other systems in this group, this would indicate an experimental fault, and might mean that further subjective testing could not be performed. With regards to the MTF analysis, it was of interest to see how the algorithm would respond to a loudspeaker that had a gradual low-frequency roll-off but was severely lacking in bass output. If the algorithm did not reflect both aspects of this behaviour, it would suggest that the method might need some further adjustment.

4.2.1 Response Measurement: Group I

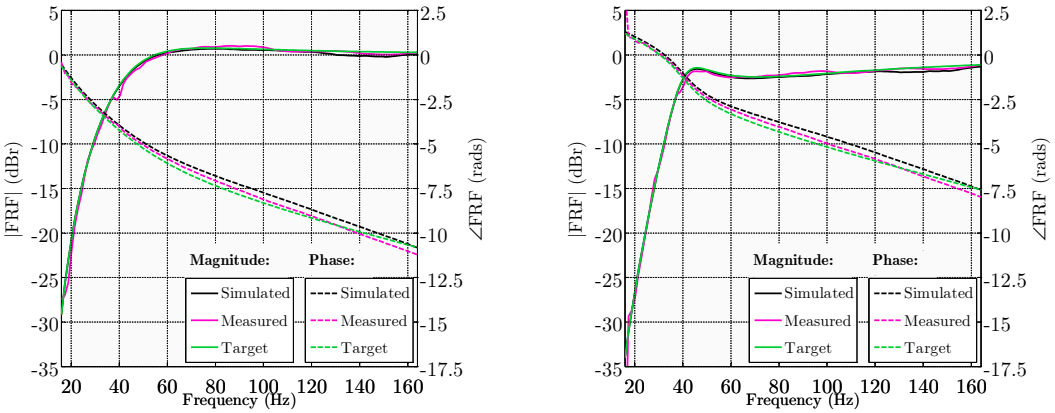
All responses shown in this chapter were measured using the method described in section 3.3.1 and the processing technique to simulate the loudspeaker models at the listening positionshown in section 3.3.4, except that a single white noise extract was used as the source signal instead of different musical extracts. With reference to Eqn. 3.6, this source signal was also used for the frequency response estimation instead of the loudspeaker input signal; this allowed the responses of the virtual loudspeakers to be measured, rather than that of the real experimental loudspeaker. All responses were computed from recordings of 178 s duration (478 averages with 2^{15}- point transforms). Figure 40 shows the simulated responses against the measured equivalents; the original targets are also shown for comparison. The target is the low-frequency alignment produced by the lumped parameter model, and the simulation is the result of imposing this target on the equalised physical loudspeaker response, described as $R_i(f)$ and $S_v(f)$ respectively in section 3.3.4. Linear frequency axes have been used with limits corresponding to that covered by the MTF algorithm.



(a) R **(b)** C



(c) D **(d)** E



(e) F **(f)** G

Figure 40: Measured, simulated and target responses for Group I virtual loudspeakers. Linear frequency limits cover the same range as the MTF algorithm (16 to 164 Hz). Magnitude is shown on the left axis and data is marked by solid lines; phase angle is shown on the right and marked by dashed lines.

Some points regarding differences between the target, simulated, and measured responses are noteworthy here. In chapter 3 it was described how an inverse filter was created to equalise only the low-frequency region of the physical loudspeaker, and this filter was created from a measurement of the loudspeaker's response taken at the listening position with all experimental equipment in place; but this inverse filter was only as accurate as the measurement upon which it was based. If there was any difference in the system transfer function between the measurements presented here and the original measurement, the equalisation would not be completely accurate and the simulated response would not exactly match the measured equivalent. Repeating the position of the loudspeaker across measurement events in this study was considered accurate; alignment of the rectangular loudspeaker stand bases against the square floor grids in the chamber made it easier to ensure that the same distance and angle relative to the listening location was repeated. Consistent on-axis positioning of the microphone capsule at the approximate location of a listener's head across different measurement sessions was more difficult; however, as described in section 3.1.1, listener head movements were not controlled during experiments, so some discrepancy between the ideal and 'human receiver' position was already accepted as a source of potential error.

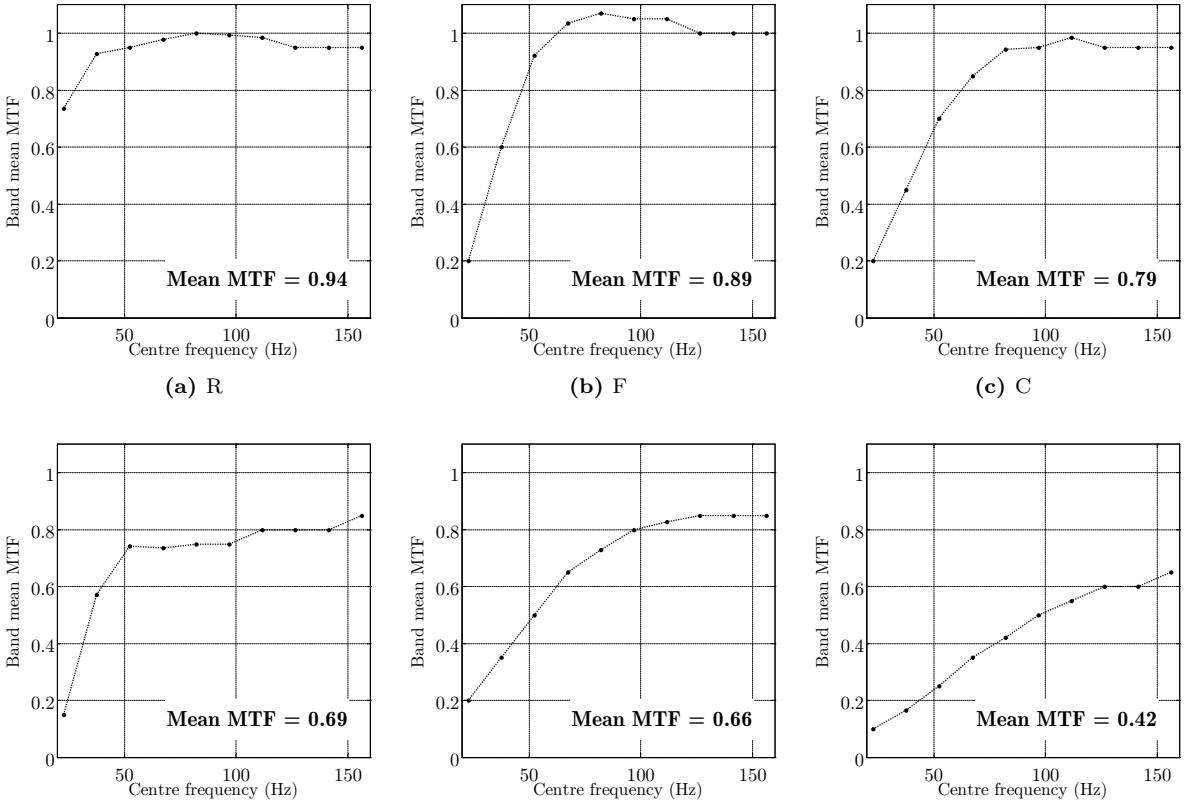
The fact that the chamber was not anechoic down to the lowest frequencies of interest is another possible reason for changes in the transfer function between measurement sessions. The results presented in section 3.3.1.4 showed that measurement of the physical loudspeaker response was generally consistent but did show some variation when measured on different occasions, especially at very low frequencies; the largest low-frequency deviations observed in those measurements occurred below 25 Hz, being approximately 0.4 rads and 2.3 dB.

On reflection, it is believed that an averaging approach to measuring the experimental responses would have been useful, perhaps repeating at the start of each day of subjective testing; this would have allowed a comparison of the simulations against an average of the responses measured throughout the entire testing period, giving greater confidence that the results were consistent for all listening sessions and were therefore not a source of bias in the subjective judgements. It should be noted that the subjective results (presented in chapter 7) did not give any reason to believe such an error existed in this study, but a more systematic approach to checking the reproductions would have been helpful if bias of this nature had been suspected.

Also related to these points is the fact that the modelled target responses were noise-free; any measurement noise or changes in the transfer function of the system being equalised would show as a deviation of the simulated and/or measured response from the target function. In Figure 40 it can be seen that by 160 Hz, the phase of the simulated and measured responses have a steeper gradient than the target; this is believed to be an effect from the crossover at 250 Hz between woofer and mid-range drivers in the experimental loudspeaker that was not modelled in the targets [128]. As the target responses were models rather than derived from measurements taken inside the real experimental environment, they were not considered sufficiently representative of listening conditions and were not used to calculate the MTF algorithm results shown in this chapter.

4.2.2 MTF Assessment: Group I

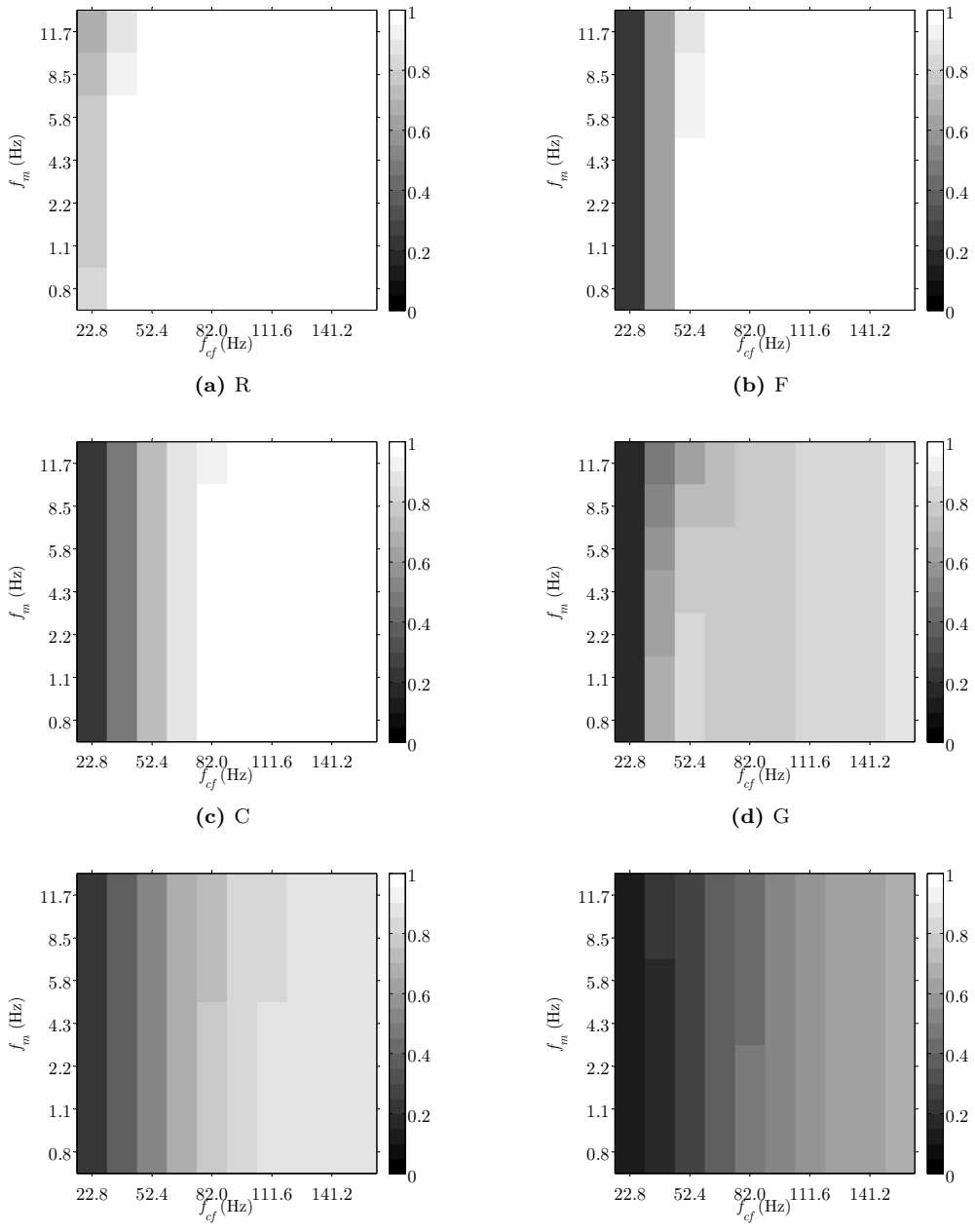
The MTF algorithm was applied to the simulated responses shown in Figure 40. The mean-band plots and intensity images are shown in Figures 41 and 42 respectively, arranged in order of descending mean score $\overline{M}$. Full numeric results are listed in Appendix G, Table 31.



(a) R **(b)** F **(c)** C

(d) G **(e)** E **(f)** D

Figure 41: Group I MTF results: mean-band plots (from simulated responses)



(a) R **(b)** F **(c)** C **(d)** G **(e)** E **(f)** D

Figure 42: Group I MTF results: intensity images (from simulated responses)

4.2.3 Discussion of Group I Objective Results

Inspection of the MTF results in comparison to the steady-state responses and design parameters listed in Table 4 led to the following conclusions:

- MTF results, numerical and visual, showed the reference system R to be the most accurate system, as intended.
- All test systems returned differences in numerical and visual MTF results, as was expected given the nature of the differences in their LF alignments.

- The algorithm is effective in illustrating differences between alignments of the sealed-box loudspeakers, R, C, and D, with the overall shading of the intensity images indicating how their relative LF extensions compare. This is seen most clearly by comparing systems C and D. They both show an intensity image with distinct vertical bands and a smooth horizontal transition from dark to lighter shades, i.e. lower to higher values for m. However, the much higher f_c of loudspeaker D gives it an overall much darker intensity image and lower value of $\overline{M}$; these results therefore indicate that loudspeaker D will not reproduce content as accurately as C in the range covered by the MTF analysis. As previously stated, it was expected that this effect should be seen when model D was included in this group.

- The algorithm is effective in distinguishing between different types of alignment. The comparison between C, a sealed-cabinet design with a well-damped response, and G, a ported model with a more extended response but underdamped system resonance, illustrates this most clearly. The mean MTF score is lower for G, and the intensity image shows more abrupt transition between dark and light bands, reflecting the extended response but rapid attenuation below f_c; it shows multiple shades of grey within a single band, signifying that m varies as a function of f_m, i.e. faithful reproduction of a signal's envelope decreases with increasing modulation frequency; it also shows an isolated 'bright spot' in an otherwise darker region, suggesting that the system has an underdamped resonance and identifying the band where it occurs. In contrast, loudspeaker C shows a gradual horizontal transition between dark and light bands, as would be expected from an alignment with very gradual attenuation below f_c; there is no variation in shading within a given band, suggesting that it transmits the input signal's envelope consistently, regardless of modulation frequency. Based on this comparison, it might be stated that the algorithm shows characteristic behaviour depending on whether the loudspeaker is sealed or ported. This is a generalisation that will not be true for all comparisons. For example, it would be possible, but unusual, for a well-designed sealed-cabinet loudspeaker to have an alignment with a high Q_{TS} , the quality factor of the overall response at f_c [18]. Also, models D and E, sealed and ported designs respectively, show different $\overline{M}$ scores and band-mean plots, but exhibit similar characteristic behaviour in their intensity images; the transition between darker and lighter bands is smooth, and there is no variation in shading as a function of f_m apart from what could be described as a 'step' change. Based on the comparison of these systems, it seems that this is the characteristic intensity image behaviour for overdamped alignments.

- The intensity image in its current form does not sufficiently illustrate the behaviour if an underdamped resonance produces $m > 1$. This should be shown in the plots because it means that a loudspeaker is not simply failing to faithfully transmit information; the resonant behaviour is adding extra information not present in the original signal. This effect was observed in loudspeaker F; from the intensity image alone it appears as if this system produces near-perfect results, but inspection of individual matrix elements shows that the highest theoretical MTF score has been exceeded in some locations. This behaviour can be seen in the band-mean plot; this means that either both forms of plot must be presented, or the colour scheme for the intensity images must be adjusted to enable such behaviour in the MTF matrix to be shown if it is encountered. Considering the

numerical output, it seems reasonable to consider any deviation from $m = 1$ as degradation in reproduction accuracy regardless of direction, i.e. both an increase and decrease in m after passing through the loudspeaker under test should be considered as distortions, but the algorithm does not currently contain a way to reflect this in the overall mean score.

4.3 Group II Models

Table 5 shows the key parameters for the Group II virtual loudspeakers. As in Group I, system labels have no inherent meaning except for the reference, R.

Label	Cabinet	Protection filter?	$\approx f_c$(Hz)
R	Sealed	N	40
C	Sealed	N	92
D	Sealed	N	55
E	Ported	Y	62
F	Ported	N	64
G	Ported	N	53

Table 5: Design (target) features for Group II virtual loudspeakers

Note that these models have a similar distribution of design strategies to those in Group I but with a much smaller spread of values for the point at which low-frequency attenuation begins. This was expected to produce smaller audible differences between models due to bass extension alone. However, the cut-off frequencies are generally higher, with the lowest value being a full octave above the Group I equivalent. Selection of these models into the same experimental group was considered to have several advantages:

i) Realistic variations in system alignments were preserved whilst reducing the likelihood that discrimination between designs was on the basis of gross magnitude variations alone.

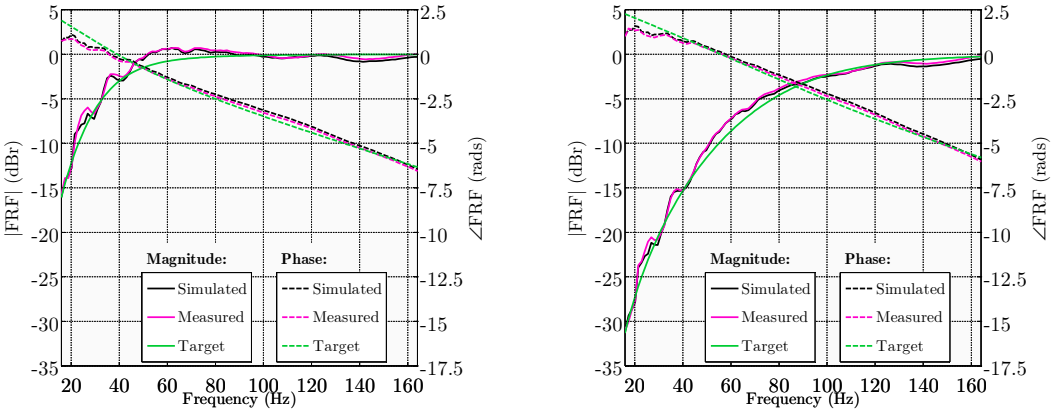
ii) Differences generally occurred higher in the frequency spectrum so that the main behaviour of interest was less likely to fall below the threshold of audibility at reproduction levels used in the listening tests [96].

iii) It was easier to find musical extracts with greater signal content in the region where differences between the alignments occurred [12].

iv) There was less risk of non-linear distortion due to excessive driver excursions when equalising the experimental loudspeaker at very low frequencies.

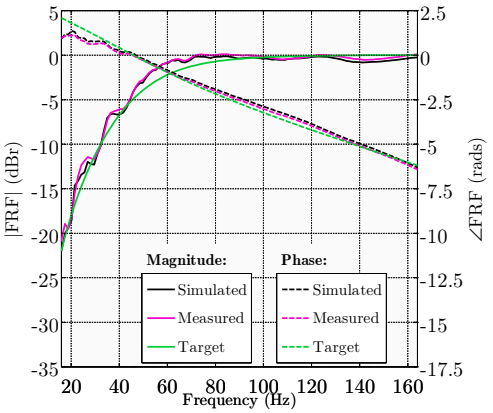
4.3.1 Response Measurement: Group II

Figure 43 shows the simulated responses against the measured versions.

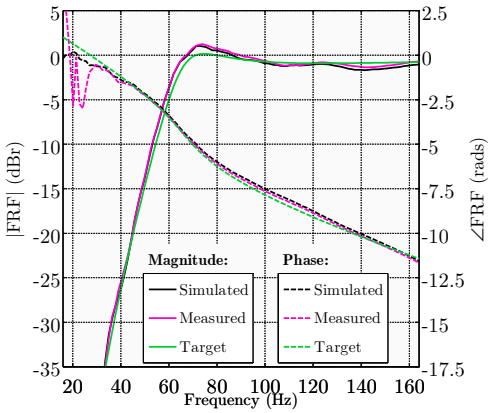


(a) R

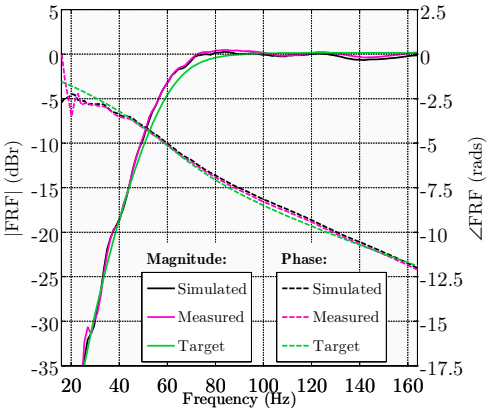
(b) C



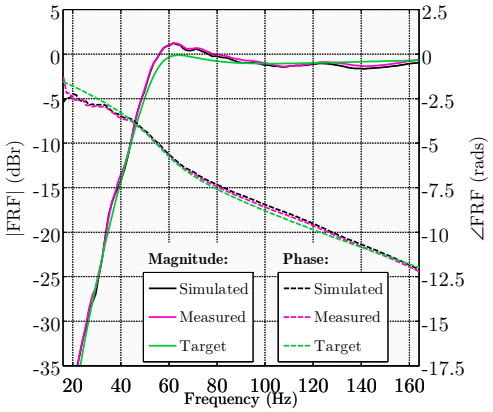
(c) D



(d) E



(e) F

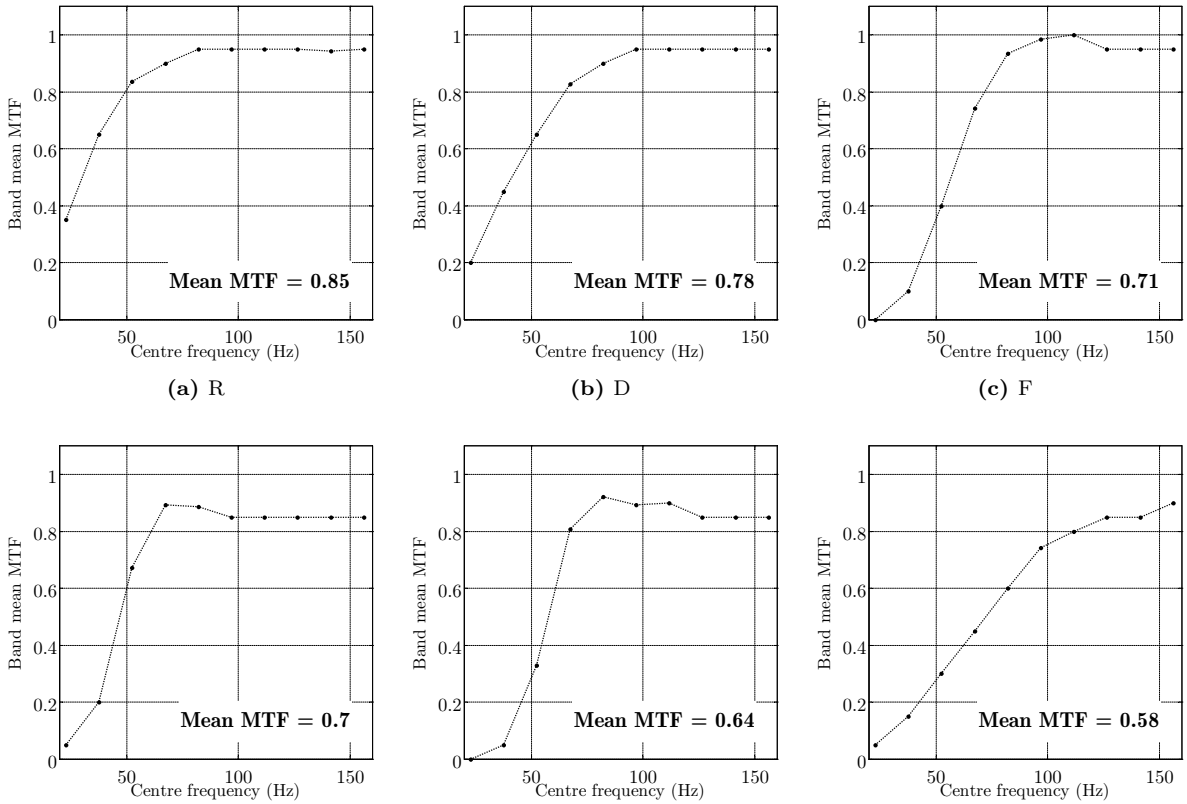


(f) G

Figure 43: Measured, simulated and target responses for Group II virtual loudspeakers. Linear frequency limits cover the same range as the MTF algorithm (16 to 164 Hz). Magnitude is shown on the left axis and data is marked by solid lines; phase angle is shown on the right and marked by dashed lines.

4.3.2 MTF Assessment: Group II

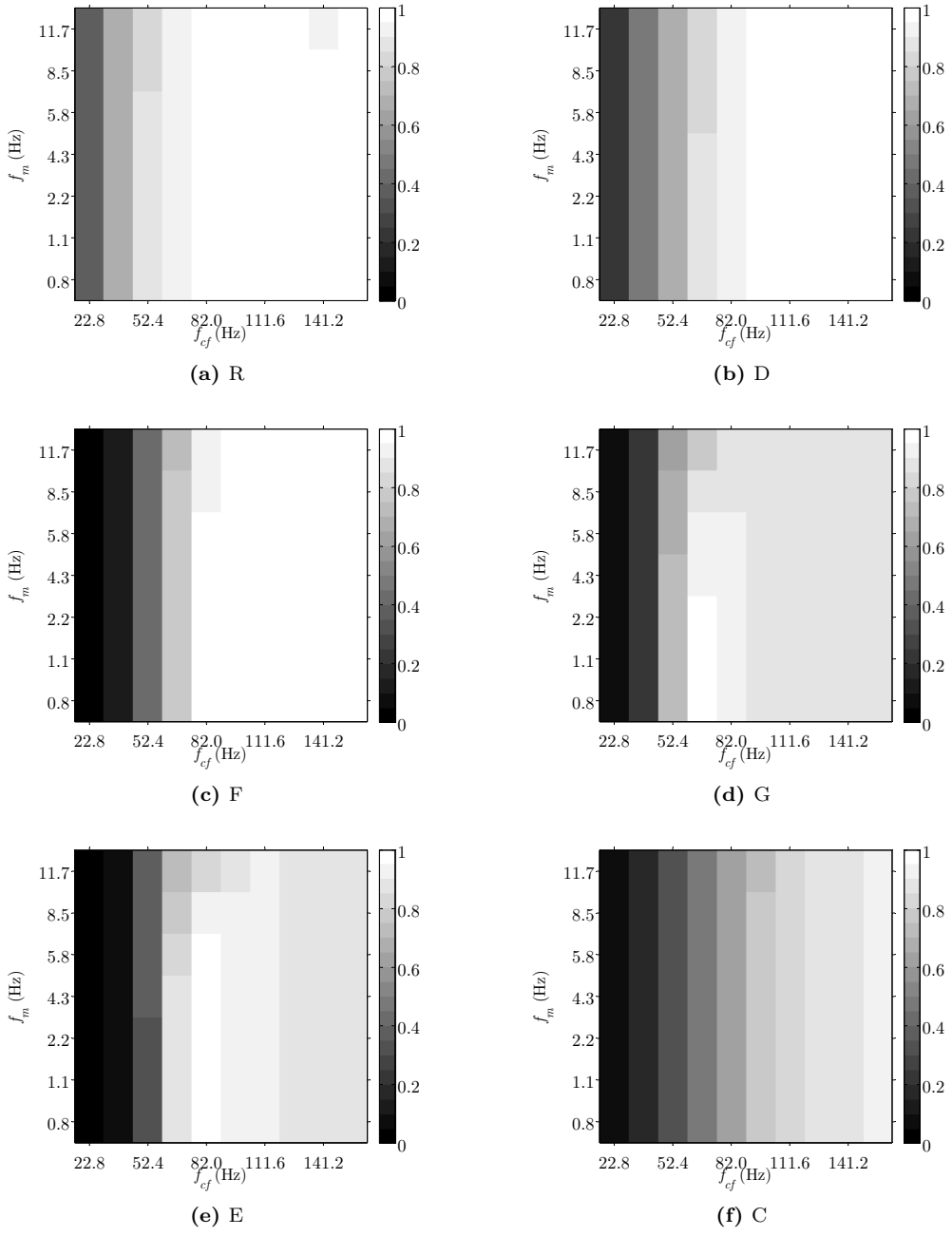
Figures 44 and 45 show the MTF results following analysis of Group II models. Results for each virtual loudspeaker are presented in order of decreasing mean MTF, $\overline{M}$. Full numeric results are listed in Appendix G, Table 32.



(a) R **(b)** D **(c)** F

(d) G **(e)** E **(f)** C

Figure 44: Group II MTF results: mean-band plots (from simulated responses)



(a) R **(b)** D

(c) F **(d)** G

(e) E **(f)** C

Figure 45: Group II MTF results: intensity images (from simulated responses)

4.3.3 Discussion of Group II Objective Results

Conclusions following inspection of the design parameters and results from analysis are presented below.

- Reference loudspeaker R produced numerical and visual MTF results that identified it as the most accurate system compared to others in the group.
- All systems in the group produced different MTF results but with an overall smaller range of $\overline{M}$ values compared to Group I: $\overline{M}_{\Delta \mathrm{I}} = 0.52$, $\overline{M}_{\Delta \mathrm{II}} = 0.27$, where $\overline{M}_{\Delta}$ denotes the

difference between highest and lowest mean MTF score within a group. This behaviour had been expected, given the intended differences between the two groups, as previously discussed.

- The sealed-cabinet loudspeakers, R, C, and D, show similar characteristic intensity image behaviour; they produce distinct vertical banding, with a gradual horizontal transition from black to white, most clearly seen in system C. The bass-reflex models, E, F, and G, show less distinct vertical banding, indicating that m is varying as a function of f_m. They also produce more abrupt changes from black to light grey or white bands, indicating a steep roll-off below the system resonance.

- The bass-reflex loudspeakers produce a similar characteristic shape in their band-mean plots but not necessarily in intensity image; for example, systems E and G show similar behaviour in intensity image but different values for $\overline{M}$, whereas systems F and G produce different intensity images but almost identical $\overline{M}$ scores. As observed with the Group I systems, the characteristic appearance in intensity image seems to reflect the different overall system Q_{TS}; the higher-Q alignments of E and G produce isolated brighter regions in the intensity image that are easily distinguished from the darker surrounding values. This cannot be seen in the intensity image for loudspeaker F; it has a broad, low-Q resonance in the transition region and this 'blends' into the surrounding values of m that are already close to 1. The band-mean plot does show this behaviour.

4.4 Collated Results: Groups I and II

This section adds to the conclusions in sections 4.2.3 and 4.3.3 regarding differences observed between the sealed- and ported-cabinet models simulated in Groups I and II. Results were collated to perform an additional comparison between these two fundamental loudspeaker design strategies. Figure 46 shows the mean score, $\overline{M}$, plotted against cut-off frequency ($f_c = -3$ dB) for the models in Groups I and II. The data was taken from the simulated responses corresponding with the computed algorithm results shown in sections 4.2.2 and 4.3.2. The data for sealed and ported models have been separated to show their contrasting behaviour.

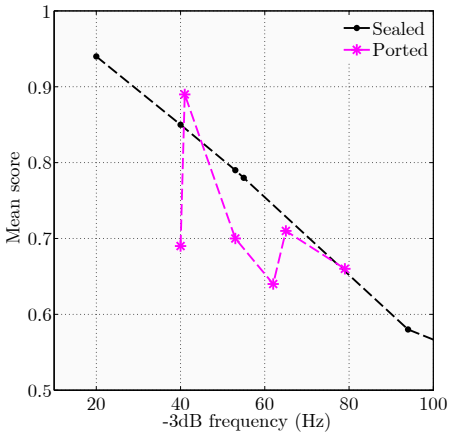


Figure 46: Changes in mean score with low-frequency extension: Group I and II models only. Data from sealed- and ported-cabinet models are plotted separately.

The sealed models have similar alignments and therefore differ primarily in low-frequency extension; it can be seen in Fig. 46 that for these systems, mean score decreases approximately linearly as f_c increases. This is not the case for the ported models. They have more complex differences between their alignments, and it can be seen that a more extended response does not necessarily return a higher mean MTF score. Note that the data point for model D in Group I has been excluded here. It caused a deviation away from the linear characteristic of the other sealed-cabinet models, as can be seen by the dotted line extending beyond the upper limit of the x-axis. The reason for this deviation is unknown, but it was treated as an outlier here because: 1) the $f_c = 166\,\text{Hz}$ is so high that it is not representative of typical professional mix monitors; 2) the low-frequency roll-off begins above the upper analysis limit of the algorithm (160 Hz). This is possibly the reason for the deviation in behaviour compared to the other models shown here, and indicates that analysis of such systems is not reliable given the current frequency range of the method.

4.5 Group III Models

The third group of experimental models was designed using a different approach to the others. Although the mid- and high-frequency response of the main virtual loudspeakers was the same, as described in section 3.1.1, the response still naturally varied in magnitude within the bass region. A set of models was developed that combined a fixed magnitude response with the phase from different order filters. This was proposed as a way to investigate the perceptible impact of differences only in phase response, representative of the behaviour that may be seen in studio monitors of different but commonly-encountered design strategies. These models allowed the sensitivity of the algorithm to phase distortion alone to be investigated.

4.5.1 Generating Artificial Systems

The experimental target responses were filters of fixed magnitude but differing phase, taken from increasing even-order Butterworth high-pass filters; the transfer function equations are given in Appendix H. Using polar notation, the target responses can be described as:

$$z_n = r_m \angle \theta_n \begin{cases} m = 2 \\ n = 2, 4, 6, 8 \end{cases} \tag{4.1}$$

These were created using:

$$z_n = a + \mathrm{j}\, b = r_m \cos\theta_n + \mathrm{j}(r_m \sin\theta_n) \tag{4.2}$$

where: a and b are the real and imaginary parts of complex number z_n;

r_m is the modulus (magnitude) of filter m: $r_m = |m| = \sqrt{a_m^2 + b_m^2}$;

θ_n is the argument (phase angle) of filter n: $\theta_n = \arg(n) = \tan^{-1}\left(\frac{b_n}{a_n}\right)$.

A cut-off frequency, -3 dB from the mean passband value, of 60 Hz was chosen; this is in a critical region for a typical rhythm section, where a bass guitar and kick drum have dominant content. It is therefore a region of particular focus for a sound engineer during the mixing process. This value is also a reasonable half-power limit for small professional monitors [135–137].

Note that the models are non-minimum phase systems, except for when $m = n = 2$; for other values of n, the model has the magnitude of a 2$^{\text{nd}}$-order HPF but the phase of a 4$^{\text{th}}$, 6$^{\text{th}}$ or 8$^{\text{th}}$ equivalent filter. Table 6 shows how these models approximated the order of phase shift that would be expected from real-life monitors, using the assumption that a loudspeaker response at low frequencies can be modelled by a high-pass filter, as discussed in section 3.2.1.

n (filter order)	Loudspeaker with equivalent phase response
2	Sealed cabinet
4	Ported cabinet
6	Ported cabinet with 2$^{\text{nd}}$-order protection filter
8	Ported cabinet with 4$^{\text{th}}$-order protection filter (extreme example)

Table 6: Approximations of filter orders to real loudspeaker systems

The effect in the time domain of this increasing low-frequency phase shift is illustrated in Figure 47. The target impulse responses are plotted together, with axes limits allowing the differences to be viewed clearly. It can be seen that ringing in the time domain increases with the order of the filters.

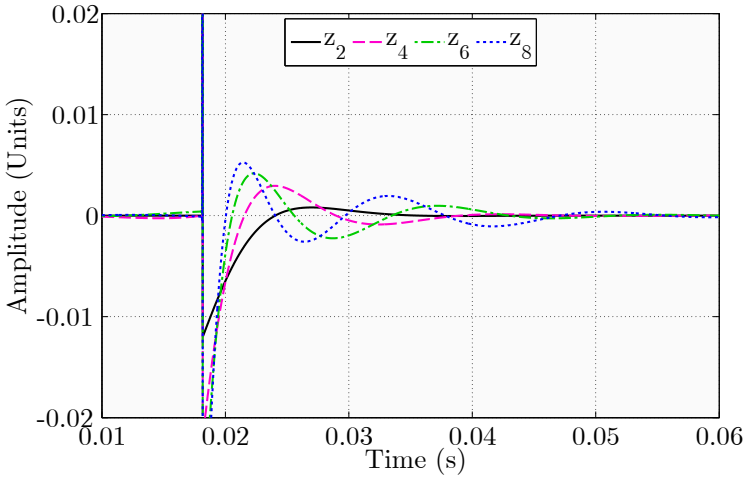
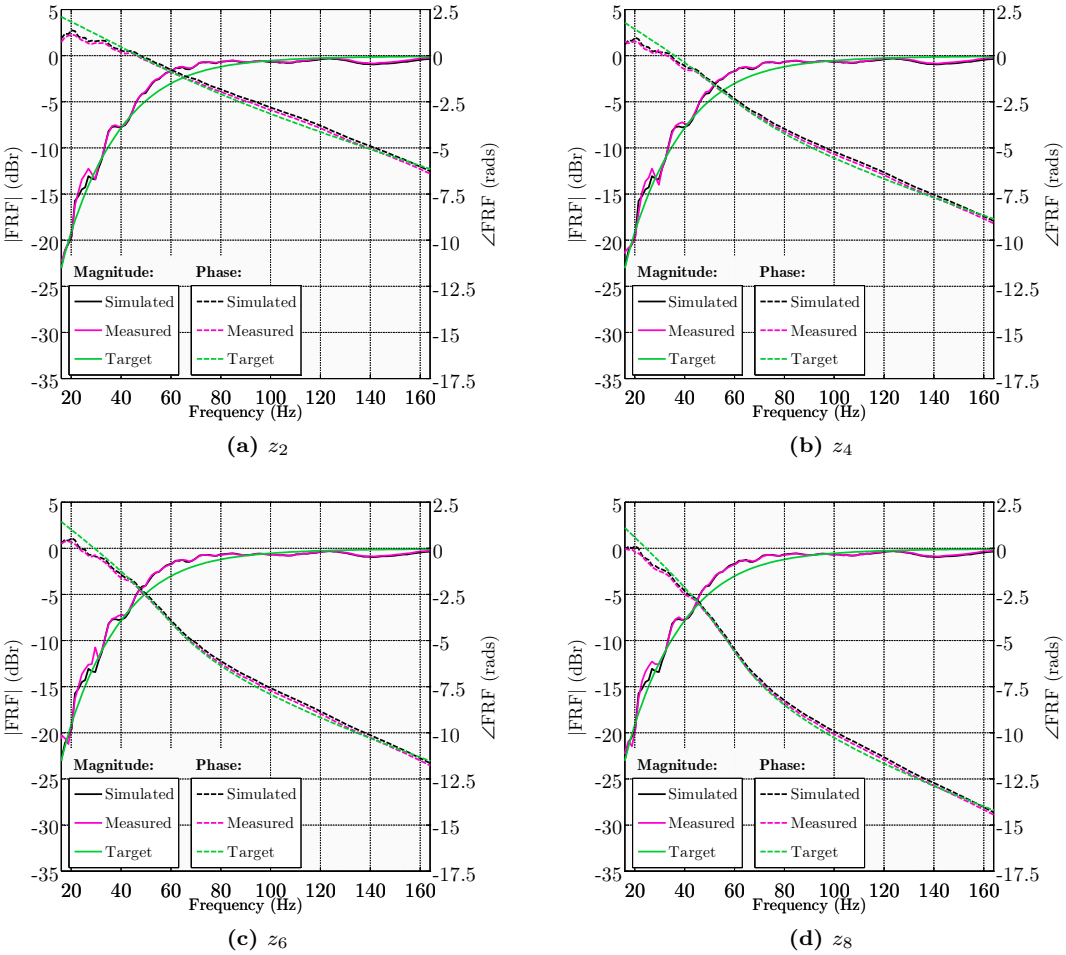


Figure 47: Partial view of impulse responses: targets for Group III models, z_2–z_4. Magnitude response is identical for all four models but their low-frequency phase shift increased from 2$^{\text{nd}}$ to 8$^{\text{th}}$ order below 60 Hz. Ringing in the time domain increases as the phase response deviates further from a linear characteristic; the system takes longer to come to rest after excitation.

Note that it is the frequency-dependent nature of this shift that degrades the reproduction accuracy of the system. As described in section 1.1.3.1, a non-linear phase characteristic means that different frequencies in a signal passing through the system will be delayed by different amounts, leading to a change in the envelope; the waveform of the reproduced signal will be different from the original.

4.5.2 Response Measurement: Group III

Figure 48 shows the expected (simulated) responses against the measured equivalents.



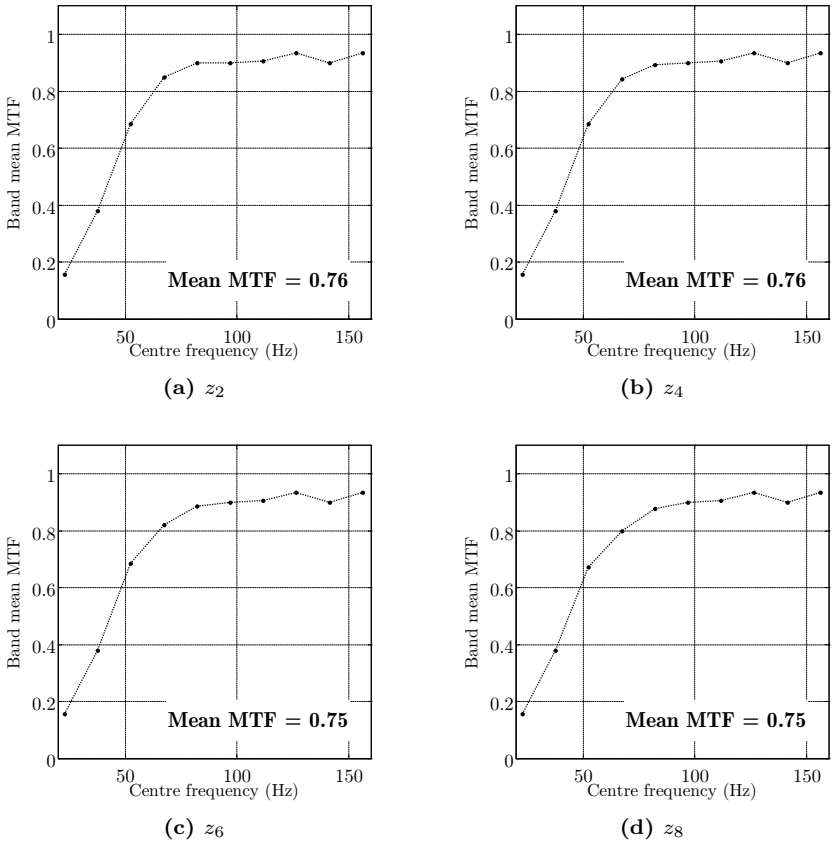
(a) z_2 **(b)** z_4

(c) z_6 **(d)** z_8

Figure 48: Measured, simulated and target responses for Group III virtual loudspeakers. Linear frequency limits cover the same range as the MTF algorithm (16 to 164 Hz). Magnitude is shown on the left axis and data is marked by solid lines; phase angle is shown on the right and marked by dashed lines.

4.5.3 MTF Assessment: Group III

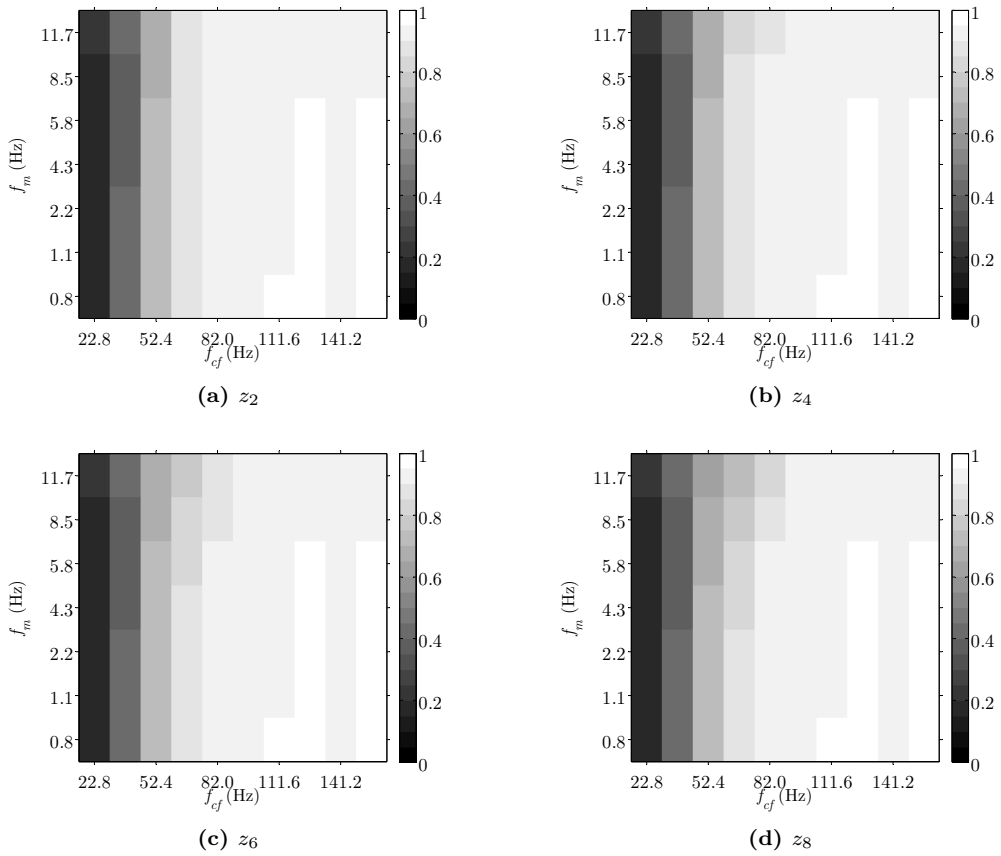
Figures 49 and 50 show the results after MTF assessment of Group III models. Full numerical results are listed in Appendix G, Table 33.



(a) z_2 **(b)** z_4

(c) z_6 **(d)** z_8

Figure 49: Group III MTF results: mean-band plots (from simulated responses)



(a) z_2 **(b)** z_4

(c) z_6 **(d)** z_8

Figure 50: Group III MTF results: intensity images (from simulated responses)

4.5.4 Discussion of Group III Objective Results

The following conclusions were made after MTF analysis of the Group III systems:

- Differences between systems were seen more clearly in the intensity images than in the band-mean plots or $\overline{M}$ scores. Changes in the phase response alone produce a 'spreading' from dark to light from the top left corner of the intensity images, i.e. from highest modulation frequency in the lowest band; increasing orders of phase shift produce spreading over an increased area. Differences are subtle compared to changing magnitude only in otherwise equivalent systems, such as the comparison between sealed-cabinet systems in Group II.

- Differences between systems were too small to be resolved when showing $\overline{M}$ to 2 d.p. Increasing resolution to 3 d.p. was sufficient to reveal the differences:
  $\overline{M}_{z2} = 0.758$; $\overline{M}_{z4} = 0.755$; $\overline{M}_{z6} = 0.752$; $\overline{M}_{z8} = 0.747$;
  This shows that increased phase shift leads to reduced mean MTF score, and suggests that $\overline{M}$ should be routinely stated to three decimal places. Based on these models, there is a decrease in $\overline{M}$ of between 0.003 and 0.005 for each increase in phase shift of two orders. This appears to be a very small change compared to the effect on mean score of increasing the cut-off frequency, as demonstrated by the sealed-cabinet model results in section 4.4. If these changes had not corresponded with judgements of audible differences from listeners

(discussed in sections 7.4 and 8.1), it might have been concluded that the algorithm was not sensitive enough to this behaviour, but data from this study showed that the changes in mean score of this size were sufficient to discriminate between the models. It must also be noted that, as described in section 4.5.1, the Group III models were not considered realistic in that only the order of their phase response about a single cut-off frequency was changed; a comparison of real monitors would therefore not present algorithm results like those returned for this group. Looking at the results for Group I and II, it was concluded that the variation in m scores with modulation frequency for monitors of realistic alignments, varying in magnitude and phase, made an important contribution to the characteristic behaviours seen in the results matrices, and therefore in the intensity images, which made the format so useful in describing and comparing different types of low-frequency response. It was concluded for these reasons that the sensitivity of the algorithm was sufficient.

- Model z_2 was intended for use as the reference model in subjective testing; the MTF results confirmed that this was the best model within the group to use for comparison.

4.6 Summary of Objective Evaluation

Three groups of experimental models were presented in this chapter. Groups I and II were considered to be virtual loudspeakers, having a range of alignments representative of the types of response that might be observed in real monitoring systems. These were developed using electroacoustic design parameters, selected to produce a range of realistic behaviours at low frequencies. The third set of models, Group III, were artificially constrained to differ only in their phase response about a single frequency, defined as being in a critical region for bass reproduction in music. These simplistic models were created to investigate sensitivity of the MTF algorithm to the range of phase shifts typically encountered in studio monitors. These models were not strictly classed as virtual loudspeakers because they were not representative of realistic monitor responses; however, they were considered an effective way to quantify the contribution of non-linear phase shifts to the algorithm results in a controlled manner that could later be assessed for audibility.

For all model groups, it was seen that the measured system responses did not exactly follow the original targets, partly due to the anechoic chamber not being completely anechoic at very low frequencies, and partly due to some deviation believed to be from a crossover in the experimental loudspeaker that was not modelled in the target response. It was therefore concluded that the target responses did not accurately represent the listening conditions during subjective testing, and should not be used for MTF analysis. However, it was observed that the playback simulations produced close approximations to the magnitude and phase of the measured equivalents. The simulation method was therefore concluded to be a satisfactory estimation of the virtual loudspeaker responses that were experienced at the listening position during subjective experimentation.

MTF analysis of each model group was considered successful in that consistent behaviour was seen for similar types of alignment, and systems within each group produced a range of MTF results in line with expected behaviour. Differences between all models within a given group could be seen; this confirmed that the selection and distribution of parameters and design strategies had been successful in producing a range of LF alignments for analysis. Further

confirmation was provided by analysis of the reference models in all three groups; inspection of the MTF results for these systems indicated that they would be the most accurate reproducers of low-frequency content within their respective model sets.

For the virtual loudspeakers in model Groups I and II, the use of realistic response simulations with known design parameters allowed the following general conclusions to be drawn about the algorithm when applied to loudspeakers with different types of LF alignment:

- Similar alignments produce similar characteristic behaviour in the MTF intensity images. Well-controlled sealed-box systems show distinct vertical banding, with a gradual transition across bands from black through increasingly lighter shades of grey. The overall shade of the plot indicates the f_c point; systems with a more extended response produce an image that is overall lighter in appearance. Loudspeakers with steep roll-offs below f_c show an abrupt horizontal transition from dark to light, with little grey separating them. Systems with well-damped alignments show little or no variation in shading in the vertical direction, i.e. the value of m does not vary with modulation frequency. The limits to this effect were not formally assessed, but it is expected to apply to alignments where the system Q is 0.707 or less. Systems with narrow, high-Q resonances produce the most complex behaviour in the intensity images; they display a more obvious rippled or mottled appearance, especially for higher modulation frequencies, and are likely to contain isolated lighter spots or patches in the region of the resonance. Broad, low-Q resonances are the most difficult to identify using the intensity image alone. This indicates that the resolution of the image colour scheme is not always sufficient to reveal this type of behaviour; however, the values currently used are believed to be the most effective for comparing all other aspects of response behaviour described here.

- All three aspects of the algorithm output were seen to be useful. The unweighted mean score $\overline{M}$ allows loudspeakers to be easily compared and ranked according to their ability to accurately reproduce a temporally varying signal at low frequencies. The two plot options provide further information about the system behaviour. Although the intensity image was generally found to be the most revealing, the band-mean plot is perhaps more intuitive and currently the only way to observe signal distortion of the type that causes $m > 1$, occurring when an underdamped system resonance increases the modulation depth of the input signal relative to its original value.

- For the sealed-box systems, having similar alignments apart from their cut-off frequency, the corresponding mean algorithm scores reduced approximately linearly as bass extension reduced; even including the result of a data point considered to be an outlier, $\overline{M}$ decreased monotonically with increasing f_c for these models. Changes in $\overline{M}$ relative to low-frequency extension for the more complex bass-reflex designs were not so easy to describe; this was due to the increased sensitivity to temporal fluctuations of the input signal, and the presence of underdamped resonances in some models. It was observed that systems with this design methodology can produce almost identical scores for $\overline{M}$ but demonstrate differing behaviour in their band-mean and intensity plots.

The 'artificial' Group III systems were modelled using Butterworth high-pass filters, combining a fixed magnitude with increasing even-order phase responses. MTF analysis of these models

revealed a slightly different type of characteristic behaviour in the intensity images. This was a rippled effect at higher modulation frequencies in lower bands, becoming more pronounced, or spreading, with increasing phase shift through the bass region. With these systems, it was found that 2 d.p. was insufficient to reflect the changes in overall MTF score; reporting these values to 3 d.p. showed a decrease in mean score between 0.003 and 0.005 for an increase of two filter orders. Intermediate odd-order responses were not considered. The Butterworth filters used to create these models are not sufficient to simulate the behaviour of all realistic loudspeaker alignments. It is expected that repeating the analysis with other common alignments, such as Chebyshev, would produce greater differences in mean score as they are known to degrade transient response to a greater extent [19].

The model groups presented in this chapter were seen to have been suitable for evaluating performance of the MTF algorithm in more depth than had been possible in the initial validation. Each group contained loudspeakers with a variety of LF alignments that were clearly differentiated by the algorithm through a combination of its numerical and visual outputs. However, to fully validate the algorithm, these results had to be compared against listener judgements of the same systems when reproducing music; the algorithm may only be considered a truly useful indicator of bass reproduction accuracy if it demonstrates some correlation with perceived performance. Preparation for gathering this subjective data is discussed in chapter 5.

5 Listening Test Design

Chapter 4 showed the results from objective analysis of the experimental loudspeaker models after application of the MTF algorithm developed in chapter 2. This chapter addresses the way in which further experimental data would be gathered through listening tests.

The primary objectives of the research project were to develop an MTF-based algorithm for evaluating loudspeaker fidelity at low frequencies, and then look for evidence that such a method had perceptual relevance, i.e. showed some correlation with listener impression. Gathering useful and credible subjective data was therefore a crucial element of the project. Selection and execution of an appropriate test strategy was a major decision within the project and determined the usefulness of the subjective data. The experimental aims had to be balanced against a number of limiting factors to develop a method that produced meaningful and robust data without excessive demand on resources or participants.

Section 5.1 describes the aims and requirements that were defined for the subjective experiments; these informed the final test strategy, described in section 5.2, along with how and why it was chosen. Other aspects critical to the success of the listening tests are described in sections 5.3 to 5.5: selection and processing of test stimuli, development of testing software that was used to execute the experiments, and consideration of practical matters affecting the test procedure.

5.1 Experimental Aims and Requirements

Very little was known before the project about how results from the MTF algorithm would relate to what listeners perceived. Two specific aims were identified as relevant to the study:

i) Demonstrating that perceptible differences existed between the loudspeaker models, differing only in their low-frequency alignment. This situation cannot be replicated when comparing groups of 'real' loudspeakers, and it was not known whether the majority of participants would be able to discriminate between all of the alignment variations presented in the experiments.

ii) If perceptible differences between alignments could be demonstrated, the findings would need to be presented in a format that could be compared with results from the MTF algorithm. Ideally, these effects would be quantified, indicating direction and magnitude of perceived differences. The ultimate aim would be to achieve reliable mapping of subjective responses to numerical results from the MTF algorithm. However, construction of a rating scale for such a comparison was considered a long-term goal; if an appreciable effect could be demonstrated in this study, further testing with systematic variation of alignment parameters over many values would be required. Thus, it was decided that establishing a direction of perceived difference was sufficient at this stage; the numerical MTF results could easily be reduced to ordinal data for comparison with subjective responses.

Two primary features of the experimental method were defined before developing specific details of the testing procedure:

Realistic presentation The acoustic experience should be as close as possible to that which would be encountered in mix monitoring situations. This required the presentation of

music, not artificial signals, reproduced through virtual loudspeakers, not headphones, as discussed in section 3.1.1. The musical stimuli should be presented at SPLs similar to those during mixing; as described in section 1.1.1, a level between 75 and 85 dB SPL would be appropriate.

Discrimination-based It was believed that a listening task based on hedonic judgements [138] would be of no value in this study, where the intended application was studio monitoring; mix engineers do not make decisions based on personal preference, but on perceived imbalances in the overall mix between the reproduction they are presented with and a known 'good' response. Although professional engineers may periodically check their mix through other monitors for comparison, they instinctively learn to make auditory judgements without direct and immediate comparison; inexperienced listeners cannot be expected to demonstrate this skill that can take many years of training and experience to acquire [5]. Therefore, a reference for comparison would need to be provided along with the systems under test so that judgements could be made based on perceived differences from this presentation.

A pragmatic view was taken when considering the test conditions. It was known that, out of necessity, experimental participants would likely be inexperienced listeners, and there was a concern that requiring a commitment of enough time to allow for training as well as testing would deter people from taking part. These restrictions led to some further requirements when deciding on a final test strategy:

Sensitivity Audible differences between stimuli were small in most cases, making discrimination a difficult task for inexperienced listeners. Highly sensitive listening conditions were required, along with a test procedure that aided detection of small perceptible variations. As discussed in section 3.1.1, listening was performed with a single loudspeaker inside an anechoic chamber; this environment minimised the presence of background noise and interference from room effects, and there is evidence that presentation using a single source elicits more critical judgements from listeners [60, 112, 139].

Difficulty A difficult listening task is further complicated by performing an onerous procedure. A test with a simple technique and method of registering responses allows participants to concentrate more fully on the listening task; it also reduces the need for extended training periods to familiarise them with the process, and reduces the likelihood of bias in the data due to misunderstanding of the procedure or response system.

Duration Ensuring short test periods had several advantages. Unlike professional studio engineers, inexperienced listeners cannot be expected to concentrate on a critical discrimination task for long periods of time. Brief sessions reduced the risk of listener fatigue, and therefore, the loss of concentration or interest in the task. The university safety and ethics guidelines imposed limits on the test duration as a function of exposure level; longer test sessions for any individual would have meant reducing the overall reproduction levels. This had to be avoided as it might reduce audibility of the differences being investigated. A less time-consuming experiment was also expected to increase participation level amongst a population that were not paid for their attendance, and

allowed a greater number of sessions to be run within the limited periods of availability for all test equipment and facilities.

5.2 Developing a Test Strategy

A good experimental procedure is efficient, robust, easily repeatable in a controlled manner, and allows strict control over test variables whilst minimising all sources of bias. Development of such a method within acoustics has been investigated in depth, and there is a vast amount of literature available on the topic; only aspects with particular relevance to this project are discussed in detail here.

Selection of a method for the listening tests was based on the requirements and limitations described in section 5.1. Sections 5.2.1 to 5.2.4 discuss the key aspects of the methodology, advantages and drawbacks compared to other common listening test strategies, and explain why it was considered the most appropriate choice for this study. Other key features of the experimental design are covered in section 5.2.5; here it is shown how the final test procedure accommodated the requirements and limitations defined in section 5.1 whilst also fulfilling the criteria for a robust experimental procedure. Finally, subsection 5.2.6 describes the inherent assumptions of the procedure and why they are important.

5.2.1 Method Definition

Discrimination tests are used to investigate whether a difference exists between two or more items; they are especially useful when differences are thought to exist but their nature is not fully understood [140]. An example of such a procedure is the 'ABX' test, and has featured in acoustics literature in relation to listening tests since at least 1950 [141]. The later use by Clark [142] describes a procedure that is more commonly associated with this terminology: listeners are presented with two audible alternatives, A and B, and are forced to choose one of them in comparison to a reference, X. Some of the widely used listening test guidelines [143, 144] describe procedures that are similar to this method, but other sources were found to be more informative when defining the experimental strategy that was used in this study. Clear and detailed standards for subjective discrimination experiments exist within the field of sensory testing; though these documents are intended for use in detection of taste and smell, they can be adapted easily for studies where hearing is the sense of interest. The label 'ABX' is not specifically mentioned in any of these standards, but three sensory testing procedures were considered when planning the experiments for this project; these are summarised briefly below, described in terms of listening:

i) Triangle test [145]: Listeners are individually presented with three audible samples (stimuli); two are identical and listeners must identify the sample that is different. There is no marked reference so this is a three-alternative forced-choice (3AFC) method.

ii) Duo-trio test [146]: Very similar to the Triangle test but one of the two identical samples is always marked as a reference, making this a two-alternative forced choice (2AFC) test. The reference may be fixed (constant in every trial), or may alternate (balanced equally across all trials). Listeners have to decide which of the remaining two sounds are identical to it, or different from it– this must be specified in the instructions.

iii) Paired comparison (PC) test [147]: Listeners are presented with just two stimuli that may or may not be identical. The task is to select the one which is perceived to have the greatest intensity of a specified attribute, e.g. which of A or B sounds loudest. The standard is intended for non-hedonic tests but the general procedure can be used for investigating a listener's preference, unlike the Triangle and Duo-trio tests which can only be used for detecting whether A and B are audibly different (or similar, as appropriate).

Following these definitions, the constant-reference Duo-trio test comes closest to defining the typical ABX listening test format, but the following differences between these methods are noteworthy. In the Duo-trio test, listeners have to choose the 'odd one out' from two alternatives compared to a known reference. The task is therefore easier for the participant than in the triangle test where they have to choose the odd sample out of three, with no marked reference [146]. However, this key distinction means that the Duo-trio test is less efficient because the chances of making the correct selection are 1/2 rather than 1/3; the consequence statistically is that more correct responses are required for a given sample size before the null hypothesis of 'no difference' can be rejected (assuming that a difference really does exist and therefore that rejection of the null hypothesis is appropriate) [148, A.4], [146, Table A.1, Note 1], [145, Table A.1, Note 1]. The other key distinction between these methods is in the expected direction of responses. The statistical implications of this issue are discussed in section 6.4.4, but it relates to the anticipated outcome of the experimental task: is the participant expected to give a certain answer or not? In the Triangle and Duo-trio tests there is always a correct answer to the task; the odd one out is known to the experimenter and this must be selected by the participant. In the PC test, however, the experimenter may or may not predict the way listeners will respond; if they expect them to choose A, then listeners must choose A to be correct. Conversely, the experimenter may not be able to predict the direction of responses in advance, so the correct answer is decided by the participants. This makes the PC method suitable for studies where it is necessary to test whether a directional difference between two items exists but the 'correct' direction is unknown.

Consideration of these standardised methods meant that the features of the ABX procedure in this study could be defined more accurately:

- The experimental method was based on the fixed-reference Duo-trio format. The reference was fixed (always X) and always identified to listeners, and they were always forced to make a choice between A and B (2AFC), even if it required guessing. This technique carries some risk of increased variability in the resulting data, but the approach is widely used because it can lead to very high levels of performance [149, 150]. Insisting on an answer encourages participants to listen more carefully when audible differences are small, and prohibits the tendency of inexperienced listeners to be conservative, i.e. say 'not sure' rather than risk giving an incorrect response [142, 151].
- Some trials contained two identical loudspeaker models and one different, making them true to the Duo-trio format. In these 'hidden reference' trials, listeners unknowingly compared the reference against itself. Therefore, these trials had a definite predefined correct answer; listeners had to indicate that they could detect a difference between A and B by selecting the one they believed to sound most like reference X.

- In all other trials, no two stimuli were identical. Listeners still had to indicate that they could detect a difference between A and B by selecting the one they believed to sound most like reference X; the crucial difference in these trials was that there was no predefined correct answer: the direction of listener responses was not predicted in advance, so a difference in favour of A *or* B was valid. This procedure shares similarities with both the constant-reference Duo-trio and PC techniques, and may be described as 'Paired Comparison with Reference' (PCwR).

From this summary it can be seen why both types of trial were classed as ABX, but featured a fundamental difference. This influenced the statistical treatment of the data returned from each type of trial, as discussed in sections 6.4.4 and 7.2.3. However, the necessity to include a reference in the PC trials meant that the different techniques were identical from a participant's viewpoint. They were not aware that they were performing two different listening tasks. Also, the instruction was the same even when the task involved two identical loudspeakers and one different: identify the one sounding *most similar* to X. This allowed the hidden reference trials to be incorporated into the primary comparisons of loudspeaker models without interruption or bias; listeners were given no reason to view any given trial as different to the next. As such, there was no reason for them to speculate on how or why the two parts of the experiment differed, and whether they should try to modify their judgement strategy accordingly. An additional feature of the method not yet covered is the nature of the differences that listeners were told to assess. They were given no advice regarding the audible attributes upon which they should make their judgements. This is an acceptable approach according to the formal PC guidelines as long as the advice is consistent across all trials [147, sec. 5.8]. It was a concern that requesting participants to focus on bass content of the musical reproductions would bias the responses; it may have been advantageous in focussing attention on the relevant aspect of the samples, therefore potentially increasing discrimination of fine differences, but there was a risk that this instruction would be interpreted simply as 'which of A or B has more bass' rather than considering the impression in relation to the reference. The instruction 'which of A or B is most accurate' was not used because it was believed that such a term would be difficult for participants to interpret correctly without extended training, and this was considered impractical (as discussed in section 5.1). The use of a Paired comparison with Reference method was therefore believed to be an appropriate way to evaluate loudspeaker reproduction accuracy at low frequencies without requiring trained listeners and whilst minimising the likelihood that judgements would be solely based on perceived level of bass content.

5.2.2 Response Bias

Systematic errors that affect listening test data are referred to as bias, and must be minimised as they can result in misleading conclusions. A key source of bias in listening tests arises when a participant has to map their 'internal judgement' onto the response scale provided in the experiment [152]. It is desirable, for conceptual and computational reasons, to try to map a subjective impression onto a linear numerical scale. It is difficult to construct such a scale [153], and whilst it is usual to interpret results as if they correspond to an interval or ratio measurement, this relationship should never be assumed without firm justification that this is an appropriate model for the underlying psychometric function [154]. Rating the performance of

objects by awarding numerical scores can be revealing, but is likely to be unreliable if perceived differences are very small or differ greatly between judges [155], as would be expected from inexperienced listeners without considerable training in the task. Bias due to mapping of numerical scores onto a scale may produce errors in the order of 10 to 40 % of its total range [152, 156]. Though the exact error figure will depend on the circumstances of the experiment, it seems that direct scaling methods for auditory evaluation, such as numerical and verbal rating scales or rankings, can give a compressed representation of the underlying perceptual function [157]; this may be due to the use of 'anchor' points or descriptive labels which aim to make listener responses more uniform, but unavoidably constrain the scale [144, 152, 154, 158]. Ranking methods, arranging groups of items according to some specified attribute, avoid any scale-mapping bias and can be fast and effective, but only if differences between items are fairly obvious; if not, there may be a greater amount of uncertainty than in paired presentation [155].

The response method in ABX tests has no inherent scale bias, provided that measures of randomisation and balancing are implemented to minimise the chance that either A or B is favoured for extraneous reasons. It is also possible in some cases to rank stimuli through the use of ABX or paired comparisons, known as indirect scaling; the data is free from the scale-compression imposed through direct methods and can therefore lead to a more realistic measure of auditory attributes [159, 160].

5.2.3 Ease of Participation

It can be a demanding task for listeners to quantify what they perceive, especially when differences between auditory stimuli are small. Inexperienced participants in particular find it difficult to assign ratings reliably, though improvement in accuracy and consistency can be reached with prolonged training in the task and correct use of the scale [161, 162]. The use of a simpler response method therefore reduces the need for extensive practise before the experiment. Ordinal scales (rankings) are generally easier for listeners to work with, and can be quicker than pairwise comparisons if differences between stimuli are clear and the group of objects is small. When differences between objects are hard to detect, a paired comparison or ABX method allows easier discrimination than either scaling or ranking procedures; the distracting effect of other items is removed, and the process is quicker than trying to confidently allocate ranks for objects that are perceived as being very similar [155].

A team of listeners that have been trained and practised over a number of years will allow more complex listening experiments to be conducted, and variability of responses can be low. As this kind of listening panel is rarely available to researchers, it is advantageous to develop an experimental method that is suitable for anyone who is willing and able to take part; this type of participant may be referred to as 'naïve' [163]. Simple discriminative methods such as ABX and paired comparisons are examples of only a few standard techniques that are ideal for naïve participants [164]. Even if the listeners are experienced, overall sound quality is known to be a multidimensional percept [165], and the ABX method is well suited to assessments with complex stimuli of this nature [164].

The cognitive strategy used in an ABX task is generally a very simple 'difference' decision: compare A with X, and B with X, then choose whichever one gives the smallest perceived difference; keeping the cognitive load low is especially important in listening tests as, unlike

visual tasks where two images may be presented side by side, it seems that the unavoidable temporal separation of stimuli in non-monadic tests increases difficulty of the task for participants [164]. This calls for a straightforward experimental method that minimises the amount of information that listeners must receive, process, and remember. The ABX method requires very few actions by the participant, except for switching between stimuli, and the need for long-term memory is minimised if fast switching between stimuli is possible [139, 145].

5.2.4 Duration and Efficiency

One drawback of the ABX method is that the number of comparisons, and therefore test duration, increases rapidly with the number of items to compare. For an experiment with M models to assess, this requires a total of N trials [166]:

$$N = \frac{M(M-1)}{2} \tag{5.1}$$

It can be seen from Eqn. 5.1 that the number of evaluations each listener must perform increases rapidly with the number of stimuli being compared. This demand can be reduced through the use of balanced incomplete block (BIB) arrangements [167], at the cost of increased design complexity. In practice, however, many comparisons can be made in a single session without exceeding reasonable time limits; the duration of a single trial can be very brief because only short audible stimuli are needed, and the simple nature of the task means that listeners can reach a judgement very quickly. The overall testing duration is also reduced because minimal training is required to familiarise participants with the task before the formal experiments.

The other disadvantage of the ABX technique compared to rankings or numerical ratings is that it reveals much less information about what listeners perceive; it merely shows which of two options they choose when presented with a choice. As such, the ABX method is generally considered to be less revealing, and therefore less useful, than direct scaling methods; Ryden [168] clearly described a study where an ABX method was rejected on these grounds. The ABX method is an inefficient test strategy if the following conditions are met:

i) The underlying effect is already well understood and can justifiably be reduced to a linear construct.

ii) The numerical scale has been developed sufficiently well to allow adequate mapping of the full effect magnitudes, thereby avoiding the perceptual compression effects described in section 5.2.2.

iii) Participants are critical listeners that are well trained in the specific grading system to be used during the tests.

None of these conditions were true for the experiments in this project, and so the relative inefficiency of the ABX technique was considered an acceptable compromise for the potential sources of bias and error that it avoids.

5.2.5 Chosen Methodology

The main procedural details for the final method are described here, with brief additional notes to justify their use in this study. Details that are specific to each set of listening tests are given in

chapter 7 before presentation of the results.

5.2.5.1 Environment Listening tests with individual participants were performed inside a large anechoic chamber. The stimuli were short musical extracts simulating playback through a number of different loudspeaker designs, reproduced through a single equalised loudspeaker.

This provided the most sensitive listening conditions, minimising ambient background noise, problems of placement and room interaction, and inter-channel cross-talk. The listener encountered no bias from conferring with other participants and was seated in the optimum listening position.

5.2.5.2 Design The test was an ABX 2AFC design, requiring listeners to say which of two loudspeaker models, presented on channels A and B, sounded most like the reference; this was clearly identified as a reference before commencing testing, and was always presented on channel X. Listeners were not given any guidance on the audible attributes they must assess to make this decision. Some trials contained a hidden reference without the listeners knowing, where one of A or B was the reference itself. All other trials featured different loudspeaker models on channels A and B. In each trial, listeners were allowed to switch freely between the three stimuli using a three-way switchbox.

An ABX method was most suitable for inexperienced participants and required minimal training in the task; rapid and direct comparison aided discrimination of small audible differences between stimuli. Forcing a decision prevented conservative responses in a task that was likely to be difficult for many participants. It was predicted prior to testing that the inefficiency from inclusion of a hidden reference would be compensated for by its usefulness in post-screening data; this is discussed further in section 6.5. The switching mechanism allowed near-instantaneous blind comparison of the three signals, minimising the reliance on acoustic memory and giving no indication regarding channel assignment for any stimulus.

5.2.5.3 Execution The test procedure was run from a bespoke MATLAB program, and controlled by the participant from the listening position using a small touch-screen interface. The listener was allowed to repeat playback as many times as desired; they were instructed to record their answer by pressing buttons on the touch-screen before moving onto the next trial.

Having a user-controlled test removed the need for any contact with the experimenter. Interaction bias was therefore avoided and the experiments were considered as being double-blind. It also allowed the participants to conduct the experiment at their own pace and give a response when they were ready, rather than being forced into responding before they had confidently reached a judgement.

5.2.5.4 Selection and Training Before taking part, participants were asked complete a short questionnaire on their listening experience and report any recent colds, ear infections or exposure to loud sounds; only people reporting significant unilateral hearing impairment were excluded from the tests. The same written and verbal instructions were provided to all participants prior to each session and they were allowed to ask questions about the experimental procedure but not about the nature of the study. All participants performed dummy trials

immediately before the experiment, some identified as practise trials, and some as part of the formal test procedure; the latter were presented to listeners at the start of the formal experiment as brief additional practise in the task without them knowing, and responses from these trials were excluded from the final test results. The musical extracts, models, presentation order and channel assignment of the dummy trials were identical for every participant.

The number of available participants was small, so further limiting this pool though excessive pre-screening was avoided. Also, measuring hearing acuity at very low frequencies is difficult [96], and it was not believed that a typical audiological screening procedure would be useful in selecting suitable participants for this study. Ideal candidates for this task were considered to be those who could listen critically; this is a learned skill rather than a reflection of an individual's hearing threshold. Providing consistent instructions and dummy trials minimised bias due to preferential preparation for the test in some participants but not others; everyone had the same opportunity to familiarise themselves with the task and software interface. Though limited due to time constraints, these dummy trials allowed participants to make mistakes and become accustomed to the procedure without biasing the final data.

5.2.5.5 Presentation Order Channel assignment of each model pair and presentation order of the musical extracts was randomised across all trials and listeners; this was executed using a personalised 'playlist' of test stimuli, generated for each new participant in the test. The playlist contained all possible A/B pair combinations for the models in the test, with complete repetitions by musical extract. The data from any participant failing to evaluate all combinations, e.g. due to leaving the test early, was excluded from the final data set.

Balancing, randomisation, and repetition help to minimise bias due to learning and presentation effects by distributing it equally across all presentations for all listeners [151, 161, 167, 169]. Performing repeated evaluations of the same models across and within listeners helped to prevent occasional spurious responses from skewing the overall cumulative result. The musical extracts themselves were not intended to be experimental variables but enabled all listeners to replicate all pair comparisons a number of times without losing interest in the task. The strategy used for stimulus arrangement is formally known as a randomised complete block design [167, 170]; whilst planning the experiments there was concern that a fully balanced arrangement, as achieved using Latin Squares, would be difficult to perform in practice due to too few participants. Randomisation was therefore used instead as a way to distribute the variability across trials.

5.2.5.6 Duration and Level Duration of the test for any participant was as short as possible and split across two non-concurrent sessions when necessary. These complied with the maximum daily exposure limits set by the university ethics committee. Reproduction SPL was calibrated before the tests and not adjusted. Participants were not able to alter the reproduction level from their listening position but were allowed to leave the experiment if they found it uncomfortable.

Keeping sessions as short as possible reduced the chance of listener fatigue and boredom with the task, and allowed more sessions to be carried out within the testing period. Fixing the SPL across all sessions was an essential requirement due to the increase in threshold of hearing at low

frequencies [96]; reducing the level for some participants and not others could significantly bias the results if signal content in the frequency region being evaluated was allowed to drop below the minimum audible threshold.

5.2.6 Assumptions in the Chosen Procedure

The listening test procedure required several key assumptions; these are explicitly stated here as they were of fundamental importance when analysing the data and drawing conclusions from the results:

- The switching mechanism has no audible effect itself, or at least has an identical impact across all three channels such that it does not bias listener selection in favour of either A or B. This is discussed further in section 5.5.1.
- Trials are statistically independent, i.e. the outcome of one trial does not influence the outcome of any other; this effect was minimised through presentation management, as described in section 5.2.5.5.
- The instruction to select whichever of A or B sounds most like the stimulus on channel X is synonymous with higher fidelity. This is based on the additional assumption that the reference is able to reproduce low-frequency content more accurately than any other model within the experimental group under test. As explained in sections 4.2.3, 4.3.3, and 4.5.4, this was believed to be a valid assumption.

5.3 Music Selection and Processing

The selection and processing of musical extracts was an important stage in preparing for the listening tests. The following subsections describe how a suitable set of signals were chosen and modified before being convolved with the experimental loudspeaker models.

5.3.1 Selection and Extraction

Critical, or revealing, programme material for a listening test must be chosen according to the aims and requirements of an individual experiment. Selection of appropriate material is usually time-consuming but it is known that poor choice of test excerpts can produce misleading or incorrect conclusions [144]. Programme dependence is well established, and is the effect whereby a device is excited in different ways due to the varying temporal and spectral nature of the input signal [153, 171]; in terms of listening tests, this means that a loudspeaker might elicit different judgements from a listener depending on the content it is reproducing. As mentioned in section 5.2.5.5, musical extract was not intended to be an experimental variable; it was assumed that a listener would reach the same judgement about a given loudspeaker model over repeated evaluations using different musical extracts. In practice, this was not a critical assumption, because the data was tested for programme dependence effects before the final analysis; however, an effort was made to select musical stimuli with similar characteristics, increasing the chance that programme dependence would not affect the cumulative responses, i.e. the replication of comparisons by extract could be treated as repetitions rather than separate experimental blocks (details of this analysis are given in section 6.3).

A shortlist of more than 50 extracts was created after auditioning many tracks from a wide range of genres. The initial selection was performed by listening through headphones connected to a computer sound card, and inspection of spectrograms in Adobe Audition. Final selection was performed in the experimental environment; this was found to be necessary for two reasons:

- The perceived level of bass was found to be greatly exaggerated, even in a fairly well-damped listening room, due to the added reverberation from wall and floor reflections.
- Listening through a large monitor at high SPLs in a very quiet environment revealed audible components that had not been heard when auditioning under less critical conditions; a surprising number of recordings appeared to have artefacts at very low frequencies that were distracting and clearly not part of the musical presentation. Fincham [16] reported the same problem when selecting music for investigations at low frequencies and suggested it might be due to the material being produced on monitors that were less revealing than those used for experimental reproduction.

Tracks for audition were extracted from the right channel of the original stereo .wav files, and saved in the same CD-quality format (16-bit, 44.1 kHz). This file quality was maintained throughout all further processing and playback.

Extracts from the shortlist of potential material were selected in the preparation for each separate set of experiments. The extracts were chosen according to several criteria and were believed to be critical material, i.e. revealing of differences between the experimental loudspeaker models. The use of a very restricted sample of extracts can lead to a fixed bias in results if the material has not been optimally chosen [112]. From the subjective results of listening test I (discussed in chapter 7) it was difficult to confirm whether the chosen extracts were the most critical material for the effects being investigated. Therefore, new extracts were selected for listening test II. A different approach was taken in listening test III, where it was of interest to compare extracts with differing characteristics (described in section 5.3.3.3), so again, new material was selected. This use of different material minimised the risk of bias being carried over from one experiment to the next due to poor selection of programme material; the disadvantage was that it introduced a confounding factor, when comparing the results of test I and II in particular: differences may have been due to differences between the loudspeakers or differences in revealing characteristics of the music. This was accepted as it was believed that:

- Testing with a slightly larger sample of extracts would make conclusions common across experiments more widely applicable than if based on an extremely restricted sample of material.
- Material for listening test II was selected according to the same criteria and process and was therefore believed to be at least as revealing as that chosen for listening test I.
- Comparison across experiments was less important than comparison of results within an experiment. Therefore, minimising any carryover bias due to the use of common extracts was considered a priority.

Potential extracts were primarily judged according to three criteria:

Content Adequate low-frequency content was a prerequisite for any track being considered as a potential test extract; differences in LF alignment between the loudspeaker models could

only be assessed if they had content to reproduce. The relevant instruments also needed to be well balanced within the overall mix so that they could be heard sufficiently and not masked by other parts of the arrangement. Excessive bass content was not an essential feature of the music selection process. Examples that were representative of commercially available music were deemed most appropriate to the research application of general studio monitoring. Results from other studies looking at the frequency content of such recordings generally reflected the findings when looking for material in this project. Fielder and Benjamin [12] were interested in the extreme low-frequency limits of music on CD; they found that some recordings feature sounds as low as 12 Hz, but most typical content is above 30 Hz. Chapman [172] analysed 400 randomly sampled CD tracks and found symphonic and classical music to contain almost entirely mid-band energy, with pop, jazz, blues and folk containing evenly distributed energy throughout the audio range. Hip-hop was found to feature a lot of LF energy, the average power spectrum peaking in the 63 Hz third-octave band. More recently, Francombe *et al.* [173] used a method to randomly sample 200 tracks from radio stations; they concluded that contemporary music featured higher levels of bass, especially around 50 Hz. In summary, few recordings in the present study were found to contain large amounts of low-frequency energy below 30 Hz, and instances were generally associated with electronic music or 'effect' type sound events; contemporary music, especially hip-hop and R&B/urban genres had appreciable LF energy in the range 40–60 Hz, making them more suitable than most classical recordings. These characteristics are reflected in the selected extracts described in section 5.3.3; they are all relatively modern recordings, being released between 1996 and 2006, featuring primarily pop and hip-hop genres with no orchestral or chamber music. They all show peak power at or above 40 Hz but were deemed to contain sufficient energy below this frequency to make them suitably revealing extracts in this experiment whilst also being representative of typical tracks that a mix engineer might be presented with.

Stationarity A reasonably homogeneous presentation throughout an extract's full duration was required. Extracts were not analysed for stationarity in the formal sense, but were checked to ensure that they did not contain any large deviations in amplitude, frequency content or tempo. Particularly revealing audible events, such as a bass run or drum fill, are brief, and are difficult to evaluate when switching between multiple stimuli [105]. Zielinski *et al.* [152] described the systematic error that can affect results if extracts do not have similar characteristics across their duration, requiring listeners to perform a sort of mental averaging when making their judgements. Neuendorf and Nagel [174, 175] referred to this requirement as 'perceptual stationarity'; they tested the effects of shortening extracts drastically to increase it, and whether the temporal location of significant audible events affected judgements. They found that despite the increased stationarity of very short musical excerpts (<5 s), listeners found comparisons easier when the duration was longer (≈19 s). They also demonstrated that a significant event occurring at the start of an extract leads to a different rating than if it occurs at the end. When selecting the extracts described in section 5.3.3, an effort was made to choose passages without any such prominent events. The need for musical sections without large temporal or spectral variations limited the maximum length of an extract for most types of music as there are

rarely long stationary periods; this was not found to be a major problem as the decision to use a discrimination-based test method meant that only short excerpts between 20 and 25 s were required. It was found easiest to select extracts of the desired stationarity and required length by using sections from the verses of songs. The chosen segments featured the same rhythmical pattern and no large deviations in loudness or spectral balance throughout their duration.

Quality Any potential extracts had to be of sufficient quality such that hiss, crackles or distortion were not audible under experimental listening conditions. The presence of any such non-musical artefacts could jeopardise the perceptual stationarity of extracts as well as the ecological validity of the presentation: it was assumed that engineers mixing modern recordings will not usually work with material that features these types of distracting audible event, although it was accepted that some variation in quality is typical when considering a wide range of recordings [112]. Whilst auditioning potential material it was found that this requirement excluded many digital transfers of older recordings that had not been remastered first to remove audible artefacts due to the original analogue medium. Tracks that appeared to have been poorly mixed were also excluded; this was usually perceived as a very apparent spectral bias, making the recording either sound muddy and lacking in clarity, or particularly bright and harsh, being very uncomfortable to listen to at the kind of SPLs that were used for experimentation. The selected extracts were all judged to be free from non-musical artefacts and sounded well balanced when reproduced in the experimental environment at the required levels.

Style and genre were not key factors in determining selection, but anything deemed to be especially evocative, offensive, or well known was rejected. This was to reduce the chance that listeners would already be overly familiar with, or distracted by, the extract such that it affected their evaluations of the virtual loudspeakers being compared [144, 173]. Arrangement complexity was also another determining factor but firm restrictions were not imposed; solo instrumental passages were not selected but any arrangement providing appreciable spectral content in the bass regions was considered.

All chosen extracts within a given experimental set were trimmed to have exactly the same length:

$$L = d\, f_\mathrm{s} \tag{5.2}$$

where: L is the signal length in samples, d is the extract duration in seconds, and f_s is the sample rate, 44.1 kHz. It can be seen from the details presented in section 5.3.3 that the extracts were shortened with each experiment. The typical duration of extracts in discrimination tasks is between 10 and 25 s [144]. The upper limit of this duration was selected for listening test I; sufficiently stationary extracts were found, and it was believed that this length would give even inexperienced listeners sufficient time to make a decision without needing to repeat playback. Informal discussion with participants after listening test I indicated that this duration was more than sufficient. Therefore, the extract duration was reduced to 23 s for listening test II, and again to 20 s for listening test III.

5.3.2 Loudness Matching

Perceived loudness of stimuli is known to be an influencing factor in listener judgements of sound quality [165]; this was not intended to be a variable in the listening tests but frequency-dependent variations existing between models due to their different alignments were preserved. The loudness was matched across musical extracts within each experimental set. This was performed on the source files, i.e. before convolution with the partial-system impulse responses as described in section 3.3.4; in this way, significant loudness variations occurring across the range of source material were removed, but those arising from the differences in LF alignment between models in a given experimental group were preserved. The process is summarised below:

- Each experimental group used n unprocessed musical extracts, $x_n(t)$.
- The mean-square value for each extract was calculated: $s_n = \overline{x_n^2}(t)$.
- The mean-square amplitudes were matched by scaling each extract with a gain factor, G_n, determined by the extract with highest mean-square value, $s_{\max}$:

$$G_n = \frac{s_{\max}}{s_n} \tag{5.3}$$

$$m_n(t) = x_n(t)\, G_n \tag{5.4}$$

  where: $m_n(t)$ is the mean-square-scaled version of extract n. Note that $G_{\max} = 1$, i.e. one extract in the group was left unaltered, and the others were amplified or attenuated to match its mean-square value.

- The group of mean-square-matched extracts were scaled by a common factor G_{o} to maximise SNR in the digital domain without clipping; this ensured that:

$$\max |m_n(t)\, G_{\mathrm{o}}| \leq 0.99 \tag{5.5}$$

  All extracts were then saved as CD-quality .wav files.

After processing, the gain-scaled extracts were auditioned under experimental conditions to ensure that no major loudness differences still existed. The method of amplitude matching was found to be effective in reducing overall loudness differences across the extracts; any remaining variations were removed by subjectively adjusting the levels during playback until all extracts in the experimental group were considered to be overall equally loud. The extracts in question were then given an additional gain adjustment in MATLAB and saved for further use; thus, they required no manual gain adjustment during experimentation. The final step in gain processing was a 0.5 s fade in and out to prevent sudden loudspeaker transients during playback.

5.3.3 Selected Extracts

The extracts chosen for each set of listening experiments are listed here with a brief description of the features that made them appropriate source material. Also shown are short excerpts from the time histories and plots of the power spectra up to 1k Hz; the relative levels between spectra for each group have been preserved, but scaled so that the peak value is 0 dB to make comparison across groups easier. With reference to the discussion about perceptual stationarity in section

5.3.1, the spectrograms of the chosen extracts are presented in Appendix I to illustrate that their spectral characteristics did not vary appreciably with time except as part of their individual rhythmical phrases.

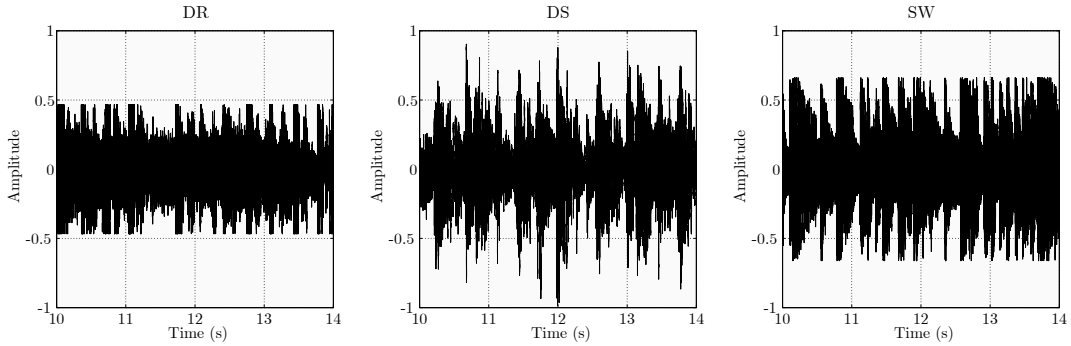
It can be seen from the time history plots that some of these extracts have undergone dynamic range compression, but this was not introduced through processing before experimentation. As described in section 5.3.1, it was considered appropriate to select extracts representative of commercially available music, and this form of compression is a feature of many modern recordings. Deruty and Tardieu [176] presented an historical review of the progression towards excessive dynamic compression, characterising what is known as the 'loudness war'. Whilst this is commonly believed to be detrimental to the perceived quality of modern music, there is evidence that listeners are relatively insensitive to this form of processing unless it is extreme enough to cause audible distortion [177]. All extracts used in this study were auditioned carefully many times before being used in experiments to ensure that any samples with apparent dynamic range compression did not clip or have any unwanted artefacts indicating that distortion was present. The digital waveforms were also inspected in the time domain to ensure that clipping was not present on the original file or introduced through processing for experimentation.

5.3.3.1 Extracts for Listening Test I

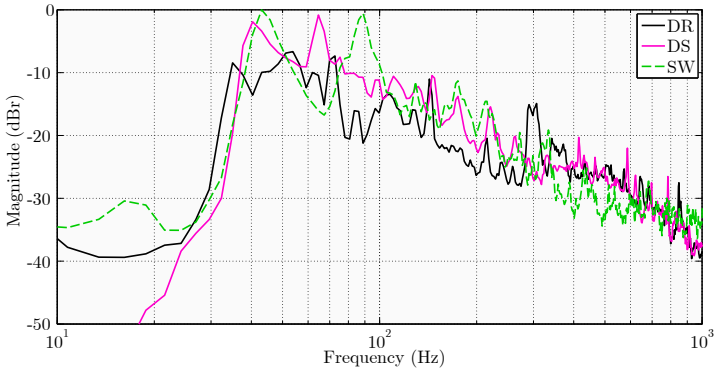
Table 7 describes the extracts used in listening test I, where the experimental models in Group I were evaluated. Characteristics of the extracts are illustrated in Figure 51.

Track	Artist	Album	Comments
Murder Was the Case (DR)	Snoop Doggy Dogg	The Chronicle: Best of the Works (Dr Dre)	Rhythm section is a powerful kick drum and a deep, resonant, synthesised bassline; both remain clear in the mix. Particularly strong in bass energy between 35 and 75 Hz.
L'Y10 Bordeaux (DS)	Daniele Silvestri	Prima di Essere un Uomo	A simple arrangement with a pronounced bass sound that is full and deep, making it seem poorly defined when played through less revealing systems. Shows spectral peaks at 40 and 75 Hz, giving a bass-heavy feel that is balanced by the upper-frequency content.
Dip and Get Low (Deekline & Wizard Remix) (SW)	Stanton Warriors feat. Rodney P.	FabricLive. 30	Complex electronic dance arrangement; has resonant, but clear, visceral synthesised low bass; combined with a punchy mid bass, this track sounds full also but also rhythmically tight. Has most energy around 40 and 90 Hz.

Table 7: Description of listening test I extracts



(a) Time histories (4 s out of 25 s extracts)



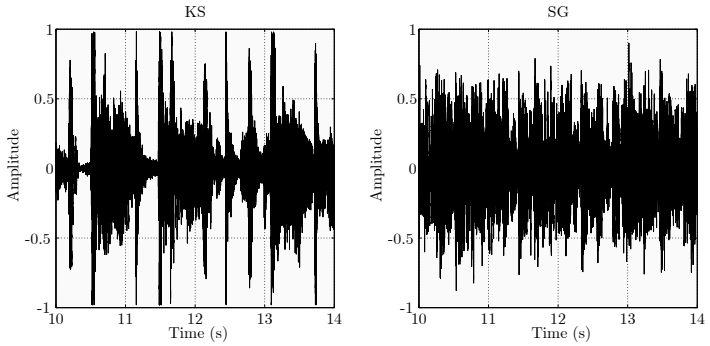
(b) Power spectra (averaged across full duration of 25 s)

Figure 51: Listening test I extract characteristics

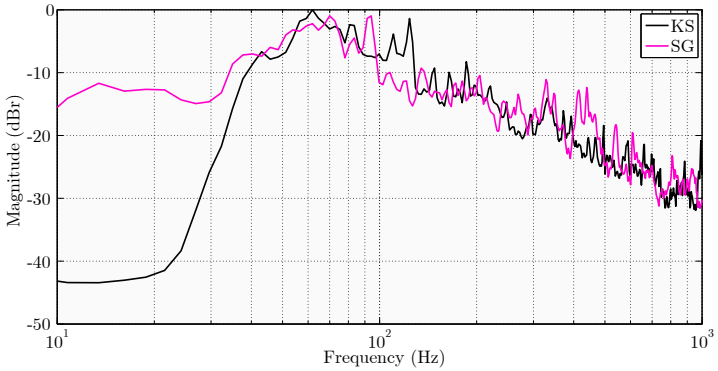
5.3.3.2 Extracts for Listening Test II Table 8 describes the extracts used in listening test II, where the experimental models in Group II were evaluated. The extracts are illustrated in Figure 52.

Track	Artist	Album	Comments
Things Ain't the Same (KS)	Kasino	Blade (soundtrack)	Slow-paced track with a simple arrangement and excellent instrument separation. Has a rounded bass sound that illustrates bass fullness well, and a prominent, tight kick drum; this rhythm section has a tendency to become muddy and overpowering if not reproduced accurately. Spectrum peaks around 60 Hz but has consistently high levels of bass energy down to 40 Hz.
Rundedance (SG)	Rudeboy	101% Speed Garage Anthems	Fast, resonant synthesised bassline that drives the track but does not overpower it. Accompanied by a powerful kick drum sound, this track gives a strong feeling of full bass even when reproduced at moderate SPLs. Unusually flat spectrum down to VLFs with no prominent dips or peaks, but strongest in bass energy between 35 and 100 Hz.

Table 8: Description of listening test II extracts



(a) Time histories (4 s of 23 s extracts)



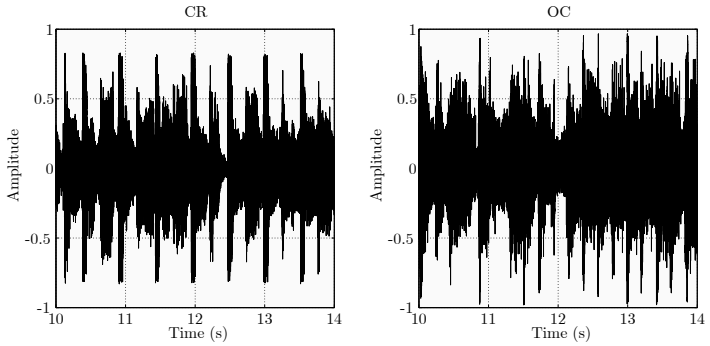
(b) Power spectra (averaged across full duration of 23 s)

Figure 52: Listening test II extract characteristics

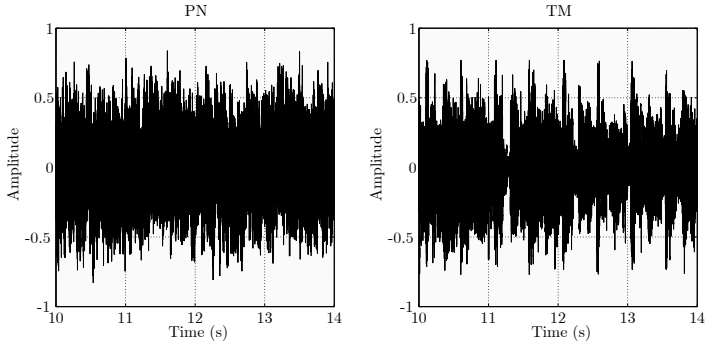
5.3.3.3 Extracts for Listening Test III Selection of extracts for listening test III was slightly different from those used in tests I and II; in these experiments, extracts with a variety of spectral and temporal characteristics were chosen. This was to investigate whether the nature of the signal being reproduced would affect perception of differences between the models in Group III. Therefore, it was expected that extract might need to be treated as an additional experimental variable, but the nature and extent of the effect was unknown prior to testing. Table 9 describes the extracts; they are illustrated in Figure 53. Note that these extracts are separated into two groups, $S_{1,2}$ and $S_{3,4}$. This is discussed further in section 7.4.

Track	Artist	Album	Comments
Pink noise (PN)	—	—	Generated in Audition.
Try My Love (TM)	Brand New Heavies	Trunk Funk - The Best of the Brand New Heavies	Balanced mix of a fairly complex arrangement. Has a very clear, tight and punchy bassline; particularly strong in energy between 45 and 70 Hz.
Crush (CR)	Jennifer Paige	Very Best of the 90's	Clear recording with slightly resonant but pronounced bassline and strong kick drum. Spectrum peaks at 70 Hz with another prominent peak around 150 Hz but the overall mix still sounds well balanced.
40 Past Midnight (OC)	Ocean Colour Scene	Moseley shoals	A steady-tempo track with a simple instrumental arrangement. Has a clear rhythm section and bassline that is tight but has a slight resonant quality. Well-balanced frequency distribution with no regions that are especially pronounced.

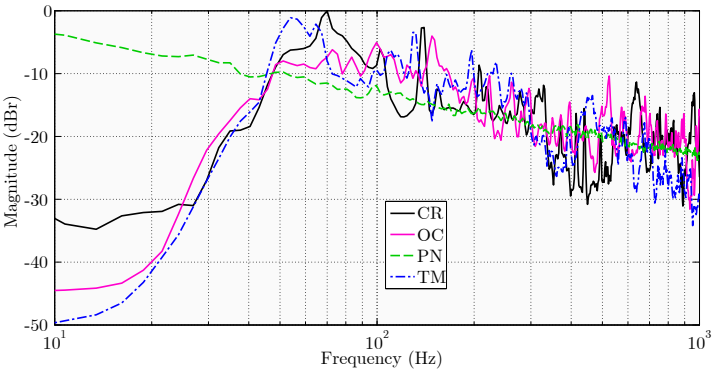
Table 9: Description of listening test III extracts



(a) Extract group $S_{1,2}$: Time histories (4 s of 20 s extracts)



(b) Extract group $S_{3,4}$: Time histories (4 s of 20 s extracts)



(c) Power spectra (averaged across full duration of 20 s)

Figure 53: Listening test III extract characteristics, separated into two groups

5.4 Software Implementation

More details of the listening test program referred to in section 5.2.5.3 are given here. It was specifically developed for this project, and was designed to accommodate simple modification for each listening test; although the overall procedure was consistent, parameters such as the number of extracts and participants were specific to each set of experiments. The two functions of the program are considered separately here: the algorithm for generating individual user playlists, and the graphical user interface (GUI) that listeners used to conduct the experiments.

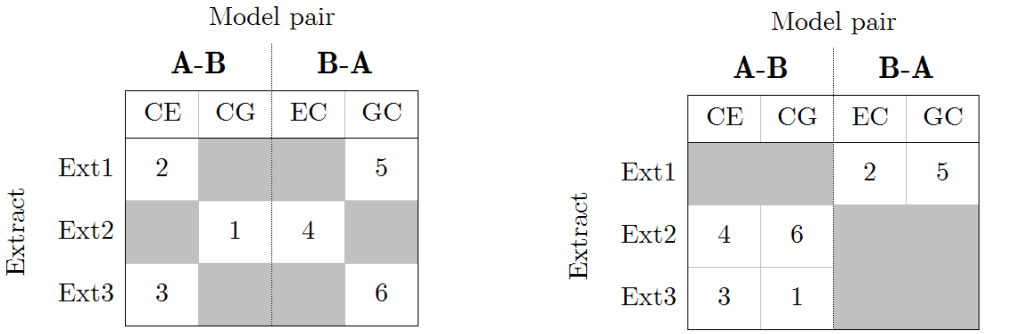
5.4.1 Generating User Playlists

User playlists were generated by randomly populating a matrix where each element corresponded to a specific presentation of extract, loudspeaker pair, and channel assignment. The MATLAB command `rand` was used to generate a pseudorandom sequence of numbers representing elements in the matrix, each time initialising the state input using the current value of the computer clock. This ensured that a different random sequence was generated for each new playlist, i.e. for each different participant; if the state is not modified in this way, it resets to the same initial value and the same random sequences will be produced every time MATLAB restarts. The algorithm was also designed to avoid consecutive replay of any extract, thereby reducing learning effects due to a listener repeatedly hearing any single stimulus.

Figure 54 shows two example playback matrices generated for three musical extracts and two loudspeaker pairs (CE and CG), where numbers populating the matrix represent listening test trial number. The extracts and models have been arbitrarily selected in this illustration. Only two pairs are shown for clarity, but for experimentation the playback matrix contained all A/B and B/A combinations of all models; in listening tests I and II this produced a 30-column matrix. Fig. 54 illustrates the key features of the playlist-generating function:

- No extract is presented twice in succession
  - Consecutive trial numbers never feature on the same row.
- All extracts are presented an equal number of times
  - Each row has the same total number of trials (elements containing numbers).
- Each pair is evaluated an equal number of times
  - Each column has the same total number of trials.
- Channel assignment is distributed across A/B and B/A presentations
  - The right half of columns are populated as well as the left.

These features were designed into the function to minimise the sources of bias described in section 5.2.5.5. Note that Fig. 54a shows equal distribution of channel assignments, first A/B, then B/A; this is the preferred arrangement, but as demonstrated by Fig. 54b, this was not always the case due to the program's random allocation of column elements within the matrix.



(a) Example with balanced channel assignment

(b) Example with unbalanced channel assignment

Figure 54: Example playback matrices. Numbers populating the matrix show the order of listening test trials, each with a specific combination of musical extract, loudspeaker pair, and channel assignment

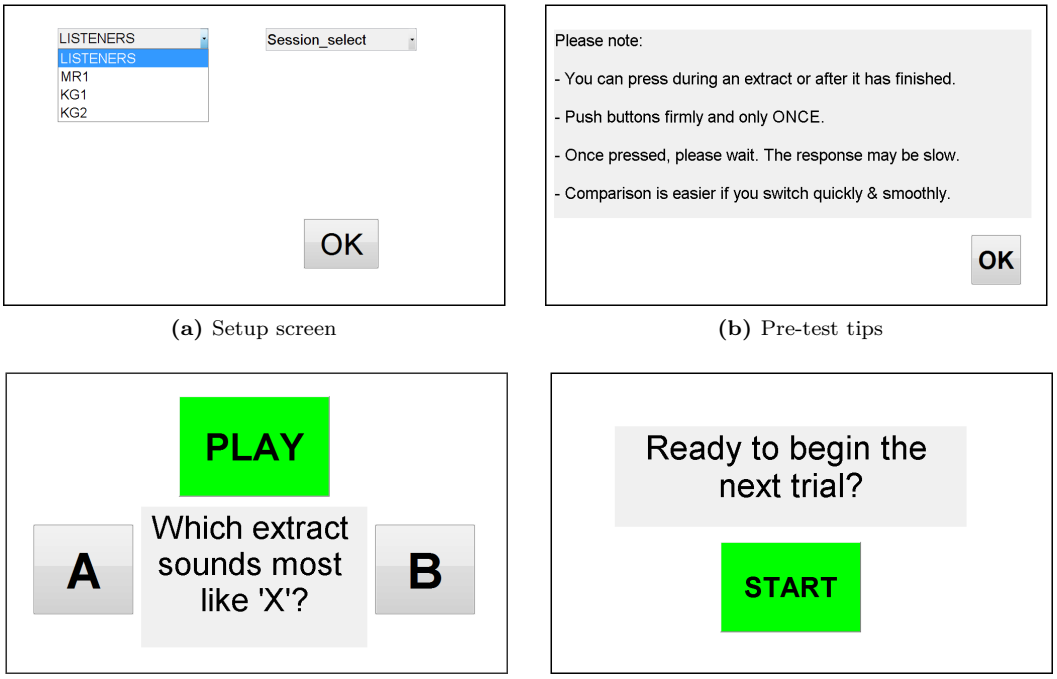
The output of the playlist-generating function was created by converting the matrix elements into a 2-by-N array where N is the total number of trials within a listening session. Each row in the array therefore defined extract, pair, and channel assignment for a given trial, identified by the row number. For example, trial 4 in Fig 54a is element $a_{2,3}$; thus, the corresponding playlist entry would show that trial 4 should use extract 2, loudspeaker models C and E, with E assigned to channel B.

Every playlist generated by this function was automatically assigned a unique identification tag that was used to recall it immediately before starting a listening session. A separate program then read the playlist array to select the correct files and assign the appropriate loudspeaker models to the relevant channels during each trial; this program is described further in section 5.4.2.

5.4.2 Experimental Interface

A major part of the testing program was the GUI that participants interacted with and use to control the experiment. Such interfaces can quickly become cluttered and complicated in multiple-stimulus listening tests [152], increasing the amount of information that a participant must receive, process, and remember. Use of an ABX method meant that a very simple user interface could be designed, minimising the cognitive load and allowing participants to concentrate on listening rather than interacting with the test software.

The interface was designed to be easy to use, having large buttons and text. Neutral colours and a symmetric layout ensured that there was no implicit suggestion that either choice (A or B) should be selected over the other. Figure 55 shows example screenshots from the key screens within the listening test GUI. Fig. 55a shows the screen used by the experimenter to select a playlist before a session; they then left the test environment and had no further interaction with the GUI. The main test screen is shown in Fig. 55c. Participants were informed before the test that they could not progress to the next trial without evaluating the presentations in the current one; the program disabled the response buttons until 'PLAY' had been pressed. This was intended to reduce bias in the final data set if participants started 'skipping' through trials and randomly selecting A or B without listening to the stimuli.



(a) Setup screen

(b) Pre-test tips

(c) Playback and response

(d) Inter-trial screen

Figure 55: Example screenshots from the listening test GUI

Once a listening session had started, the playback program read from the selected playlist, loaded the corresponding audio files (one per channel), and simultaneously sent them to outputs on the designated channels, A, B, and X, via the DACs and switchbox. When a participant registered their answer, the response was written to a results database. This process automatically repeated until all trials had been completed. Automating the procedure in this way removed the need for any interaction between participant and experimenter. In any given trial, neither party knew which stimulus was being evaluated; the tests were therefore performed under double-blind conditions. Automatic electronic storage of the results also saved a substantial amount of time and removed the risk of errors due to manually transcribing a large amount of data.

5.5 Execution: Hardware and Practical Matters

The equipment setup was virtually identical to that shown in Figure 28 for the initial impulse response measurement of the experimental loudspeaker, except that the microphone and preamp were removed. This section briefly discusses two elements of the experimental setup in more detail: assessment of the three-way switchbox, and level calibration of the musical extracts.

5.5.1 Switchbox Testing

As mentioned in section 5.2.6, it was crucially important that no bias was introduced by the method of switching between different models in the ABX presentations. Figure 56 shows the physical switchbox created for the experiments. The channels were spaced equally, with participants having to switch through the reference in the centre position to alternate between A

and B. It was located on the right-hand side of the listener, as shown by point (4) in Fig. 28, section 3.3.1.1. A manual method of switching was used to allow rapid comparison between channels, thereby minimising the reliance on acoustic memory as discussed in section 5.2.3. The touch-screen display used at the time of testing was a relatively slow interface to the desktop computer located away from the listening position; pressing buttons on the interface was followed by a delay of approximately 2–3 seconds due to the speed of the interface and execution of MATLAB commands. It was therefore a concern that such a delay would cause frustration to the participant and make comparisons difficult. Use of a current tablet computer capable of executing the test program directly would remove the need for a manual switchbox, especially given the more advanced audio-handling capabilities of the latest MATLAB versions[††].

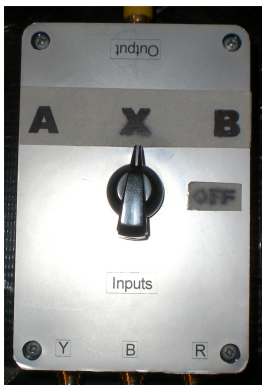


Figure 56: Three-way switchbox

Four ways were identified in which the switchbox might modify the audio signals:

i) Amplitude distortion between channels.

ii) Crosstalk.

iii) Spectral distortion.

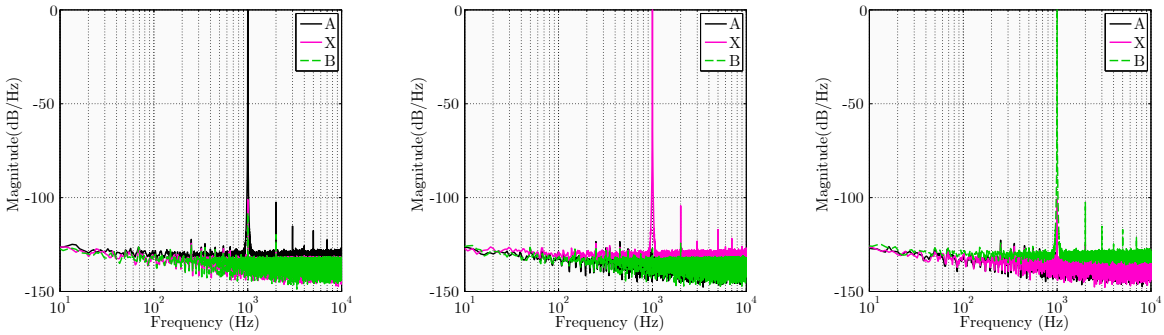
iv) Channel noise.

Before the measurements, a 1 kHz calibration tone was set at a level representative of that passing through the switchbox during experimental playback. The tone was played through the switchbox at the same level, one channel at a time, with all channels connected. The power spectrum for each separate channel recording was analysed to look for differences between channels that should have been presenting identical reproductions of the input signal.

Results of the switchbox spectral analysis are shown in Figure 57. All spectra in the plots have been normalised to the peak signal level so that the tones peak at 0 dB, making it easier to see the level of equivalent noise floor on each channel. In each plot, the active channel (through which the test tone was played) can be identified by the large peak at 1 kHz. In practice, only one channel was ever available to the listener at any time, but all three channels were active, i.e.

[††] *Audio Support from DSP System Toolbox*:
http://uk.mathworks.com/hardware-support/audio-dst.html?s_tid=srchtitle

signals sent to all three channels simultaneously so that the user could switch between them at will without any delay other than that of the switching instant; it was therefore important to check for evidence of crosstalk, the pickup of a signal on one channel onto another. For this reason, plots of the two 'quiet' channels during playback are presented for comparison against the active channel in each plot of Fig. 57.



(a) Signal on A **(b)** Signal on X **(c)** Signal on B

Figure 57: Switchbox analysis spectra, calculated from 6 s recordings using 2^{15}-point transforms and Hann-windowed segments with 50 % overlap. Results have been normalised to peak signal level on the active channel

It was concluded that the switchbox did not introduce appreciable distortion to the audio signals. Specific notes on each type of distortion are given below.

Amplitude distortion Tone peaks without 0 dB scaling (dBr to 3 d.p.):
$A = -9.588$; $X = -9.621$; $B = -9.606$;
The maximum amplitude difference between any two channels was 0.033 dB, considered negligible, and inaudible when switching between channels.

Crosstalk Presentation on channel B showed most evidence of crosstalk, with some pickup on channel X; this was 100.9 dB below the peak level of the main signal and therefore considered negligible and found to be inaudible, i.e. no signal could be heard on the inactive channels during playback of the signal tone.

Spectral distortion The tone peaks at 999.9 Hz on all channels; the difference from 1 kHz of 0.1 Hz is within the frequency resolution of the analysis and therefore considered accurate. All channels showed some harmonic components of the 1 kHz tone, but further analysis confirmed that these were present in the input signal and were not evidence of spectral distortion by the switchbox.

Channel noise The minimum SNR on any channel was found to be 93.21 dB (to 2 d.p., averaged across 6 s segments from the time histories) and therefore considered satisfactory; no noise was audible on channels during playback, regardless of whether or not a signal was present, i.e. it was inaudible even on 'quiet' channels.

5.5.2 Calibration of Music Levels

Prior to conducting the listening tests, the levels of all processed musical extracts were measured at the listening position with an integrating sound level meter to ensure that the university safety and ethics levels for human experimentation were not exceeded; the SPL limit was 85 dB(A), provided that no individual was exposed for more than 60 minutes within 24 hours [178]. The power amplifier level was set such that the extract with the highest measured SPL at the listening position met this requirement.

After conducting these checks and calibrating the playback level, the system gain settings were not adjusted further. The SPL was routinely monitored throughout each period of testing, usually at the start of each day, to ensure that the gain was still correctly calibrated, primarily a check that the power amplifier output level had not been adjusted. Table 10 lists the playback levels for extracts in each experimental group, measured at the listening position and averaged across the full duration; extracts with the highest levels of low-frequency energy were used for level calibration, i.e. the extracts within each group were replayed through their corresponding reference virtual loudspeaker. The extract labels as defined in section 5.3.3 have been used here for brevity. The A-weighted levels have been shown first as these had to be monitored to ensure compliance with the safety and ethics requirements. The equivalent Z-weighted levels have also been shown for interest; these are the unweighted SPLs [179]. The sound level meter used to measure them was not available for the final set of experiments so these values are missing from Table 10.

Experiment	Extract	SPL (dB LAeq)	SPL (dB LZeq)
I	DS	74.3	87.1
	DR	75.2	83.6
	SW	75.7	87.6
II	SG	78.3	86.9
	KS	76.7	87.1
III	CR	76.0	*Not measured*
	TM	73.5	
	OC	75.5	
	PK	75.5	

Table 10: Extract playback SPLs; values shown to 1 d.p.

5.6 Summary of Listening Test Design

A listening test strategy was developed in order to assess whether the MTF algorithm was useful in predicting subjective impression of bass reproduction accuracy. Before developing the strategy in detail, a number of aims and constraints for subjective experimentation were defined. Acoustic presentation should be representative of real mix-monitoring conditions whilst ensuring that all possible sources of bias were minimised. Highly sensitive listening conditions were required to allow detection of differences in LF alignment between the experimental models; a

discrimination-based method was the most effective way to make the task appropriate for all potential participants, and would be most likely to allow fine discrimination between audible stimuli. The chosen test method had need to return data at least at the ordinal level so that the direction of any observed effects could be compared with the corresponding MTF results. This level of data was considered sufficient for this stage of investigation; if a correlation between algorithm and listener results could be established, it would justify further investigation into the technique, including the considerable extra complexity of trying to map numerical subjective ratings to the MTF scores.

After consideration of other popular listening test strategies, the ABX method was identified as being most suitable for experimentation in this study. Although it was acknowledged that the technique may be considered inefficient, it was found to offer considerable advantages over direct scaling or rankings. The format was strictly defined after considering the possible standardised variations from the field of sensory testing. A fixed reference, X, was always identified to participants when presented with two alternative choices, A and B, but two types of trial were actually used. They were both based on the standardised fixed-reference Duo-trio method but only the hidden reference trials were 'true' ABX according to this definition; in other trials A and B were both different from reference X, described as 'Paired Comparison with Reference'. In all trials, listeners had to state which of A or B sounded most like X, based on any criteria; they were not aware that they were performing two different listening tasks where one was a direct comparison, and the other involved some judgement. The comparison against a reference assumed to be of greater reproduction accuracy than any other model obviated the need for potentially ambiguous terms such as 'better than' or 'more accurate than' when comparing A and B, and removed the need for extended training in being able to identify a more accurate reproduction without the presence of a reference for direct comparison. It was inferred that selection of a model overall more like the reference was equivalent to selecting the model with greater accuracy of reproduction.

Following clear definition of the chosen experimental strategy, a number of key design features were considered to develop the final test method. These factors included randomisation and balancing of test stimulus presentation, minimising interaction between participant and experimenter, and management of session duration to prevent listener fatigue. Identifying these features and incorporating ways to control them in the experimental technique increased the likelihood that bias in the final data set was minimised, or at least evenly distributed across all trials.

Selection of appropriate musical extracts was a critical part of listening test preparation. With the experimental method fully defined, programme material with suitable characteristics was selected. Adequate bass content, perceptual stationarity, and recording quality were key criteria upon which potential material was evaluated. It was found that final selection of extracts from a preliminary shortlist could only be performed under near-experimental conditions, i.e. reproduction through a high-quality professional monitor at high SPLs in an anechoic environment; other methods of audition proved to be insufficiently revealing of non-musical artefacts and mixes that sounded thin and unbalanced without additional reinforcement from room reflections. Eight musical extracts were eventually selected for use across three separate sets of listening tests, plus a pink noise extract, the only artificial test stimulus. The chosen extracts were divided into three groups, one for each set of experimental loudspeakers to be

evaluated. Gain adjustment of individual extracts was performed to remove differences in overall loudness within a given group; this provided a more consistent audible presentation to participants throughout a listening session, thereby reducing the likelihood that judgements were biased by inherent loudness differences due to the use of multiple extracts.

A dedicated program was created in MATLAB to enable automated execution of the listening experiments. This incorporated several key aspects of the experimental design and consisted of two parts. The first function of the program was to generate a unique playlist for each participant; a pseudorandom sequence was used to distribute each combination of extract, loudspeaker pair, and channel assignment across all trials within a given listening session. The second part of the program was a graphical user interface (GUI) that participants used to control the test independently. Activation of the GUI controls automatically recalled and assigned the relevant audio files to the correct channels, then stored listener responses to a database. The program was considered a key component of the experimental procedure as it minimised several potential sources of bias and error; it ensured that each listening session featured a unique and randomised presentation of stimuli, and excluded any errors through manual transcription of results. It also removed the need for any interaction between participant and experimenter, therefore ensuring that the tests could be classed as double-blind.

The final preparations for testing involved practical aspects of execution that might be considered further sources of bias or error. Investigation of the three-way analogue switchbox showed no evidence that this introduced audible distortion. Replay levels were calibrated for all extracts at the listening position with the full experimental setup; it was thus ensured that reproduction level was as high as possible without making participants feel uncomfortable, and without exceeding the university regulations for exposure to loud stimuli over the duration of any listening session.

Through consideration of experimental design factors and sources of bias, the final test procedure was considered to be the most appropriate given the aims and constraints initially identified. Statistical techniques used for analysing the data returned by these experiments are described separately in chapter 6.

6 Statistical Methods for Analysis of Subjective Data

Chapter 5 described the way in which carefully designed subjective experiments were performed to obtain listener judgements of the loudspeaker models presented in chapter 4. The methods used to summarise results and quantify the level of confidence in conclusions are now addressed.

It was established during planning of the listening tests that the type of data they returned could be analysed using standard statistical methods based on sources from the social sciences, sensory testing guidelines, and a number of publications in the audio literature. However, it was necessary to develop a framework for analysis that addressed the key experimental questions. These questions and explanation of their relevance to the topic under investigation is presented in section 6.1. Based on these questions, it was possible to select statistical tests that were appropriate for the class of data, and produced meaningful results for interpretation and comparison with findings from separate but similar experiments. Section 6.2 describes how a suitable class of statistical methods were chosen; the specific analysis techniques are presented in detail in sections 6.3 to 6.5, and summarised in section 6.6.

6.1 Primary Aims for Data Analysis

Three questions were of primary interest to the study, and therefore influenced design of the listening tests and the selection of methods for data analysis:

1. *Were results influenced by any particular programme material?*
   As discussed in section 5.3.1, musical extracts were selected according to a number of criteria. The source material was chosen to be revealing of differences in low-frequency behaviour between the loudspeaker models in each experimental group whilst minimising bias towards any model in a given trial, such as due to a bass run or drum fill that would be difficult to compare when switching between them. For listening tests I and II, extracts had been selected with the intention that they would elicit the same judgements about the same loudspeakers. Conversely for listening test III, extracts with dissimilar spectral and temporal characteristics were chosen; they were expected to show some variation in responses from listeners, but the nature of the effect was unknown prior to testing. It was of interest to confirm whether the selection of material had been successful, and whether any types of music made it less difficult to compare the loudspeaker models; any content found to be especially revealing of differences would indicate that this type of music must be mixed on the most accurate monitors. In the experimental context, it was essential to test whether any extract in a given experiment influenced the judgements of listeners, leading them to choose loudspeaker A when, with a different extract, they chose loudspeaker B, or vice versa; the outcome of this test would determine whether programme material had to be treated as an additional variable in further analysis.
   The method chosen to address this question is described in section 6.3.

2. *Were statistically significant differences found between all tested pair combinations?*
   The main purpose of the experiments was to gather subjective data for a range of loudspeakers with differing low-frequency alignments; findings could then be compared against objective results for the same systems to see if the algorithm had predicted the

outcome. This comparison was based on results from the pairwise subjective evaluations of the loudspeaker models. It was believed that differences between any two of the models were audible as listening conditions were very sensitive, but it was not known whether there would be consensus within a group of listeners. When an apparent split in favour of loudspeaker A or B existed in any pair, it was essential to be confident that a real effect was present in the data; pairs which produced insufficient confidence in this conclusion could not be used to address the primary experimental aim, direct comparison of algorithm results with subjective judgements.
The procedure chosen to address this question is described in section 6.4.

3. *Were results any different for the more critical listeners?*
Successful mix monitoring depends on being able to first detect subtle audible differences, then adjust the mix accordingly. The participants in this study were not expected to have the necessary skills to make a judgement on the latter, but it was considered desirable to be able to identify the most critical listeners; these were deemed to be the participants who demonstrated increased acuity and consistency in the experimental task. It was then of interest to see whether judgements from only these listeners showed any better correlation with the algorithm results; if so, it would give increased confidence that the method is well suited to the intended application, evaluation of professional studio monitors. As described in section 5.2.5, it was proposed that performance in hidden reference trials might be a suitable method for post-screening the experimental data; the assumption was that the most accurate listeners in direct auditory comparisons would be more reliable judges in less simplistic evaluations, i.e. when A and B were both different from X. Inspection of the paired loudspeaker judgements before and after post-screening allowed a conclusion about whether this assumption was correct, and then enabled the inspection of whether there was any subsequent increase in correlation between listener judgements and algorithm results.
The method used to perform the post-screening is described in section 6.5.

6.2 Identifying an Appropriate Class of Methods

If an appropriate test is used correctly, statistical methods can be very reliable, both in describing features of a data set and inferring whether an experimental effect will be observed in a population based on a smaller sample. It is up to the experimenter to choose an appropriate test; the choice should depend on a number of things, such as the number of samples, whether they are related, and the objective of the experiment [167], but must primarily be based upon the characteristics of the data. For the ABX procedure described in section 5.2.5, the primary data set would be the number of times each loudspeaker design was chosen over the other, i.e. frequencies, or counts, of the number of times loudspeaker A or B was selected. This type of data is discrete and dichotomous, being fundamentally different from the type of data returned by listening tests employing grading scales or numerical ratings where the implicit assumption is that the data values are continuous and linearly distributed. This distinction leads to the classification of statistical methods as either parametric or nonparametric; the categorisation is not absolute and sometimes contentious [180], but is used here as a way to demonstrate the fundamental criteria that were considered when selecting appropriate methods for analysis of the listening test data.

6.2.1 Classification

Parametric techniques are powerful and efficient, allowing conclusions about the presence or absence of effects in the data to be stated with greater certainty for a given sample size. These methods are based on the assumption that the data are normally distributed, or at least conform to a known distribution [181]. Parametric statistical techniques are often used for data analysis and are referred to in several current audio listening test standards [143, 144, 158]; these all focus on using rating scales for evaluation and therefore provide useful advice regarding parametric methods for statistical analysis. By comparison, nonparametric methods are generally considered to be inefficient and less powerful than their parametric alternatives, though this is not necessarily always true [182]. In general, their reduced power means that for a given sample size, an observed trend in the data must be greater to conclude that it is due to a real experimental effect. However, nonparametric techniques have some considerable advantages. They are not based on any prior assumptions about the data and can therefore be readily applied in most situations, including when the distribution is unknown or known not to be normal. As such, nonparametric methods can be used whenever parametric equivalents would also have been appropriate; Bech [183] demonstrated an example with listening test data. However, the opposite is not true, meaning that misleading conclusions can be drawn if the normality assumption is violated; Raffin [182] showed this clearly with paired-comparison data obtained from speech discrimination tests. Another significant advantage of nonparametric tests is their simplicity. They are comparatively easy to understand and apply without the need for specialist software packages, making them accessible even to researchers with little experience in statistical analysis.

6.2.2 Selection

Detailed information and guidance on selecting an appropriate statistical test is available elsewhere; for example, Kemp *et al.* [167, pp. 26–29] give useful decision charts for researchers conducting sensory experiments that could also be applied to listening tests. In general, parametric methods should be selected whenever they are appropriate for the data. Nonparametric procedures might still be used, but would be favoured mainly when fast and simple preliminary analysis is needed, for example during pilot studies.

For listening tests where numerical ratings or scale gradings are awarded, the data returned will be at the interval or ratio level. It is usually assumed that the responses are sampled from a normally distributed population; if this assumption is shown to be valid, parametric techniques such as ANOVA can be used for analysis. These requirements would not be met for the ordinal data generated by the ABX experiments in this study; nonparametric methods were therefore selected for analysis. The chosen methods are described in sections 6.3 to 6.5.

6.3 Investigating Programme Dependence

The test signals, or stimuli, used in listening tests will typically be extracts of speech, music, tones, or noise. In many studies, the different extracts form part of the experimental hypothesis; in this study, as discussed in section 5.2.5.5, they were used to allow replication of listening trials but were not intended to be an experimental variable. In either case, the researcher should be able to reach a justifiable conclusion about the suitability of the test signals they selected for the

experiment; it must be clear whether or not to include source material as a variable in the data analysis, particularly if this was not the intention.

If the data satisfies the requirements to perform parametric analysis, the confounding effect of test signals may be readily checked using ANOVA. There does not seem to be an equivalent test that is routinely used for ABX or paired comparison listening test data; it was concluded that the chi-square (χ^2) test for independence was a suitable method to address the first question described in section 6.1. The following subsections briefly describe the method and how it may be applied to detection of programme dependence in listening tests.

6.3.1 The Chi-Square (χ^2) Test for Independence

The χ^2 test for independence looks at the difference between categorical variables, i.e. the number of counts in each of two or more categories. The 'observed' frequencies, total counts from the experiment, are arranged by category in a special form of table known as a crosstabulation. For each cell in the table, an 'expected' frequency is also calculated; this is the number of counts in that category that would be expected simply due to chance, i.e. if listeners were voting randomly due to no experimental effect. The null hypothesis, H_0, for this test is that the categorical variables are independent of each other; in other words, the number of times a loudspeaker was selected in a given pair is not related to the programme material used for audition. As such, the observed and expected frequencies in each cell should be very similar, differing only due to the random variations of sampling error. The alternative hypothesis, H_A, is that loudspeaker selection was not independent of extract; from this it is inferred that listener judgements were biased by at least one of the different musical extracts.

The overall magnitude of the differences between observed and expected frequencies is reflected in the value of the χ^2; the larger the differences, the larger the value of χ^2. Referring to a table of critical values for the χ^2 distribution will show the probability of getting a figure of that magnitude due to sampling error alone. From this, it can be concluded whether differences between variables are large enough to have been caused by some association between them or if they are just due to small natural variations e.g. did listeners really tend to select loudspeaker *A* when listening to extract *1* more than extract *2*, or did they select *A* approximately the same number of times per extract. It is important to note that this test does not tell you anything about the nature of the relationship between the variables or how strong it is, only the likelihood that it exists. This was not considered a serious limitation for the investigation of programme dependence in this study; it was sufficient to establish whether or not the effect existed so that a decision could be made about whether musical extract should be classed as an experimental variable.

Some caution must be exercised when using a χ^2 test as it is particularly sensitive to sample size. As a general statistical rule, conclusions are more reliable when drawn from tests on large samples. However, χ^2 will always show a statistically significant result if the sample size is large enough; this can be misleading as the actual differences between samples in a practical sense may be trivial [184]. Conversely, χ^2 can also be unreliable if the tested samples are too small, leading to low 'expected' count values in the crosstabulation; it is generally accepted that 20 % of cells at most may have a count less than 5. In the context of listening tests, small samples are likely to be the most common of these two issues; if this is the case, a correction may be applied or categories collapsed together to increase the number of cell counts [185]. Despite these notes of caution, the

chi-square analysis is an extremely useful check to perform on the data before any other analysis where statistical independence of trials is a critical assumption; violating this assumption is a serious problem and it may be difficult to proceed with analysis at all in any depth.

If the result of a χ^2 analysis is statistically significant, the conclusion is that listener responses were not independent of stimulus; this is interpreted as saying that they judged a loudspeaker differently depending on which programme material it was reproducing at the time of evaluation. For the listening tests performed in this study, this meant that subsequent statistical analysis classed musical extract as an additional experimental variable; listener responses could not be collated across extracts, but were treated as separate experimental groups, i.e. samples taken from different populations. However, if there was good reason to assume that the use of different musical extracts did not influence listener judgements, responses from all trials could be collated together for analysis, increasing the sample size and allowing conclusions to be made with greater confidence.

6.3.2 Calculating the Value of χ^2

The steps involved in calculating χ^2 are not described here in great detail as it is a standard statistical technique and descriptions of the method can be found in many textbooks; Snedcor and Cochran [186] gave a particularly clear example based on agricultural testing. However, an example crosstabulation is presented in Figure 58 to demonstrate how such a table was constructed using the listening test variables considered in this study, loudspeaker model vs musical extract.

			Extract			
			1	2	3	Σ_{rows}
Loudspeaker	A	*f*	13	17	14	44
		F	15	15	15	
	B	*f*	7	3	6	16
		F	5	5	5	
		$\Sigma_{columns}$	20	20	20	60

Figure 58: Example crosstabulation

In each cell, f is the observed value, i.e. counts returned by the listening test. The expected value, F, is calculated using:

$$F_{ij} = \frac{\Sigma_i}{n}\Sigma_j \tag{6.1}$$

where: Σ_i is the total for row i, Σ_j is the total for column j, and n is the sum of row totals; n should also be equal to the sum of column totals.
The value of χ^2 is then found using:

$$\chi^2 = \sum \frac{(f - F)^2}{F} \tag{6.2}$$

Note that if the observed and expected frequencies in each cell are very similar, this is an indication that the value of χ^2 will be low.
The other piece of information needed when performing a χ^2 analysis is the degrees of freedom, determined by the number of rows and columns in the crosstabulation; this must be used when consulting tables of critical values for the χ^2 distribution and is calculated using:

$$\text{d.f} = (N_{\text{rows}} - 1)\,(N_{\text{columns}} - 1) \tag{6.3}$$

For the example shown in Figure 58, d.f. = 2.

6.3.3 Summary of Procedure for Detecting Programme Dependence

The procedure for investigating the presence of programme dependence using the χ^2 test for independence is summarised here. This method is applied to the listening test data in sections 7.2.1, 7.3.1, and 7.4.1.

i) State null and alternative hypothesis:

- H_0: Selected design is independent of stimulus; observed count = expected count, χ^2 small.
- H_A: Selected design is not independent of stimulus; observed count $\neq$ expected count, χ^2 large.

ii) Choose an acceptable significance level (e.g. $\alpha = 0.05$).
Note: the χ^2 distribution only has positive values. This means that the test statistic can only be in one direction and thus, tests on this distribution are always one-tailed.

iii) Construct the crosstabulation and populate observed and expected frequencies.

iv) Check that no more than 20 % of cells have an expected count less than 5; refer to suggestions in section 6.3.1 if this check fails.

v) Calculate the value of χ^2 using tabulated values.

vi) Consult table of critical values for the χ^2 distribution using the corresponding d.f.:

- Reject H_0 in favour of H_A if the calculated result, χ^2_c, is equal to or exceeds the critical value for the chosen α level: Conclude that programme dependence is observed; the difference in counts between stimuli for at least one design is too great to be attributed only to sampling error.
- Do not reject H_0 if χ^2_c is less than the critical value for the chosen α level: Conclude that programme dependence is not observed; any difference in counts between stimuli for a given design are due to random sampling variations.

The outcome of step vi) determined whether musical extract was counted as an additional experimental variable in further analysis of the listening test data.

6.4 A/B Pair Results

Some of the more unusual, and complicated, techniques that have been suggested for analysis of ABX listening test data include Bayesian methods [187], preference trees [159], and use of the Polya-Eggenberger distribution [188]. Methods derived from theory by Thurstone [189], and Bradley and Terry [190] are occasionally used and are slightly more accessible, though it is difficult to find the procedures explained in practical detail. It also seems that these tests can be laborious to apply and prone to invalidation due to their reliance on unrealistic assumptions [160, 183, 191–193]. Other methods sometimes cited for ABX analysis are based on the binomial distribution. These are simple to understand and apply, and their use in subjective testing is already well established; they are widely used in audiology, commonly for establishing hearing thresholds [149, 182], and are the basis of the standardised procedure used in sensory paired-comparison tests [147]. Leventhal [194, 195] and Burstein [196] produced useful papers discussing application of the binomial distribution specifically within the context of listening tests on audio equipment; this work provides a very useful basis for statistical analysis of ABX data of the type collected in this study. These papers, plus sources from the social sciences, were used to clearly define a suitable method for analysing the experimental data; this enabled credible conclusions to be drawn about the audibility of differences between a single pair of virtual loudspeakers, and allowed comparison across multiple pairs. This method is discussed and summarised in the following subsections.

6.4.1 Hypothesis Testing Using the Binomial Distribution

Participants in a 2AFC ABX listening test are assumed to vote randomly, or 'due to chance', in the absence of an experimental effect: the null hypothesis in this study. The probability that the listener selects A is equal to the probability that they select B in any given trial:

$$\mathrm{P}(A) = \mathrm{P}(B) = 0.5 \tag{6.4}$$

In listening experiments for this study, A and B are two virtual loudspeakers being compared against an ideal fixed reference, X. The objective of the experiment is to establish that an effect exists, i.e. that listeners can firstly detect a difference between A and B, and then decide which of them is most similar to X. There are some implications with statistical power if the objective is to establish absence of an effect [147], but this is not discussed here further as it was not relevant to the experiments in this project.

The raw A/B data consists of counts, or frequencies, of the total number of times listeners 'voted' for each loudspeaker in the pair. This type of data is dichotomous, meaning that the outcome can only be one of two possible alternatives, A or B; tests on the binomial distribution are therefore appropriate [197, 198]. The total number of listener responses for A and B within a pair is the sample size, n; this depends on the number of listeners and the number of repeats, including the use of more than musical extract, as already discussed. Intuitively, it is sufficient to compare the relative counts for each tested pair and conclude that any result other than an equal split between A and B indicates some audible effect. For a large sample, e.g. several hundred participant responses, simply comparing the total counts might be sufficient to draw reliable conclusions about the experiment. In listening tests, however, the amount of data is typically

small in statistical terms and sampling error is likely to affect the conclusions. An example of sampling error is when a coin is tossed 20 times and the result is not 10 heads, 10 tails, even though probability suggests that this should be the outcome. It is because of this random variation in the sample of results that binomial tests are useful when analysing ABX listening test data. The researcher can then conclude, with a specified level of confidence, whether a deviation from 50 % of the results is due to a real audible effect or just random fluctuations within the data.

6.4.2 Analysing One Pair

In the single-pair case, a simple binomial technique can be used to test an experimental hypothesis about the audible difference between A and B; the null hypothesis, H_0, assumes no effect, i.e. listener votes should be equally distributed between A and B. The purpose of the analysis is to establish how great the split of listener votes between A and B must be before it can be concluded that they were really reaching a consensus about the judgement. Direct computation using the expression for the binomial distribution can be cumbersome beyond very small values of n, so it is convenient to use tabulated values. These can be found in many statistical textbooks, but Leventhal [194] lists values for a finite number of sample sizes and explains clearly how they should be used. In the absence of such tables, a normal approximation to the binomial distribution may be used if the sample is large enough, and $p = 0.5$, as assumed in this study. 'Large enough' is generally taken to mean $n > 30$ [199], but it has been suggested that the normal approximation remains accurate with a sample size as low as 15 [196]. Expressions for the normal approximation can be solved using Equation 6.5 [196]:

$$z = \frac{c - 0.5 - np_1}{\sqrt{np_1\,(1 - p_1)}} \tag{6.5}$$

where: p_1 is the proportion of correct responses in population due to chance alone, assumed to be 0.5 in this study; n is the sample size, total counts for A + total counts for B, and c is the number of correct responses, e.g. counts for A.

The quantity z is a normalised measure for the number of standard deviations from the mean value of a distribution; thus, once Eqn. 6.5 has been evaluated, z must be found in a table of critical values for the normal distribution to find the associated p-value. If p is equal to or less than the chosen significance level, α, the null hypothesis can be rejected.

6.4.3 Extending to Multiple-Pair Hypothesis Tests

If multiple pairs are to be evaluated, the single-pair test can be performed repeatedly to find the exact p-value for each one. Alternatively, a binomial 'threshold' can be established, using a single chosen value of α; the split of A/B results in a pair must meet or exceed this threshold to consider the result significant, i.e. to conclude that there is an audible effect within that pair that caused listeners to consistently choose A over B, or vice-versa. Burstein [196] refers to this count threshold as *critical* c (c') and gives the expression to calculate it; this is Eqn. 6.5 rearranged to make c the subject:

$$c' = z\sqrt{np_1\,(1 - p_1)} + 0.5 + np_1 \tag{6.6}$$

As described in section 5.2.5, experiments in this study were designed such that all listeners

evaluated all experimental models in all possible pair combinations an equal number of times; this arrangement is therefore well suited to analysis with the binomial critical threshold technique. The value for c' is compared against the paired count totals to conclude whether there is a statistically significant difference between the number of times listeners selected A compared to B. Once calculated, a single value for c' can be applied to the results from all A/B pairs in the experiment, as long as they all share the same sample size, n. This is a very important consideration. Eqn. 6.6 shows that the absolute value of c' increases with sample size n, but critical count as a proportion of sample size, c'/n, decreases as the sample size gets larger; this means that as the pool of listener responses gets smaller, consensus in their responses must be greater before a significant result can be concluded because the critical count threshold is proportionally higher than for a larger sample.

6.4.3.1 Correcting Alpha If comparing results across different A/B pairs, a correction must be made to the chosen significance level, α [200, 201]; this is to account for the fact that the increased amount of data also means an increase of the number of 'false positive' results, and increases the probability of making a Type I error when considering the results across all pairs. This correction is especially important when the sample size is small, such as in listening tests, as a greater effect size is needed to reach statistical significance. The maximum α level must be reduced in order to say, with the same level of confidence, that the experimental findings are significant across all cases being compared. Leventhal [194] discussed the issue in relation to ABX listening tests; this is developed here so that it can be used in the case when evaluating the normal approximation formula for c' to compare against multiple A/B pairs.

Considering a case where repeated analysis of single A/B pairs had been performed using Eqn. 6.5 , the probability that at least one of these instances will commit a Type I error is:

$$\alpha = (1 - \alpha_1)\,(1 - \alpha_2) \cdots (1 - \alpha_G) \tag{6.7}$$

where: α_1 to α_G are the α values associated with pairs 1 to G being compared.

If trying to find a value for c' that can be used for multiple A/B pairs, Eqn. 6.7 must be rearranged to find α_{func}, a 'functional' value of α that is used to find the corresponding value of z for substitution into Eqn. 6.6. Using α_{func} rather than α holds the Type I error at a fixed value for comparisons across all pairs.

If the same significance level is selected for all pairs, Eqn. 6.7 becomes:

$$\alpha = 1 - (1 - \alpha_{\text{func}})^G \tag{6.8}$$

Rearranging Eqn. 6.8 to get α_{func}:

$$\alpha_{\text{func}} = 1 - (1 - \alpha)^{\frac{1}{G}} \tag{6.9}$$

where: α is the selected significance level, and G is the total number of results being compared.

Note from Eqn. 6.9 that α_{func} decreases as G, the number of pairs being compared, increases; thus, the effective value of α used to find the binomial threshold (c') for each A/B pair is likely to be much smaller than the significance level initially chosen. Consequently, the threshold increases; the split of A/B results must therefore be much greater to return a statistically

significant result when comparing multiple pairs than if only a single pair is being considered.

6.4.4 Directional Hypothesis Testing

The issue of directional hypothesis testing is addressed here briefly as it is pertinent to the listening test method described in section 5.2.5. In practice, the distinction between one- and two- tailed testing may seem trivial, as $\alpha_{\text{one-tail}} = \alpha_{\text{two-tail}}/2$ [202]. Formally however, the distinction should be observed because deciding whether to conduct a one- or two-tailed test depends on whether or not there is an a-priori correct response for a given trial, i.e. whether a definite correct response is identified before the experiment, or is simply the decision that gains the majority of responses from participants, also called the 'consensus' vote [147]. The distinction was important for analysis in this study, where not all trials contained a predefined correct response. Leventhal and Huynh [195] presented a clear case for using a one-tailed test in ABX analysis. If it is known that the experimental effect should be in a specific direction, then a one-tailed test will be appropriate; this applies to all hidden reference trials in listening tests in this study, where identifying the hidden reference was considered to be a correct response. As described in section 5.2.5, all other pair comparisons presented two different virtual loudspeakers; neither A nor B was identical to the reference. The consensus vote was then considered to be the correct answer and a two-tailed test conducted.

6.4.5 Summary of Procedure for Multiple A/B Pair Analysis

The procedure for establishing a binomial threshold for multiple A/B pair comparisons is summarised here. Application of this method to the listening test data is shown in sections 7.2.3, 7.3.3, and 7.4.3.

i) State the null and alternative hypothesis:

- H_0: Listeners selected A the same number of times as B ($c_A = c_B$; assume no experimental effect).
- H_A: Listeners selected A more or less than B ($c_A \neq c_B$; assume some audible effect). Note: the non-directional alternative hypothesis here implies that a two-tailed test will be used. For hidden reference trials, the alternative hypothesis is directional: $c_A > c_B$, assuming A is the hidden reference.

ii) Choose a significance level; in this study, $\alpha = 0.05$ unless otherwise stated.

iii) Sort all data into pairs, listing the A/B count totals.

iv) Define the number of pair results to be compared, G.

v) Calculate α_func using Eqn. 6.9 and look up the value of z corresponding to this significance level.

vi) Calculate the critical threshold, c', using Eqn. 6.6 and the z-score associated with α_func.

vii) Compare c' with each pair of A/B results:

- If A is equal to or exceeds c': significant in favour of A; reject H_0 in favour of H_A at α level.

- If B is equal to or exceeds c': significant in favour of B; reject H_0 in favour of H_A at α level.
- If neither A nor B reaches c': conclude that there is no significant result for the pair; fail to reject H_0 at α level.

In this study, rejection of the null hypothesis for a given pair of loudspeaker models was treated as a clear directional outcome from the listener assessment e.g. A>B. These results formed the basis for comparison with the MTF results.

6.5 Post-Screening by Individual Performance

It is useful for a researcher to be able to assess the performance of listeners taking part in an experiment; they may be tested for consistency, whether they give the same answer when repeatedly presented with the same set of stimuli, as well as accuracy, whether they are able to give 'correct' answers and thus, if they are correctly performing the experimental task. In order to carry out statistical analysis on these aspects of listener performance, the experiment must have been designed with this in mind; there must be a sufficient number of repeated trials per listener to check their consistency, and a certain number of trials with a pre-defined 'correct' response in order to investigate accuracy. As described in section 5.2.5, the experimental method used in this study permitted such analysis, having both repeated trials by musical extract, and the inclusion of a hidden reference as one of the experimental systems to be evaluated.

6.5.1 Requirements and Assumptions

In a balanced experiment, each listener performs the same number of hidden reference trials. Each of these trials has only two outcomes: correct or incorrect. As the number of hidden reference responses from each listener is the same, a binomial threshold method, very similar to that described in section 6.4, was used to test the hypothesis that a given listener correctly identified the hidden reference often enough to be confident that they were performing the task at, or above, the required level; the inference is that they selected the hidden reference in a sufficient majority of trials due to accurate audible discrimination. If an individual was able to consistently detect that the hidden and known references were identical, even in pairs where audible differences were subtle, it was assumed that they would be a competent judge of other pair combinations. If a participant was unable to identify a sound as being identical to itself in a direct presentation, they could not be expected to reliably perform a more complex auditory comparison. This use of a hidden reference to investigate intra-listener performance may be seen as inefficient, as extra trials are 'wasted' on a model not intended to be part of the experimental group under investigation. Provided that the number of extra trials is carefully balanced against session duration, this apparent inefficiency is worthwhile, especially if the test participants have widely varying critical listening abilities.

It is acknowledged that post-screening a data set to produce a more favourable outcome in subsequent analysis is controversial and should generally be avoided; a listener's data certainly should never be discarded or excluded from analysis without very clear justification. Assessing a participant based on their hidden reference performance was considered to be a legitimate benchmark of listening competence. Excluding data from participants who failed to correctly

identify the hidden reference a predefined number of times was viewed as an additional stage of analysis that was relevant to the study; both pre- and post-screened outcomes are shown in presentation of the results in chapter 7. As demonstrated there in more detail, this method of post-screening to find the best performing listeners, rather than preselection through pilot testing, is a problematic strategy if the group of participants show greatly varying abilities. It is decided before analysis that listeners must identify the hidden reference a given number of times to be performing the task sufficiently well; if it is found that the majority do not meet this criterion, the responses from most listeners must be excluded to form the post-screened data set. The sample size in analysis of this data may be so small that it makes it difficult to draw reliable conclusions. It must then be decided if and how it is appropriate to reduce the performance threshold to include the responses of more participants, therefore limiting the reduction in the post-screened sample size.

6.5.2 Summary of Procedure for Post-Screening

For analysis of intra-listener performance, the principles described in section 6.4 remain the same but with some key differences. Firstly, the sample size, n, in the intra-listener case becomes the total number of trials featuring the hidden reference for a given listener. In this study, n was too small to use the normal approximation, so the exact binomial distribution was used instead. Unlike the pairwise analysis, where a threshold was set to compare results across many sets of collated results, performance was assessed on an individual-listener basis; they were viewed as having participated in a separate individual experiment. Therefore, there was no adjustment of α to compensate for comparison across multiple listeners. As this separate experiment only contained trials where one of the loudspeakers was different from the reference, there was always an a-priori correct response; a one-tailed test would therefore be used. It has been suggested that a two-tailed test might be more appropriate if there is reason to suspect that individual listeners will demonstrate 'statistically significant poor performance' [203]; this would indicate performance worse than chance, perhaps due to an experimental fault, but there was no reason to suspect such performance in this study.

The procedure for evaluating intra-listener performance is summarised here. Application of this method to the listening test data is shown in sections 7.2.2, 7.3.2, and 7.4.2.

i) State the null and alternative hypothesis:

- H_0: Listener selected A the same number of times as B, $c_A = c_B$ assuming A is the hidden reference. Conclude listener cannot detect the hidden reference.
- H_A: Listener selected A more than B, $c_A > c_B$. Conclude listener can detect the hidden reference.

ii) Choose an acceptable value for α.
As discussed in section 6.5.1, this must be selected carefully for each experiment. It may require a compromise between the desire for a high level of confidence, minimising Type I error, and avoiding unnecessary reduction in subsequent sample size.

iii) Use the exact binomial distribution to find the minimum number of correct responses, c_R, out of n trials required to give a p-value lower than α.

iv) Arrange the data by individual listener; separate out only trials containing the hidden reference.

v) Find the total number of times the listener correctly identified the hidden reference, c_A.

vi) Compare c_R with participant's results:

- If $c_A \geq c_R$: Listener reliably detected the hidden reference; reject H_0 in favour of H_A at α level and conclude their acuity meets the required standard.
- If $c < c_R$: Listener did not reliably identify the hidden reference; do not reject H_0 and conclude that acuity is below the required standard.

Further analysis was performed only on the data from listeners found to have correctly identified the hidden reference at least c_R times.

6.6 Summary of Statistical Methods

Three statistical methods were described for analysis of listening test data. It was established that these methods were appropriate for the type of data generated by all listening tests conducted in this study. The procedures were clearly defined, combining evidence from the social sciences, sensory testing, and the audio literature.

The effect of programme dependence was investigated using the chi-square (χ^2) test for independence. This test was used to decide whether programme material was biasing results for a given loudspeaker. A significant result from this test meant that musical extract must be treated as an additional experimental variable. If the χ^2 test was not significant, it was concluded that programme dependence did not influence the experimental results. Data collected across all extracts within a given experiment were considered as samples from the same population and thus, could be collated for further analysis.

When inspecting results from comparisons of the loudspeaker models, a normal approximation to the binomial distribution was used to calculate a critical count threshold. A result was considered statistically significant if either loudspeaker in the pair met or exceeded this threshold. Failure to reach a statistically significant result with this method meant that neither model in the pair received a great enough majority of votes. The data did not allow a conclusion as to whether listeners could detect a difference between the two systems; failure to reject the null hypothesis only permitted the conclusion that listeners were not in sufficient consensus regarding the experimental question: which of the two sounded most like the reference.

Intra-listener performance was assessed using the results from hidden-reference trials; here the listener unknowingly compared the reference against itself, along with one different loudspeaker model. For this analysis, the exact binomial distribution was used to determine a reference count; an individual listener had to correctly identify the hidden reference at least this many times to be confident that they were performing the listening task to a predefined standard. This test was used to post-screen the data set before repeating the pairwise analysis to inspect results from only those participants judged to be the most critical listeners.

Application of these methods to the experimental data is demonstrated in chapter 7.

7 Subjective Evaluation of Loudspeaker Models

The general experimental method and procedures for statistical analysis were described in detail in chapters 5 and 6 respectively. This chapter presents results from the listening tests performed separately on each group of loudspeaker models. A short summary of the key procedural details is presented in section 7.1 to remind the reader of the general method. Results from each test are presented in sections 7.2 to 7.4. Each of these begins with a brief description of test-specific details; analysis of the pairwise model comparisons is presented after investigation of programme dependence and intra-listener performance. A summary of analysis for every listening test is presented at the end of each individual section. The overall findings are compared and discussed in section 7.5, with reference to the primary experimental questions defined in section 6.1.

7.1 Recap of Procedure

A brief summary is given here to clarify and highlight the main features of the experimental method that were developed and described in chapter 5.

Listening tests I and II compared the virtual loudspeakers, model Groups I and II, arbitrarily labelled C–G plus reference R. Listening test III evaluated the 'artificial models', Group III; these were labelled z_2, z_4, z_6, z_8, to identify them as being a different type of model from the virtual loudspeakers and provide a reminder of their characteristics (systems with fixed magnitude but phase shifts from order 2 to 8).

Trials in all experiments were a two-alternative forced choice where participants selected which of A or B they judged to sound most like the reference X. In each trial, listeners compared A, B, and X using a physical three-way switchbox to swap between the models before recording their judgement on a small touch-screen interface. The models were evaluated using musical extracts, each of which was given a two-letter identification tag. Model and extract labels were not visible to participants.

There were four types of trial overall, as illustrated in Figure 59; two types of dummy trial, allowing participants to practise the task, and two types of formal trial which were the main part of the experiment.

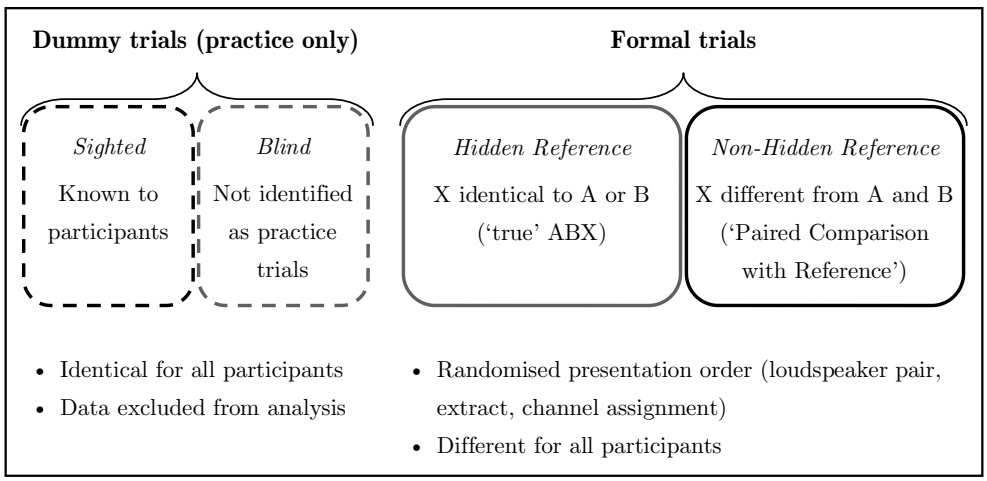


Figure 59: Types of listening test trial

7.2 Listening Test I: Evaluation of Group I Models

The first set of listening tests evaluated the virtual loudspeakers in Group I, as presented in section 4.2. Twenty-three participants completed the experiment ($L = 23$), each evaluating the six virtual loudspeakers (R, C–G) in every pair combination ($N = 15$, from Eqn. 5.1). Each pair was evaluated using three musical extracts ($M = 3$), all 25 s in duration, as described in section 5.3.3.1. This gave $NM = 45$ trials per listener. Each participant was instructed in how to conduct the experiment before performing three 'sighted' practise trials, identified to them as not being part of the formal test; they were then allowed to ask any further questions about the procedure before starting. Trials were split across two listening sessions, each with three 'blind' dummy trials at the start of the formal experiment, one per extract; listeners were not aware that they were performing these additional practise trials. They were identical for all participants and results were excluded from analysis.

7.2.1 Extract Analysis (Programme Dependence)

The distribution of all listener responses across extracts for each loudspeaker is shown in Figure 60; the width of each bar represents the percentage of responses but the exact count values are also shown. The total sample size was given by $n = NML = 1035$.

A chi-square test for independence was performed on this data and is summarised below. All values for χ^2 are given to 3 d.p.

$$H_0\text{: } \chi_c^2 < \chi_\alpha^2\text{; loudspeaker selection is independent of extract.}$$
$$H_A\text{: } \chi_c^2 \geq \chi_\alpha^2\text{; loudspeaker selection is influenced by extract.}$$

$$\alpha = 0.05; \quad n = 1035; \quad \text{d.f.} = 10; \quad \chi_{0.05}^2 = 18.307;$$
$$\text{Calculated value of } \chi^2 \text{ for listening test I: } \chi_c^2 = 3.994.$$

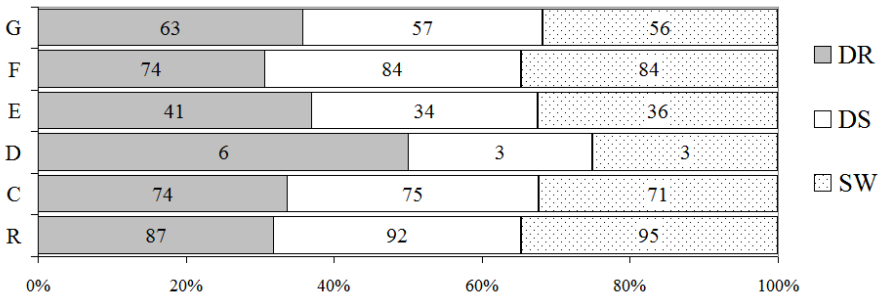


Figure 60: Group I extract distribution; data from 23 participants, $n = 1035$. The bars show allocation of listener responses for loudspeaker models R–G; absolute count totals are shown inside the bars. The horizontal divisions show distribution of votes for each model according to the musical extracts used for evaluation: DR, DS, and SW

The calculated result did not exceed the critical value of χ^2 at the 5 % level for a table with 10 degrees of freedom; the null hypothesis was not rejected and it was concluded that there was no evidence of programme dependence. The listener responses were therefore collated across all three extracts for each loudspeaker pair in further analysis.

7.2.2 Post-Screening (Intra-Listener Performance)

Responses from individual participants in only the hidden reference trials were inspected (bordered by the grey solid line in Fig. 59). The distribution of results showed that participants had identified the reference in the majority of cases but with varying consistency. Figure 61 shows the distribution, having the following statistics:

$$n = 15 \text{ (number of trials)}; \; c_{\min} = 7; \; c_{\max} = 15;$$
$$\text{mean} = 11.9 \; ; \; \text{median} = 12; \; \text{mode} = 13.$$

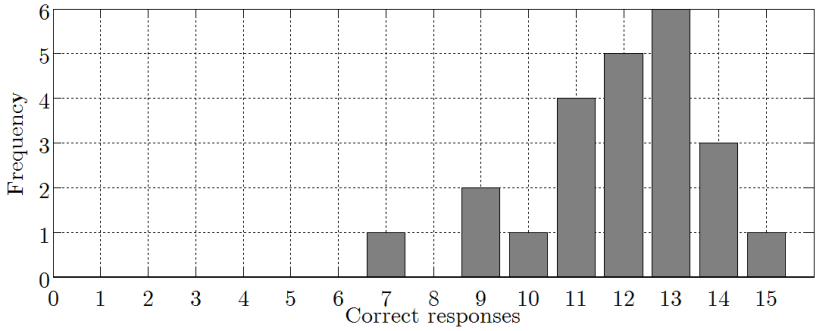


Figure 61: Distribution of correct responses from Group I hidden reference trials; data from 23 participants, $n = 15$.

A performance limit for post-screening was established using the exact binomial distribution. To be significant at the 2.5 % level, the minimum required number of correct hidden reference responses was 12 out of 15 (80 %):

$$\alpha = 0.025; \qquad P\,(c \geq 12) = 0.0176 \text{ (4 d.p., 1-tailed)}$$

Of all 23 participants, 15 (65.2 %) correctly identified the hidden reference in at least 12 trials. Data from these listeners formed the post-screened data set analysed in section 7.2.4.

7.2.3 Pairwise Results Based on Total Sample (All Listeners)

Pairwise results were first inspected using data from all participants. The individual pair analysis was conducted in two groups. Comparisons containing the hidden reference were treated separately to reflect the fact that these trials featured a direct comparison; the listening task should therefore have been easier, assuming that an audible difference between A and B existed. The significance level was halved in analysis of these pairs, and a one-tailed test conducted owing to the a-priori correct response, as discussed in section 6.4.4.
The sample size for each pair was $n = ML = 69$ responses. Following the procedure described in section 6.4.5, the critical count threshold was computed for each set of pairs as follows:

Pairs not containing the reference (A and B different from X):

H_0: $p_{\mathrm{A}} = p_{\mathrm{B}}$; proportion of responses across loudspeakers A and B is equal.
Conclude neither sounds more like the reference.

H_A: $p_\mathrm{A} \neq p_\mathrm{B}$; proportion of responses across loudspeakers A and B is different.
Conclude an audible difference exists and one sounds more like the reference.

$\alpha = 0.05$; $\quad G = 10$; $\quad \alpha_\mathrm{func} = 0.0051$, two-tailed.
$c' = 47$; either A or B must receive at least c' votes (68.1 % of responses for $n = 69$) within a pair for the split to be considered significant; H_0 rejected in favour of H_A.

The non-directional alternative hypothesis means that no a-priori statement can be made about which way the split of results might go. Either A or B could get the majority of responses and the result would be considered significant as long as the critical threshold is reached; the correct result is therefore the consensus vote, as judged by listeners. Note that failure to reject the null hypothesis in these pairs does not lead to the conclusion of no audible difference; the data does not permit this question to be answered directly. According to the experimental question, it can only be concluded that listeners were not in sufficient consensus as to which of A or B sounded most like the reference X.

Pairs containing the reference (A or B identical to X):

H_0: $p_R = p_{\widetilde{R}}$; proportion of responses is equal for the hidden reference (R) and the other loudspeaker model (not R).
Conclude that they were not audibly different.

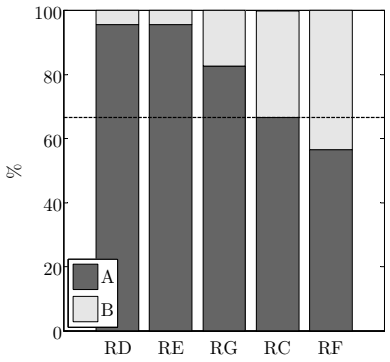
H_A: $p_R > p_{\widetilde{R}}$; proportion of responses for hidden reference is greater than for the other loudspeaker model.
Conclude an audible difference exists between A and B, and listeners were able to identify which one was the hidden reference.

$\alpha = 0.025$; $\quad G = 5$; $\quad \alpha_\mathrm{func} = 0.0051$, one-tailed.
$c' = 46$; hidden reference must receive at least c' votes (66.7 % of responses) within a pair for the split to be considered significant; H_0 rejected in favour of H_A.

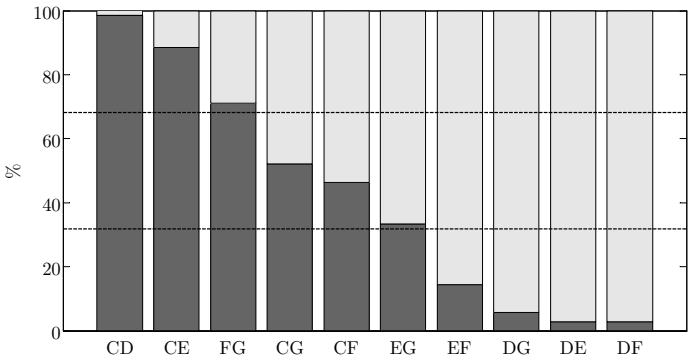
Note that the null and alternative hypotheses in these pairs leads to a conclusion about audibility. This is an inference based on the fact that the listening task was a direct comparison. The conclusions might therefore be stated more accurately as: the difference was, or was not, sufficiently audible to the majority of listeners to be confident that they were really identifying the hidden reference.

Figure 62 illustrates the results of pairwise analysis. The bar shading represents the percentage split of counts between loudspeakers A and B within a pair. Note that although channel assignment was randomised during experiments, it is fixed here for consistency of data presentation such that A is always the first model in the pair label. For hidden reference pairs, shown in Figure 62a, there is only one threshold line due to the directional alternative hypothesis; the null hypothesis was only rejected in pairs where the split of A/B bars lay on or above the threshold. In Figure 62b, the horizontal dotted lines show the critical count thresholds. When the split of A/B results falls between these lines, the difference is not considered statistically significant; the null hypothesis was not rejected, and it was concluded that the majority of listeners could not decide which of A and B sounded more like the reference. The count results and corresponding exact p-values for each pair in listening test I are listed in

Table 11, quantifying the probability of making a Type I error (the risk of wrongly rejecting the null hypothesis). All p-values were calculated from the normal approximation (Eqn. 6.5) and show $P\left(c \geq c_{\mathrm{A}}\right)$ for pairs containing the hidden reference, and $P\left(c \geq c_{\max(\mathrm{A,B})}\right)$ for all others, where c is the number of listener votes; for example, in pair RD from Table 11, p shows the probability of R getting at least 66 out of 69 votes if the null hypothesis is true. Values less than α_{func} are considered statistically significant at the α level. Pairs where H_0 was not rejected are highlighted by grey shaded columns in the results tables.



(a) Hidden reference pairs; critical threshold marked at 66.7 %



(b) All other pairs; upper and lower thresholds are marked at 68.1 and 31.9 %

Figure 62: Group I pair results; data from 23 participants. Horizontal dotted lines mark the critical thresholds. For pairs featuring the hidden reference, Fig. 62a, a significant result at the 2.5 % is returned if the split between A and B bars meets or lies above the threshold line. Where the split between A and B bars falls between the lines in Fig. 62b, the result is considered not significant at the 5 % level

A	66	66	57	46	39
B	3	3	12	23	30
p	<0.0001	<0.0001	<0.0001	0.0040	0.1678
pair	RD	RE	RG	RC	RF

$n = 69$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.025$, $G = 5$)

(a) Hidden reference pairs

A	68	61	49	36	32	23	10	4	2	2
B	1	8	20	33	37	46	59	65	67	67
p	<0.0001	<0.0001	0.0007	0.8097	0.6301	0.0081	<0.0001	<0.0001	<0.0001	<0.0001
pair	CD	CE	FG	CG	CF	EG	EF	DG	DE	DF

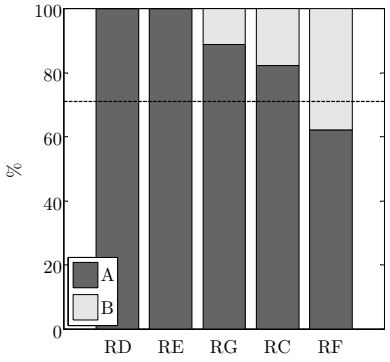
$n = 69$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.05$, $G = 10$)

(b) All other pairs

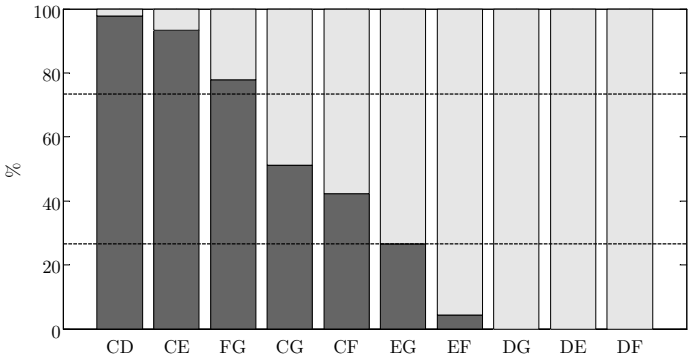
Table 11: Group I pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.2.4 Pairwise Results Based on Post-Screened Sample

Based on the intra-listener evaluation in section 7.2.2, pair analysis was repeated with the data from 15 participants who were identified as performing the listening task at a higher level. The analysis was identical to that in section 7.2.3 except that the smaller sample size, $n = 45$, increased the critical count thresholds. For hidden reference pairs, $c' = 32$ (71.1 %), otherwise $c' = 33$ (73.3 %). The results are summarised in Table 12 and Figure 63.



(a) Hidden reference pairs; critical threshold marked at 71.1 %



(b) All other pairs; upper and lower thresholds are marked at 73.3 and 26.7 %

Figure 63: Group I post-screened pair results; data from 15 participants. Horizontal dotted lines mark the critical thresholds. For pairs featuring the hidden reference, Fig. 63a a significant result at the 2.5 % is returned if the split between A and B bars meets or lies above the threshold line. Where the split between A and B bars falls between the lines in Fig. 63b, the result is considered not significant at the 5 % level

A	45	45	40	37	28
B	0	0	5	8	17
p	0.0000	0.0000	0.0000	0.0000	0.0680
pair	RD	RE	RG	RC	RF

$n = 45$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.025$, $G = 5$)

(a) Hidden reference pairs

A	44	42	35	23	19	12	2	0	0	0
B	1	3	10	22	26	33	43	45	45	45
p	0.0000	0.0000	0.0003	1.0000	0.3711	0.0029	0.0000	0.0000	0.0000	0.0000
pair	CD	CE	FG	CG	CF	EG	EF	DG	DE	DF

$n = 45$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.05$, $G = 10$)

(b) All other pairs

Table 12: Group I post-screened pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.2.5 Listening Test I Results Summary

Findings from listening test I are summarised below, grouped according to the separate stages of analysis.

7.2.5.1 Programme Dependence The distribution of listener responses by musical extract showed similar proportions across all virtual loudspeakers. This indicated that all three extracts elicited similar judgements from listeners. It was formally concluded that there was no evidence of programme dependence after a χ^2 test in which the null hypothesis was not rejected at the 5 % level ($\chi^2_c = 3.994$, d.f. $= 10$). Responses across all extracts were therefore combined in further analysis and treated as repeats of the same pair comparisons.

7.2.5.2 Intra-Listener Performance Inspection of intra-listener performance in 15 hidden reference trials showed a range of responses ($c_{\min} = 7$, $c_{\max} = 15$). Approximately two-thirds of participants, 15 out of 23, correctly identified the hidden reference enough times to meet the chosen performance level: $p < 0.025$, $c'_R \geq 12$. This level was deemed an acceptable compromise between selecting the most accurate listeners and maintaining an adequate sample size in the post-screened data set.

7.2.5.3 Pairwise Analysis The pairwise analysis was first conducted using data from all participants (no post-screening). The split of listener votes within each pair of virtual loudspeakers was compared against a critical count threshold. Sample size was 69 listener responses per pair. In pairs containing the hidden reference, the critical count was $c' = 46$ (one-tailed, $\alpha = 0.025$), a threshold of 66.7 %. For all other pairs, $c' = 47$ (two-tailed, $\alpha = 0.05$); H_0 was not rejected in favour of H_A in pairs where the split of listener votes lay between 68.1 and 31.9 %. Based on these limits, the null hypothesis was rejected in 11 out of 15 pairs. In all of these pairs it was concluded that listeners could detect a difference between the two alignments being compared, and select one as sounding most like, or identical to, the reference.

One of the four non-significant pairs featured the hidden reference; from this it was concluded that listeners could not distinguish loudspeaker F from the reference. Failure to reject H_0 in the remaining three pairs lead to the conclusion that listeners could not reach a consensus as to which of A or B sounded most like X; this may have been due to lack of audible differences, or the fact that both alignments were perceptibly different but neither was deemed as being overall more similar to the reference.

7.2.5.4 Post-Screened Pairwise Analysis The pairwise analysis was repeated using the post-screened data set; this reduced sample size per pair to $n = 45$. The critical threshold therefore increased relative to n. For pairs featuring the hidden reference, $c' = 32$ (one-tailed, $\alpha = 0.025$), a threshold of 71.1 %. For all other pairs, $c' = 33$ (two-tailed, $\alpha = 0.05$); H_0 was not rejected in favour of H_A in pairs where the split of listener votes lay between 73.3 and 26.7 %.. The increased threshold meant that listeners had to show a greater consensus within each pair to return a significant result.

After post-screening, 13/15 pairs showed a greater proportional A/B split of results; the selected listeners were overall in better agreement about their judgements. One pair, CG, remained exactly the same before and after post-screening at 51/49 %. One, CD, saw the A/B split reduce after post-screening, but this was not considered noteworthy as the counts in both data sets were only one vote away from complete unanimity: 68–1 pre-, and 44–1 post-screening. It was known that the pairs containing a hidden reference would show less variability after post-screening as listeners had been selected on this basis, but results showed that these

participants were overall in better agreement for the other pairs, where there was no hidden reference. It appeared that the assumption about performance in the hidden reference trials was correct; participants who performed well in a direct comparison were also better at more challenging auditory judgements. Despite the increased threshold, H_0 was rejected in one further pair, E vs G. Otherwise, significant pair results were the same as for the non-screened analysis.

Two pairs were of particular interest when compared to the original analysis. In pair R vs F, p_R reduced from 0.1678 to 0.0680; the higher-performing listeners identified the hidden reference more often. This indicated that F was not audibly identical to the reference, but distinguishing between them was difficult even for the more accurate listeners. However, the split of results was still within the limits of sampling error and, as for the all-listener group, H_0 could not be rejected; a conclusion about audibility could therefore not be made with any certainty. The other pair of interest was C vs G. Here the proportional A/B split of results remained the same after post-screening; the higher-performing listeners were in no greater consensus when choosing between these two models and the results remained almost equally divided ($c_C = 23$, $c_G = 22$). As the null hypothesis was not rejected, it is possible that C and G were audibly identical and the split of results reflects listeners' voting being 'due to chance'. However, this pair was conspicuous as it appeared to contradict the effect seen in all other pairs after post-screening; this group of participants otherwise reached a greater consensus in pair judgements compared to the other listeners. It is therefore also possible that the equal split of votes across C and G reflects a genuine difference in opinion; even the more accurate listeners could not decide which of these loudspeakers was overall most like the reference.

7.3 Listening Test II: Evaluation of Group II Models

The second set of listening tests evaluated the virtual loudspeakers in Group II, as presented in section 4.3. This group was developed to present more challenging auditory comparisons, as described in section 4.1. The first set of experiments has shown no evidence of programme dependence, and informal feedback from participants indicated that extract duration had been more than sufficient to make a judgement. In an effort to reduce test duration, and therefore allow more participants to complete the experiment within the given testing period, the Group II experiments were conducted with fewer trials and slightly shorter extracts, as explained in section 5.3.1. Otherwise the procedures for listening tests I and II were the same.

Twenty-six participants completed the experiment ($L = 26$). Eight of these had taken part in listening test I, so were familiar with the procedure; however, the risk of bias due to learning effects for these common participants was considered to be low due to the use of different models, extracts, and a separation of nearly eight months between the two experiments. Each person evaluated the six virtual loudspeakers (R, C–G), in every pair combination ($N = 15$); each pair was evaluated using two musical extracts ($M = 2$), all 23 s in duration, as described in section 5.3.3.1. This gave $NM = 30$ trials per listener. All trials were completed in one listening session, with two blind dummy trials at the start, one per extract; the dummy trials were identical for all participants and results were excluded from analysis. Each participant was instructed in how to conduct the experiment before performing two separate practise trials. They were then allowed to ask any further questions about the procedure before starting the test.

7.3.1 Extract Analysis (Programme Dependence)

The distribution of all listener responses across extracts for each loudspeaker is shown in Figure 64; the width of each bar represents the percentage of responses but the exact count values are also shown. The total sample size was given by $n = NML = 780$.

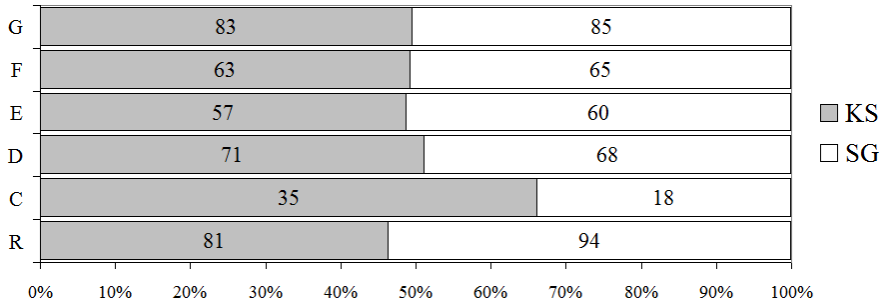


Figure 64: Group II extract distribution; data from 26 participants, $n = 780$. The bars show allocation of listener responses for loudspeaker models R–G; absolute count totals are shown inside the bars. The horizontal divisions show distribution of votes for each model according to the musical extracts used for evaluation: KS and SG

A chi-square test for independence was performed on this data and is summarised below. All values for χ^2 are given to 3 d.p.

$$H_0\text{: } \chi_c^2 < \chi_\alpha^2\text{; loudspeaker selection is independent of extract.}$$
$$H_A\text{: } \chi_c^2 \geq \chi_\alpha^2\text{; loudspeaker selection is influenced by extract.}$$

$$\alpha = 0.05; \quad n = 780; \quad \text{d.f.} = 5; \quad \chi_{0.05}^2 = 11.070;$$
$$\text{Calculated value of } \chi^2 \text{ for listening test II: } \chi_c^2 = 6.615.$$

Although the extract distribution results appeared to show some difference in allocation of responses for loudspeaker C, the calculated χ^2 result did not exceed the critical value at the 5 % level for a table with 5 degrees of freedom; the null hypothesis was not rejected and it was concluded that programme dependence did not affect the results. The listener responses were therefore collated across both extracts for each loudspeaker pair for further analysis.

7.3.2 Post-Screening (Intra-Listener Performance)

Responses from individual participants in only the hidden reference trials were inspected. The distribution of results indicated that some participants found this a difficult task and identified the reference in less than half of trials. Figure 65 shows the distribution, having the following statistics:

$$n = 10 \text{ (number of trials); } c_{\min} = 2; \; c_{\max} = 10;$$
$$\text{mean} = 6.7 \text{ ; median} = 7; \text{ mode} = 8.$$

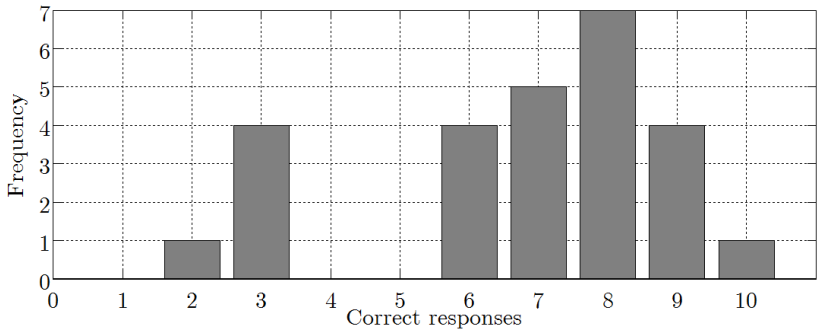


Figure 65: Distribution of correct responses from Group II hidden reference trials; data from 26 participants, $n = 10$

A performance limit for post-screening was established using the exact binomial distribution; to meet a significance level of 2.5 %, the minimum required number of correct hidden reference responses was 9 out of 10 :

$$\alpha = 0.025; \qquad P\,(c \geq 9) = 0.0107 \text{ (4 d.p., one-tailed)}$$

Of all 26 participants, only five (19.2 %) correctly identified the hidden reference at least nine times. Considering the impact on the main pairwise analysis after post-screening at this level, resulting sample size per pair was very small, $n = 10$; the corresponding critical threshold based on the exact binomial distribution was 100 % ($\alpha_{\text{func}} = 0.0051$, $c' = 10$). This was considered a problem as complete unanimity was required within and between listeners in order to reject the null hypothesis; it left no room for error or disagreement in the judgement of even one listener in a single trial. Therefore, a decision was made not to proceed with analysis based on this sample; post-screening was instead performed using a level of hidden reference performance providing some tolerance, requiring 8 out 10 correct responses. Using the exact binomial distribution, $P\,(c \geq 8) = 0.0547$ (4 d.p., one-tailed). The post-screened data set therefore contained responses from 12 of the original 26 participants.

7.3.3 Pairwise Results Based on Total Sample (All Listeners)

As for the Group I analysis in section 7.2.3, the individual pair analysis was conducted in two sets; trials containing the hidden reference were treated separately from those where both virtual loudspeakers in the pair were different from the reference. The hypotheses and subsequent conclusions are the same as presented in section 7.2.3; therefore, only the parameters for computing the critical count threshold are presented here. The sample size for each pair was $n = ML = 52$ responses.

Pairs not containing the reference (A and B different from X):

$$\alpha = 0.05; \quad G = 10; \quad \alpha_{\text{func}} = 0.0051, \text{ two-tailed.}$$

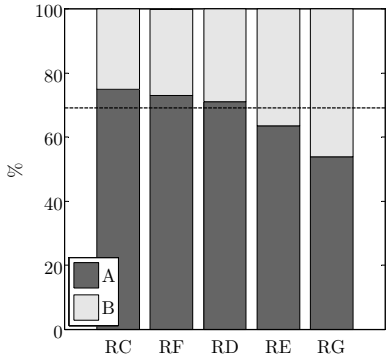
$c' = 37$; either A or B must receive at least c' (71.2 % of responses for $n = 52$) within a pair for the split to be considered significant; H_0 rejected in favour of H_A.

Pairs containing the reference (A or B identical to X):

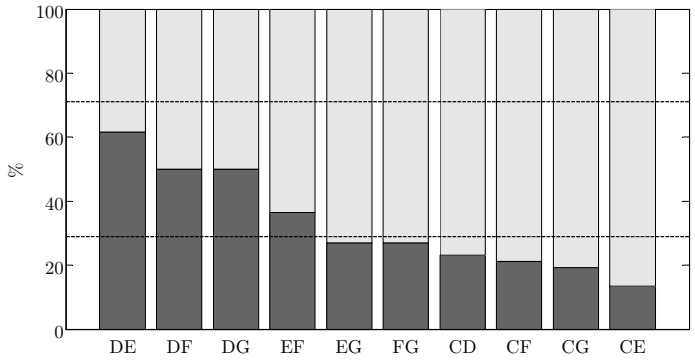
$$\alpha = 0.025, \text{ one-tailed}; \; G = 5; \; \alpha_{\text{func}} = 0.0051, \text{ one-tailed.}$$

$c' = 36$; hidden reference must receive at least c' (69.2 % of responses) within a pair for the result to be considered significant.

Figure 66 and Table 13 summarise the results of pairwise analysis of the data from listening test II. The presentation format is identical to that shown in section 7.2.3. Values for p less than α_{func} are considered statistically significant at the α level. As in Table 11, grey-shaded columns in the results tables highlight pairs where H_0 was not rejected.



(a) Hidden reference pairs; critical threshold marked at 69.2 %



(b) All other pairs; upper and lower thresholds are marked at 71.2 and 28.8 %

Figure 66: Group II pair results; data from 26 participants. Horizontal dotted lines mark the critical thresholds. For pairs featuring the hidden reference, Fig. 66a, a significant result at the 2.5 % level is returned if the split between A and B bars meets or lies above the threshold line. Where the split between A and B bars falls between the lines in Fig. 66b, the result is considered not significant at the 5 % level

A	39	38	37	33	28
B	13	14	15	19	24
p	0.0003	0.0007	0.0018	0.0357	0.3387
pair	RC	RF	RD	RE	RG

$n = 52$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.025$, $G = 5$)

(a) Hidden reference pairs

A	32	26	26	19	14	14	12	11	10	7
B	20	26	26	33	38	38	40	41	42	45
p	0.1272	0.8897	0.8897	0.0714	0.0014	0.0014	0.0002	0.0001	<0.0001	<0.0001
pair	DE	DF	DG	EF	EG	FG	CD	CF	CG	CE

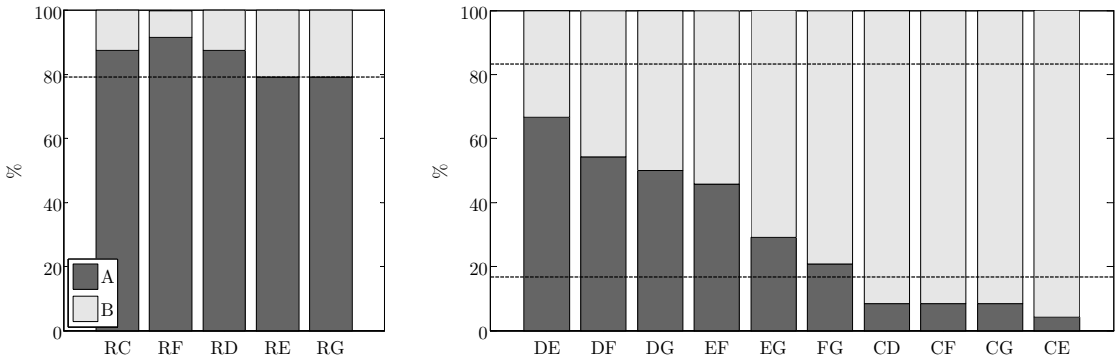
$n = 52$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.05$, $G = 10$)

(b) All other pairs

Table 13: Group II pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.3.4 Pairwise Results Based on Post-Screened Sample

Based on the intra-listener evaluation in section 7.3.2, pair analysis was repeated with the data from 12 participants who were identified as performing the listening task at a higher level. The analysis was identical to that in section 7.3.3 except that the smaller sample size, $n = 24$, increased the critical count thresholds. For hidden reference pairs, $c' = 19$ (79.2 %), otherwise $c' = 20$ (83.3 %). The results are summarised in Table 14 and Figure 67.



(a) Hidden reference pairs; critical threshold marked at 79.2 %

(b) All other pairs; upper and lower thresholds are marked at 83.3 and 16.7 %

Figure 67: Group II post-screened pair results; data from 12 participants. Horizontal dotted lines mark the critical thresholds. For pairs featuring the hidden reference, Fig. 67a, a significant result at the 2.5% level is returned if the split between A and B bars meets or lies above the threshold line. Where the split between A and B bars falls between the lines in Fig. 67b, the result is considered not significant at the 5 % level

A	21	22	21	19	19
B	3	2	3	5	5
p	0.0003	0.0001	0.0003	0.0040	0.0040
pair	RC	RF	RD	RE	RG

$n = 24$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.025$, $G = 5$)

(a) Hidden reference pairs

A	16	13	12	11	7	5	2	2	2	1
B	8	11	12	13	17	19	22	22	22	23
p	0.1530	0.8383	0.8383	0.8383	0.0662	0.0080	0.0001	0.0001	0.0001	<0.0001
pair	DE	DF	DG	EF	EG	FG	CD	CF	CG	CE

$n = 24$; $\alpha_{\text{func}} = 0.0051$ ($\alpha = 0.05$, $G = 10$)

(b) All other pairs

Table 14: Group II post-screened pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.3.5 Listening Test II Results Summary

Findings from listening test II are summarised below, grouped according to the separate stages of analysis.

7.3.5.1 Programme Dependence The distribution of listener responses by musical extract showed similar proportions across all but one of the six loudspeakers. This indicated that one of the extracts may have influenced some listeners' judgements. A χ^2 test was performed and the null hypothesis was not rejected at the 5 % level ($\chi^2 = 6.615$, d.f. $= 5$). It was therefore concluded that the apparent deviation in distribution was not great enough to be evidence of programme dependence; responses across both extracts were combined in further analysis and treated as repeats of the same pair comparisons.

7.3.5.2 Intra-Listener Performance Inspection of intra-listener performance in ten hidden reference trials showed a wide range of responses: $c_{\min} = 2$, $c_{\max} = 10$. Only five listeners performed at the preferred accuracy level: $p < 0.025$, $c'_R \geq 9$. This threshold was not used for post-screening as the resulting sample size was deemed to be unacceptably small. A reduced performance threshold was therefore adopted: $c'_R \geq 8$, $p = 0.0547$. Almost half of the participants, 12 out of 26, correctly identified the hidden reference this many times.

7.3.5.3 Pairwise Analysis The pairwise analysis was first conducted using data from all participants (no post-screening). The split of listener votes within each pair of virtual loudspeakers was compared against a critical count threshold. Sample size was 52 listener

responses per pair. In pairs containing the hidden reference, the critical count was $c' = 36$ (one-tailed, $\alpha = 0.025$), a threshold of 69.2 %. For all other pairs, $c' = 37$ (two-tailed, $\alpha = 0.05$); H_0 was not rejected in favour of H_A in pairs where the split of listener votes lay between 71.2 and 28.8 %. Based on these limits, the null hypothesis was rejected in 9 out of 15 pairs. In all of these pairs it was concluded that listeners could detect a difference between the two systems being compared, and select one as sounding most like, or identical to, the reference. Two of the non-significant pairs featured the hidden reference; from this it was concluded that listeners could not distinguish loudspeakers E and G from R. Failure to reject H_0 in the remaining four pairs lead to the conclusion that listeners could not reach a consensus as to which of A or B sounded most like X; this may have been due to lack of audible differences, or the fact that both loudspeakers were perceptibly different but neither was deemed as being overall more similar to the reference.

7.3.5.4 Post-Screened Pairwise Analysis The pairwise analysis was repeated using the post-screened data set; this reduced sample size per pair to $n = 24$. The critical threshold therefore increased relative to n. For pairs featuring the hidden reference, $c' = 19$ (one-tailed, $\alpha = 0.025$), a threshold of 79.2 %. For all other pairs, $c' = 20$ (two-tailed, $\alpha = 0.05$); H_0 was not rejected in favour of H_A in pairs where the split of listener votes lay between 83.3 and 16.7 %. The increased threshold meant that listeners had to show a greater consensus within each pair to return a significant result.

After post-screening, 12/15 pairs showed a greater proportional A/B split of results. One, DG, remained exactly the same at 50/50 %; two, EF and EG, saw the A/B split reduce after post-screening. It was known that the pairs containing a hidden reference would show less variability after post-screening as listeners had been selected on this basis, but results in the other pairs showed similar behaviour. Inspection of the post-screened results showed that the selected participants agreed better as a group about most of the judgements; in pairs where a significant result was returned after post-screening, the percentage split of A/B votes was always greater, i.e. further from 50/50, compared to the same pairs before post-screening. It appeared that participants who performed well in a direct comparison also showed greater consistency in more complicated auditory judgements. Despite the increased threshold, H_0 was rejected in 9 out of 15 pairs. This was the same number as before post-screening, but the conclusion in four pairs was different: RE, RG, EG, FG. After post-screening, H_0 was rejected in all pairs featuring the hidden reference R; it was therefore concluded that the reference was audibly different from all of the other loudspeakers, but only the more accurate listeners were able to make this distinction reliably.

Excluding the hidden reference trials, only pairs containing loudspeaker C returned a significant result in the post-screened analysis. Six out of ten pairs therefore had no defined outcome as to which loudspeaker sounded most like the reference when evaluated by the higher-performing listeners. It is possible that the majority of these participants could not hear differences between A and B and therefore voted randomly. These participants had been selected based on their increased discrimination ability in direct comparison trials, and they had demonstrated low inter-listener variability when evaluating loudspeaker C; group consensus was over 90 % in all four cases (CD, CE, CF, CG). It therefore seems unlikely the same group of listeners consistently failed to hear differences between A and B in 40 % of tested pairs. Where H_0 could not be rejected, it is considered likely that listeners could discriminate between A and

B but could not choose either as being overall more like the reference. As stated in section 4.3, the virtual loudspeakers in Group II had intentionally been designed to present more subtle audible comparisons; the subjective data suggests that this kind of evaluation was very difficult for the majority of participants. Greater acuity in detecting audible differences in a direct comparison did not appear to make a listener better at performing a critical judgement about relative differences between two loudspeakers' LF alignments.

Pair DG was of particular interest as the split of listener votes remained at exactly 50 % after post-screening; the accurate listeners showed no greater consensus in their judgement of these loudspeakers. Also of interest were EG and FG; both returned a significant result before, but not after, post-screening. The proportional split of A/B results for FG was greater after post-screening, but the critical threshold was 12 % higher in this case due to the reduction in sample size; it is therefore expected that the same result, significant in favour of G, would have been concluded if the sample had been larger.

7.4 Listening Test III: Evaluation of Group III Models

The final set of listening tests evaluated the artificial loudspeaker models in Group III, presented in section 4.5. The motivation for developing these models was different from Groups I and II, as described in section 4.1; the aim was to investigate whether listeners could detect audible differences between the simulated loudspeakers when only their phase response at low frequencies was altered. From a participant's perspective, the procedure was the same as listening tests I and II, but session duration was shorter due to time restrictions. To reduce the number of trials, only three pair combinations were evaluated ($N = 3$); comparisons were only made against the reference system, z_2:

$$z_2 \text{ vs } z_4; \quad z_2 \text{ vs } z_6; \quad z_2 \text{ vs } z_8.$$

Each pair was evaluated using four musical extracts ($M = 4$), all 20 s in duration, as described in section 5.3.3.3. This gave $NM = 12$ trials per listener. All trials were completed in one listening session, with four hidden dummy trials at the start, one per extract; the dummy trials were identical for all participants and results were excluded from analysis. Eleven participants completed the experiment; one was available to perform it again at the end of the testing period seven days later, so they were classed as an additional listener ($L = 12$). The risk of bias due to this listener being better trained in the task than the other participants was accepted because the sample size in this experiment was so small; the risk was reduced by ensuring that this person was allocated a different playlist, i.e. presentation order, on their second attempt. One person had participated in listening test I; another had taken part in both I and II. These listeners were therefore familiar with the procedure, but the risk of bias due to learning effects was considered minimal due to the separation of tests II and III of more than 12 months, and the use of different models and extracts. Each participant was instructed in how to conduct the experiment and were then allowed to ask any further questions about the procedure before starting the test.

7.4.1 Extract Analysis (Programme Dependence)

In this experiment, extracts with differing temporal as well as spectral characteristics had been selected to investigate how variations in signal content affected detection of audible differences

between the loudspeaker models. It was therefore expected that the results would show some evidence of programme dependence, but the nature of the effect was unknown prior to the experiment. The distribution of all listener responses across extract for each loudspeaker is shown in Figure 68. The width of each bar represents the percentage of responses but the exact count values are also shown. The total sample size was given by $n = NML = 144$.

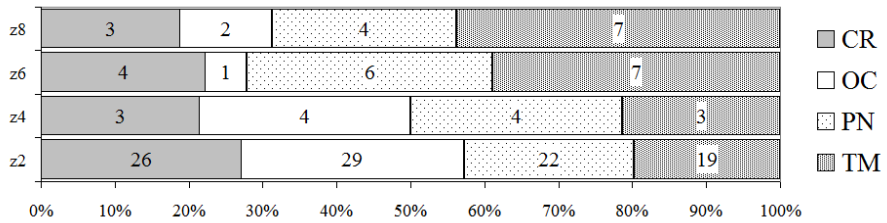
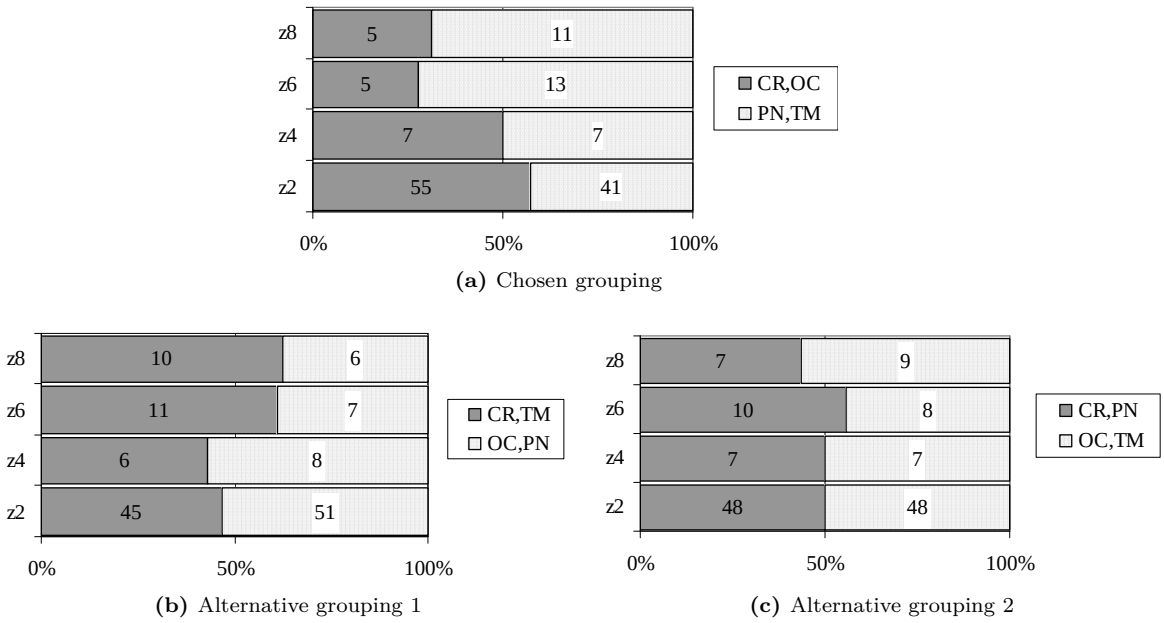


Figure 68: Group III extract distribution; data from 12 participants, $n = 144$. The bars show allocation of listener responses for models z_2–z_8; absolute count totals are shown inside the bars. The horizontal divisions show distribution of votes for each model according to the different extracts used for evaluation: CR, OC, PN, and TM

The initial χ^2 crosstabulation showed that 75 % of cells had an expected count less than 5. As discussed in section 6.3.1, this meant that the analysis could not be performed if treating each extract as a separate category. Inspection of the extract distribution in Figure 68 indicated similarities in allocation between extracts CR and OC compared to PN and TM; the categories were therefore collapsed by combining listener responses for extracts CR with OC, and PN with TM. Different arrangements were not analysed, but the possible alternatives are shown in Figure 69. It can be seen that the chosen arrangement appeared to give the clearest overall separations of vote allocation by extract compared to the other options.



(a) Chosen grouping

(b) Alternative grouping 1

(c) Alternative grouping 2

Figure 69: Chosen arrangement for collapsing extract groups in listening test III compared to alternatives. Bar widths reflect percentage split of listener responses; the total count values are shown inside each bar

A valid χ^2 test for independence could be conducted after collapsing the extract categories, as summarised below. All values for χ^2 are given to 3 d.p.

H_0: $\chi^2_c < \chi^2_\alpha$; loudspeaker selection is independent of extract.
H_A: $\chi^2_c \geq \chi^2_\alpha$; loudspeaker selection is influenced by extract.

$$\alpha = 0.05; \quad n = 144; \quad \text{d.f.} = 3; \quad \chi^2_{0.05} = 7.815;$$

Calculated value of χ^2 for listening test III (collapsed into two categories): $\chi^2_c = 7.847$.

The calculated result exceeded the critical value of χ^2 at the 5 % level for a table with 3 degrees of freedom. The null hypothesis was rejected in favour of H_A and it was concluded that programme dependence was present; selection of loudspeaker models was influenced by the extracts used to evaluate them. Further analysis was conducted treating the original data set as two separate samples, one for each collapsed extract group. The following notation will be used to identify the separate groups elsewhere:

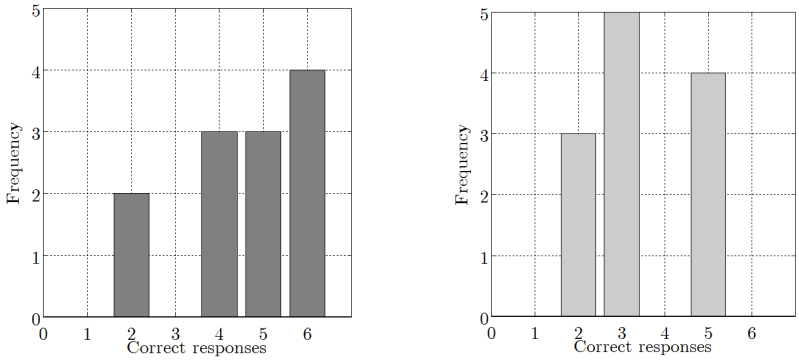
$S_{1,2}$: Extracts CR and OC; $S_{3,4}$: Extracts PN and TM.

7.4.2 Post-Screening (Intra-Listener Performance)

Every trial in listening test III could be considered as a hidden reference comparison as listeners were always directly comparing the reference (z_2) against itself without knowing. Individual listener responses were inspected after collating across extract groups $S_{1,2}$ and $S_{3,4}$. Figure 61 shows the distributions, having the following statistics:

Extracts $S_{1,2}$ (Fig. 70a): $n = 6$ (number of trials); $c_{\min} = 2$; $c_{\max} = 6$;
mean = 4.6; median = 5; mode = 6.

Extracts $S_{3,4}$ (Fig. 70b): $n = 6$; $c_{\min} = 2$; $c_{\max} = 5$;
mean = 3.4; median = 3; mode = 3.



(a) Extracts $S_{1,2}$ **(b)** Extracts $S_{3,4}$

Figure 70: Distribution of correct responses from Group III hidden reference trials; data from 12 participants, $n = 6$

The performance of individual listeners differed across the two extract groups; this had been expected, based on the findings from analysis of programme dependence. The distributions indicated that extracts $S_{3,4}$ made detection of differences between the two models in a given pair very difficult, perhaps impossible. If true, even the most critical listeners could not be expected to correctly identify the hidden reference. Therefore, only listener performance in trials containing extracts $S_{1,2}$ were used for post-screening.

A performance limit for post-screening was established using the exact binomial distribution; to meet a significance level of 2.5 %, the minimum required number of correct hidden reference responses was 6 out of 6:

$$\alpha = 0.025; \qquad P\,(c \geq 6) = 0.0156 \text{ (4 d.p., one-tailed)}$$

Of all 12 participants, only 4 correctly identified the hidden reference in every trial. The remaining sample size per pair after post-screening at this level was extremely small, $n = 8$, due to the separation into two extract sets. As a compromise for performance level against remaining sample size, a lower post-screening threshold was selected, requiring 5 out of 6 correct responses per listener when evaluating models with extracts $S_{1,2}$. Using the exact binomial distribution, $P\,(c \geq 5) = 0.1094$ (4 d.p., one-tailed). The post-screened data set therefore contained responses from 7 of the original 12 participants.

7.4.3 Pairwise Results Based on Total Sample (All Listeners)

The individual pair analysis was conducted as two groups, collated by the collapsed extract categories, $S_{1,2}$ and $S_{3,4}$. All evaluations in listening test III were direct comparisons; the reference X was always model z_2, and every A/B pair contained this model. Therefore, there was always an a-priori correct response; the expected result was in a specified direction (reference z_2

expected to get the majority of responses within a pair) and a one-tailed hypothesis test was used.

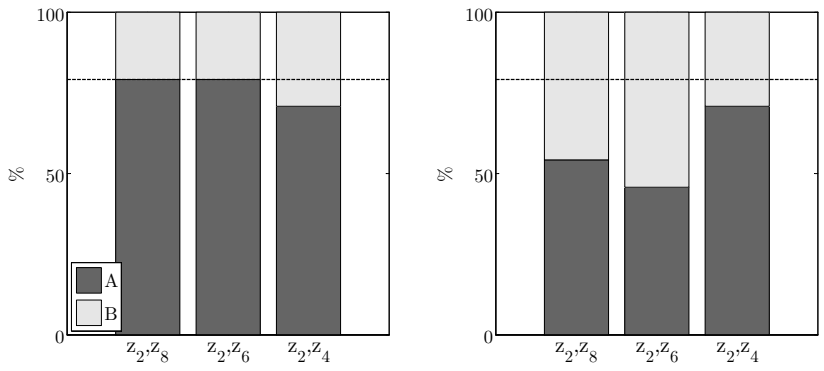
Due to division of the data into the two extract groups, sample size for each pair was $n = \left(\frac{M}{2}\right) L = 24$ responses. The critical count threshold was determined using the exact binomial distribution:

H_0: $p_{z_2} = p_{\tilde{z_2}}$; proportion of responses is the same for hidden reference and other model (not z_2). Conclude that a difference between them was not audible.

H_A: $p_{z_2} > p_{\tilde{z_2}}$; proportion of responses for hidden reference is greater than for the other model. Conclude that an audible difference exists between A and B, and listeners were able to identify which one was the hidden reference.

$\alpha = 0.025; \quad G = 3; \quad \alpha_\mathrm{func} = 0.0084$, one-tailed.
$c' = 19$; hidden reference must receive at least c' votes (79.2 % of responses for $n = 24$) within a pair for the split to be considered significant; H_0 rejected in favour of H_A.

Figure 71 and Table 15 summarise the results of pairwise analysis for the data from listening test III. The presentation format is consistent with that in previous sections. All p-values show $P\left(c \geq c_\mathrm{A}\right)$, the probability of reaching at least c_A counts by chance, i.e. by randomly guessing in the absence of an experimental effect; values less than α_func are considered statistically significant at the 2.5 % level.



(a) Extracts $S_{1,2}$ **(b)** Extracts $S_{3,4}$

Figure 71: Group III pair results; data from 12 participants. The hidden reference (z_2) is represented by the label A. Horizontal dotted lines mark the critical count thresholds; if the split of counts between A and B bars meets or lies above the threshold line marked at 79.2 %, the result is considered significant at the 2.5 % level

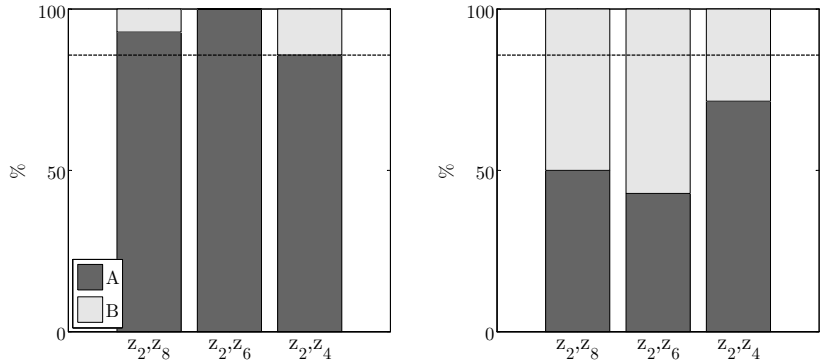
	Extracts $S_{1,2}$			Extracts $S_{3,4}$		
A	19	19	17	13	11	17
B	5	5	7	11	13	7
p	0.0033	0.0033	0.0320	0.4194	0.7294	0.0320
pair	z_2,z_8	z_2,z_6	z_2,z_4	z_2,z_8	z_2,z_6	z_2,z_4

$n = 24$; $\alpha_{\text{func}} = 0.0084$ ($\alpha = 0.025$, $G = 3$)

Table 15: Group III pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.4.4 Pairwise Results Based on Post-Screened Sample

Based on the intra-listener evaluation in section 7.4.2, pair analysis was repeated with the data from seven participants who were identified as performing the listening task with greater accuracy and consistency. The analysis was identical to that in section 7.4.3 except that the smaller sample size, $n = 14$, increased the critical threshold. For all pairs, $c' = 12$ (85.7 %). The results are summarised in Figure 72 and Table 16.



(a) Extracts $S_{1,2}$ **(b)** Extracts $S_{3,4}$

Figure 72: Group III post-screened pair results; data from seven participants. The hidden reference (z_2) is represented by the label A. Horizontal dotted lines mark the critical count thresholds; if the split of counts between A and B bars meets or lies above the threshold line marked at 85.7 %, the result is considered significant at the 2.5 % level

	Extracts $S_{1,2}$			Extracts $S_{3,4}$		
A	13	14	12	7	6	10
B	1	0	2	7	8	4
p	0.0009	0.0001	0.0065	0.6047	0.7880	0.0898
pair	z_2,z_8	z_2,z_6	z_2,z_4	z_2,z_8	z_2,z_6	z_2,z_4

$n = 14$; $\alpha_{\text{func}} = 0.0084$ ($\alpha = 0.025$, $G = 3$)

Table 16: Group III post-screened pair results. Shaded columns in the results tables highlight pairs where the critical count was not reached. Values for p and α_{func} are given to 4 d.p.

7.4.5 Listener Comments

Some general comments from participants after they had completed the experiment are summarised here. Although this questioning was not formally conducted as part of the experiment, the feedback is relevant to the findings from the data analysis.

It was hard to detect differences between the models in any pair. Differences were easier to detect with certain extracts, but the nature of the difference was hard to describe; it was not as simple as an overall loudness difference or change in particular parts of the spectrum, e.g. bass boost. The impression seemed to be a change in relative levels of individual instruments. It appeared that for tracks OC and CR, group $S_{1,2}$, the resonant bass content became louder or more forward, giving the impression of a greater difference between mid-/high- and low-frequency instruments, e.g. vocals/guitar and bass. This observation indicates that the experimental models in Group III were replicating the effect illustrated in section 4.5.1 and discussed in section 1.1.3.3, where increasing amounts of phase distortion through the bass region degrade performance in the time domain and change the perceived balance between instruments in the rhythm section. The general consensus from listeners was that differences using the pink noise extract, PN in group $S_{3,4}$, were nearly always imperceptible, though two participants said that they actually found it easier to detect differences using this extract. The experimental data had been anonymised to ensure confidentiality, so it was not possible to check whether these participants were actually correct in their identification of the hidden reference. No particular comments were made regarding extract TM in group $S_{3,4}$. As described in Table 9, this was a relatively complex arrangement with a fast, tight, and punchy bassline. The informal comments and formal pairwise results indicate that phase shifts in a loudspeaker's low-frequency alignment will have little impact on the audible impression when reproducing this type of programme material. Based on the characteristics of extracts $S_{1,2}$, music featuring simpler band arrangements with a slower, slightly resonant rhythm section is highly susceptible to phase distortion of the orders encountered in typical monitoring loudspeakers. In this experiment, the perceived change in instrument balance could be detected because an undistorted reference was present for comparison. In a real-life monitoring situation, an engineer would not be able to make this kind of immediate and direct comparison, and would therefore not realise that they were mixing on a distorted system.

7.4.6 Listening Test III Results Summary

Findings from listening test III are summarised below, grouped according to the separate stages of analysis.

7.4.6.1 Programme Dependence The distribution of results by musical extract showed differing proportions of listener responses. A high number of expected counts below 5 in the initial χ^2 crosstabulation forced the collapse of data categories into two groups. Two extract pairs were formed, based on an apparent natural grouping within the original data. The null hypothesis was rejected at the 5 % level in the subsequent χ^2 analysis ($\chi^2 = 7.847$, d.f. $= 3$). It was concluded that programme dependence was observed in the data; the two extract groups elicited different responses from listeners about the same loudspeaker models. Further analysis was conducted treating the data as two separate groups; responses from extract CR and OC were kept separate from those returned by extracts PN and TM, groups $S_{1,2}$ and $S_{3,4}$ respectively.

7.4.6.2 Intra-Listener Performance All trials in this experiment were classed as containing a hidden reference, as the reference z_2 was always compared against itself and one other model. Intra-listener performance in six trials was viewed separately for extract groups $S_{1,2}$ and $S_{3,4}$. The results indicated that extracts $S_{3,4}$ might not be revealing of differences between the models even in a direct comparison; therefore, only data from extracts $S_{1,2}$ were used for analysis. The number of correct hidden reference responses with these extracts ranged between $c_{\min} = 2$ and $c_{\max} = 6$. Only 4 out of 12 participants performed at the preferred accuracy level: $p < 0.025$, $c'_R = 6$. This threshold was not used for post-screening as the resulting sample size was deemed to be unacceptably small. A reduced performance threshold was therefore adopted: $c'_R \geq 5$, $p = 0.1094$. Seven of the twelve listeners correctly identified the hidden reference this many times.

7.4.6.3 Pairwise Analysis The pairwise analysis was first conducted using data from all participants (no post-screening). Responses from extract groups $S_{1,2}$ and $S_{3,4}$ were analysed separately. The split of listener votes within each model pair was compared against a critical count threshold: $c' = 19$ (one-tailed, $\alpha = 0.025$). Sample size was 24 listener responses per pair, making the critical threshold equal to 79.2 % of responses. Results supported the hypothesis from the investigation of intra-listener performance; H_0 was not rejected in any pair evaluated with extracts $S_{3,4}$. The null hypothesis was rejected in two out of three pairs when evaluated using extracts $S_{1,2}$. It was concluded that the majority of listeners could not detect an audible difference between the reference z_2 and model z_8, but could otherwise detect differences when extracts $S_{1,2}$ had been used to make the comparisons.

7.4.6.4 Post-Screened Pairwise Analysis The pairwise analysis was repeated using the post-screened data set; this reduced sample size per pair to $n = 14$. The critical threshold therefore increased relative to n: $c' = 12$ (one-tailed, $\alpha = 0.025$), or 85.7 %. The increased threshold meant that listeners had to show a greater consensus within each pair to return a significant result. Despite the increased threshold, H_0 was rejected in all three pairs when evaluated with extracts $S_{1,2}$, but in none when extracts $S_{3,4}$ had been used. The most critical listeners were better able to detect differences with extracts $S_{1,2}$, but did not demonstrate that

differences were any more audible with extracts $S_{3,4}$ compared to the original analysis. It is concluded that, under highly sensitive listening conditions, and in the absence of accompanying magnitude variations, phase distortion in a loudspeaker's low-frequency alignment is imperceptible with some types of programme material but detectable with others.

7.5 Summary and Discussion of Subjective Evaluation

Three sets of listening test were performed, one for each group of loudspeaker models. The aim was to gather subjective data for comparison with MTF results derived from the same models. Three aspects were of particular interest when analysing the subjective data: identifying which pairs gained a significant consensus from listeners, investigating whether any musical extracts seemed to make the decision more or less difficult, and, finally, whether the pair judgements were any different for listeners who performed well in hidden reference trials.

The experimental method was fundamentally the same for all experiments; listeners were presented with two loudspeaker models at a time, A and B, and forced to make a judgement about which was most like an 'ideal' reference, X. The assumption was that selection of A or B was a vote in favour of the one deemed to reproduce musical bass content most accurately. Apart from featuring different loudspeaker models, each experiment used different musical extracts; within an experiment their duration was the same, but this length was reduced across the three tests. The number of repeats performed by participants was not formally monitored, but reduction of duration did not appear to make it any harder for participants to form their judgements; it had the advantage of considerably reducing overall session duration, as a large number of trials are required to make all comparisons in this type of experiment. Only two extracts were used in listening test II, compared to three in the first experiment; the motivation was to reduce the number of trials per listener so that the test could be completed within a single listening session. It was believed that more people would take part if they only had to commit to one session. Although more participants did complete the second experiment, 26 compared to 23, there were not enough to compensate for the reduction in trials per listener from repeating by another extract; the second experiment therefore had a smaller sample size than the first. Based on this experience, it is concluded that a better approach would have been to maintain the number of trials per listener but make the duration of each trial shorter by using shorter extracts; participants would have needed to attend two sessions but the duration of each would have been reduced, therefore making their commitment less onerous.

Listening tests I and II were the primary focus of the study. These evaluated virtual loudspeakers: models with different low-frequency alignments, representative of responses that might be seen in professional mixing monitors. Listening test III was a secondary investigation looking at whether listeners could detect phase distortion alone. This evaluated artificial loudspeaker models with a fixed magnitude but different low-frequency phase responses, approximating the range of behaviours observed in typical monitor design strategies. The MTF assessment of these models had shown differences between them, but the overall score changed very slightly, in the order of 0.001. If it was shown that changes of this size were perceptible, it would indicate that the observed MTF behaviour was meaningful in relation to subjective impression.

The overall findings from subjective evaluation are summarised and discussed in relation to the original aims of analysis defined in section 6.1.

1. Were results influenced by any particular programme material?
In listening tests I and II it was expected that the characteristics of the chosen musical extracts were similar enough to produce a consistent decision, e.g. a listener's judgement for loudspeaker A with extract x would be the same as when evaluated using extract y.

Extract selection for listening tests I and II appeared to have been successful; following a χ^2 analysis, it was concluded that neither set of experiments was affected by programme dependence. It is possible that the extracts used in listening test I were more revealing of differences between the loudspeaker alignments because this experiment returned clear outcomes for a greater number of pairs. However, it is not possible to state with any certainty whether this is true, or whether the discrepancy was actually due to the more complicated differences between alignments within the Group II models. The latter is considered more likely.

Selection of extracts for listening test III was also considered successful. This was the only set of experiments where material with differing temporal and spectral characteristics was specifically chosen, with the expectation that some variation in listener responses might be seen. Using extracts with different temporal qualities was deemed especially important for this experiment, as the magnitude response of the the Group III models was fixed; they only varied in their amount of low-frequency phase distortion. The χ^2 analysis showed that these extracts did affect listener judgements differently. This was further confirmed in the pair analysis. It was seen that audibility of low-frequency phase shifts, representative of those encountered in common loudspeaker designs, was strongly material-dependent; mid-tempo music containing simple arrangements and slightly resonant bass content was most revealing of differences. It should be noted that this conclusion is based on one reproduction level, and phase shifts in one frequency range. However, it is interesting to compare this finding with the study reported by Fazenda *et al.* [7] who, coincidentally, appeared to use extracts of very similar characteristics to those used here and found similar results; although that study was looking at low-frequency decay times in studio control rooms, it was demonstrated that as well as sufficient low frequency energy, musical extracts chosen for this type of investigation must also have sufficient temporal gaps between bass notes and a degree of transient and sustained sounds. That conclusion is fully corroborated by the findings from listening test III. Reference to Figure 53 in section 5.3.3.3 adds further support to this finding; the power spectrum of extract TM was very similar to that of extracts OC and CR ($S_{1,2}$), so inspection of the signal content alone could not have indicated that TM would fail to be revealing of phase distortion in the bass region. Based on this evidence, it is concluded that temporal characteristics must be considered as carefully as adequate bass content when selecting suitable programme material for any further investigation of the MTF algorithm in relation to studio monitoring.

2. Were statistically significant differences found between all tested pair combinations?
All three experiments returned a majority of pairs where the null hypothesis was rejected, giving clear outcomes for direct comparison with the MTF data. Listening test II returned a larger proportion of non-significant pairs compared to the first set of experiments. This was not

surprising, given that the Group II models were designed to have more subtle alignment differences (as described in section 4.3). Failure to reach a significant result in any pair could be explained by one of the following:

i) A and B were audibly identical.

ii) A and B were audibly different but it was difficult to choose which one sounded overall more like X.

iii) Participants formed a consensus but there were too few trials to reach a confident conclusion (an effect is present but not strong enough for the given sample size).

It was considered possible, but highly unlikely, that i) was the reason for all of the pair results which failed to return a clear directional outcome. As described in sections 7.2.5.4 and 7.3.5.4, a trend was observed where the proportional A/B split of results increased compared to the same pairs before post-screening; the more acute listeners were in better agreement as a group about their judgements. Of the 30 pairs evaluated across listening tests I and II, 25 showed this trend (83 %). Changes in the split of results before and after post-screening were therefore treated as potential evidence supporting ii), with pairs where the A/B split *reduced* after post-screening being most interesting. This happened in two pairs from listening test II:

EF – split reduced from 63 to 54 % in favour of F after post-screening.

EG– split reduced from 73 to 71 % in favour of G after post-screening.

If failure to reach a significant result was always due to option i), these results imply that the otherwise more acute listeners became worse at detecting differences in these particular pairs, contradicting the trend observed in the majority of other comparisons. Based on the observations from post-screening across both experiments, it is plausible that the reduction in split of results is due to the more acute listeners detecting differences between A and B but being more divided as a group about which was more like X. It must be acknowledged that lack of certainty about this issue is due to a limitation of the experimental method; there was no option for participants to respond with 'not sure'. It is believed that a forced-choice strategy was the best option in this study, as the judgements were clearly not always straightforward for many participants; the number of pairs failing to reach a clear directional outcome in listening test II is taken to be evidence in support of this statement. However, it would have been interesting to be sure whether these outcomes were simply due to inaudible differences; it is relevant to the research topic to know if two different alignments reproduce musical signals that are audibly identical. An alternative would have been to retain the 2AFC strategy but require participants to record that their answer was a guess, a recommended approach in sensory (food) testing standards [145, 146]. In this study, such an approach would not have been sufficient; the 'additional information' would need to take the form of:

1) Guess; no audible difference.

2) Guess; audibly different.

This could be further extended to:

1) Guess; no audible difference.

2 a) Guess; audibly different, both very similar to X.

2 b) Guess; audibly different, both very different from X.

Including these options would have complicated the test GUI and potentially made the task more difficult for participants. A different approach might have been used whereby models C–G were evaluated in direct comparisons, similar to the hidden reference trials. This would have permitted a firm conclusion about audibility to be drawn in every pair comparison; any pair not returning a significant result would be excluded from further testing under the conclusion that they were not sufficiently different from each other. Subsequent testing with the remaining models could then be made with confidence that listeners failing to reach a consensus in a given pair was due to lack of agreement about which sounded most like the reference. This is considered to be an inefficient approach due to the high number of trials required to compare all combinations of models directly. Perhaps the best alternative solution would be to combine these options into a two-stage response procedure; the first question presented on the test interface would require listeners to respond whether A or B are the same or different; if they replied 'same', the next trial would begin. If they responded 'different', the screen used in this study would be presented, where they had to choose which of A or B was most like X. Lack of agreement between listeners in answer to this question must then be due to a difference in opinion about similarity to X rather than inaudibility between the models. This would have made data analysis more complicated, but potentially more useful.

Given the small sample sizes and lack of unanimity between listeners in most pairs, maintaining a low Type I error risk was considered a priority in this study. This reduced the chance of wrongly rejecting H_0 and thus, incorrectly concluding a significant experimental effect. However, it is considered likely that Type II errors may have been committed (incorrectly failing to reject the null hypothesis) as all three experiments contained pairs showing evidence of iii). The most notable example is pair FG after post-screening; although the proportional split of A/B results increased compared to pre-screening, maintaining the same criteria for analysis meant that the critical threshold increased with the reduction in sample size and a significant result was no longer returned.

3. Were results any different for the more critical listeners?

Analysis of performance in hidden reference trials appeared to be successful in identifying the most acute listeners. Variability initially seen in the paired responses reduced, indicating improved intra- and inter-listener consistency; the most accurate listeners were overall in better agreement with themselves and each other. The effects of post-screening on pairwise results were partially discussed in the preceding point. After comparing findings pre- and post-screening in all three experiments, it was concluded that the assumption underlying the hidden reference analysis was valid. If a listener was able to consistently identify a loudspeaker when compared 'blind' against itself and another stimulus, they would be more discerning in auditory evaluations where there was no direct comparison. Despite the apparent inefficiency introduced by inclusion of a hidden reference, these direct comparison trials were seen to be a very useful feature of the experiments. The method was considered highly effective, and it is suggested that it could be used before similar experiments as a reliable way to select participants. This might also help to avoid the problems seen in all three experiments, where the initial data set reduced in size considerably after post-screening because many participants were not performing at the chosen

level of accuracy.

It should be noted that participants selected through hidden reference performance were considered to be critical listeners only within the context of the experiment. It was not assumed that a person's ability to accurately detect differences between two loudspeaker alignments would make them capable of performing a detailed evaluation of bass reproduction accuracy. In the proposed application of the MTF algorithm, the target audience is professional sound engineers; these listeners will have developed critical listening abilities over many years of mixing experience, and will therefore be skilled in making this type of critical judgement. The number of non-significant pair results returned in listening test II seems to support this hypothesis. The post-screened data showed that the higher-performing listeners were accurate in direct comparisons (the hidden reference trials), and agreed more closely as a group in the majority of pair judgements where neither A nor B was identical to the reference; conversely, in some pairs their data showed no more consensus than was observed before post-screening. It was suggested that this apparent contrast in behaviour was due to these listeners being able to detect differences between the two loudspeakers, but unable to consistently decide which one was reproducing bass content most accurately overall.

Discussion of the findings from subjective evaluation in relation to the MTF results is presented in chapter 8.

8 Assessment of the MTF-Based Method for Loudspeaker Evaluation

Chapter 7 presented the findings from subjective evaluation of the experimental loudspeaker models. This chapter compares the results with the outcomes of the objective assessment, shown in chapter 4, to consider the efficacy of the proposed method in more detail. Section 8.1 considers performance of the MTF algorithm in relation to the significant pairwise results and partial rankings of experimental loudspeakers obtained from listening tests. In section 8.2 the comparison is to extended pairs where a significant result was not returned through subjective assessment. This informs section 8.3, where two types of adjustment are reviewed that might be applied to the algorithm results when considering its ability to predict subjective judgements. Finally, section 8.4 looks at features of the method compared to some alternatives measures that might be used for evaluation in the target application.

8.1 Comparing Significant Pair Results (Listeners vs Algorithm)

8.1.1 Pairwise Outcomes

Initial comparison was performed with the numerical results, i.e. scores for unweighted overall mean modulation index, $\overline{M}$. The majority of tested pairs returned clear directional outcomes that were used in this comparison; the following summary is based on the post-screened data:

Listening test I: 12/15 pairs (80 %; 4 ABX, 8 PCwR);

Listening test II: 9/15 pairs (60 %; 5 ABX, 4 PCwR);

Listening test III: 3/3 pairs with extract group $S_{1,2}$,
0/3 pairs with extract group $S_{3,4}$ (all ABX);

where: ABX denotes direct comparison trials (one of A or B identical to X), and PCwR denotes paired comparison with reference (A and B different from X). Table 17 summarises the listening test results and compares them alongside the corresponding algorithm output. This initial comparison only considers the statistically significant pair results. The generic labels A and B have been used to indicate which models are under comparison, where A represents the system to the left of the inequality sign, and B the system on the right. The symbol $>$ separating them signifies that a statistically significant result was returned in this pair, where A was judged as sounding closer to the reference model than B at the given significance level; recall from chapter 7 that α was 0.025 in hidden reference comparisons, i.e. pairs containing R or z_2, and 0.05 otherwise. As described in section 5.2.6, it was assumed that if model A was deemed as sounding closer to the reference model, it would be expected to return a higher value for $\overline{M}$ than produced by model B.

Only the post-screened data for each set of listening tests has been shown here; as discussed in chapter 7, data from participants falling into this category generally showed less inter-listener variability and, in some cases, a greater ability to discriminate between the virtual loudspeakers. Comparison with results from the unscreened data are listed in Appendix J. Using the full data set for the comparison in this section does change any of the conclusions presented here except

for one pair, FG in Group II, and that is discussed in section 8.2.2. All values for $\overline{M}$ are shown to 3 d.p. as it was shown in section 4.5.4 that this precision was required to differentiate between the systems in Group III.

A>B	$\overline{M}_A$	$\overline{M}_B$	$\overline{M}_\Delta$
R>C	0.939	0.795	0.144
R>D		0.419	0.520
R>E		0.656	0.283
R>G		0.692	0.247
C>D	0.795	0.419	0.376
C>E		0.656	0.139
E>D	0.656	0.419	0.237
F>D	0.893	0.419	0.474
F>E		0.656	0.237
F>G		0.692	0.201
G>D	0.692	0.419	0.273
G>E		0.656	0.036

(a) Listening test I (Group I models)

A>B	$\overline{M}_A$	$\overline{M}_B$	$\overline{M}_\Delta$
R>C	0.845	0.576	0.269
R>D		0.775	0.070
R>E		0.645	0.200
R>F		0.706	0.139
R>G		0.696	0.149
D>C	0.775	0.576	0.199
E>C	0.645		0.069
F>C	0.706		0.130
G>C	0.696		0.120

(b) Listening test II (Group II models)

A>B	$\overline{M}_A$	$\overline{M}_B$	$\overline{M}_\Delta$
$z_2 > z_4$	0.758	0.755	0.003
$z_2 > z_6$		0.752	0.006
$z_2 > z_8$		0.747	0.011

(c) Listening test III (Group III models)

Table 17: Comparison of listening test significant pair results (post-screened data, column $A > B$) with corresponding algorithm mean scores. The difference between $\overline{M}_A$ and $\overline{M}_B$ is given by $\overline{M}_\Delta$

From the results presented in Table 17, it is seen that all significant pairwise results in all three listening tests were predicted by the mean algorithm scores alone. For each pair, the difference in mean scores, $\overline{M}_\Delta$, was inspected. For any pair returning a statistically significant result, it was concluded that differences between the loudspeaker models were audible; inspection of the minimum $\overline{M}_\Delta$ values in these pairs therefore shows the smallest change in mean score between two models that coincided with a conclusion of perceptible difference in the subjective experiments. Based on the results listed in Table 17 the smallest values for $\overline{M}_\Delta$ are as follows:

Group I: $\overline{M}_{\Delta} = 0.036$ (G vs E);

Group II: $\overline{M}_{\Delta} = 0.069$ (C vs E);

Group III: $\overline{M}_{\Delta} = 0.003$ (z_2 vs z_4);

These results show that the majority of listeners could discriminate between loudspeakers with low-frequency alignments that returned very small differences in $\overline{M}$. For comparison, Table 18 shows the equivalent differences in $\overline{M}$ for model pairs in Groups I and II where no clear outcome was returned, defined here as meaning that it was not concluded with the chosen level of confidence that listeners had chosen A over B, or vice versa; there was no directional outcome for direct comparison with the algorithm results

Inspection of the values in Table 18 shows that six of the nine non-directional pair results returned $\overline{M}_{\Delta}$ equal to or greater than the values observed to correspond with a perceptible difference. This supports the conclusion stated in section 7.5; failure to reach a significant result within a given pair was not always simply due to imperceptible differences between the loudspeakers, but reflects an inability to select which model sounded most like the reference. However, discrimination of differences is likely to be affected by a number of factors that the algorithm does not account for, such as individual hearing acuity and musical content. The results in pairs where the null hypothesis was not rejected are discussed further in section 8.2.

AB	$\lvert\overline{M}_{\Delta}\rvert$
RF	0.046
CG	0.103
CF	0.098

(a) Group I models

AB	$\lvert\overline{M}_{\Delta}\rvert$
DE	0.130
DF	0.069
DG	0.079
EF	0.061
EG	0.051
FG	0.010

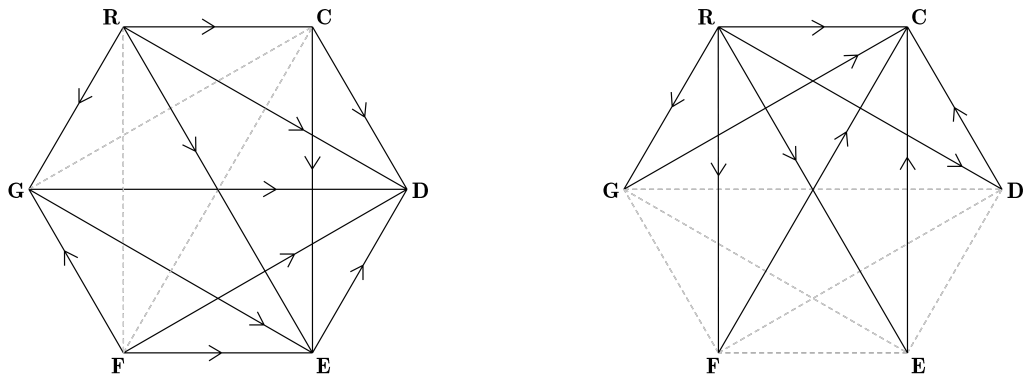
(b) Group II models

Table 18: Difference in $\overline{M}$ scores for pairs with no directional outcome; Group I and II models based on post-screened data

8.1.2 Indirect Ranking

An indirect ranking method was used to investigate how the objective ranking of Group I and II loudspeakers compared to the subjective results. As mentioned in section 5.2.2, paired comparisons can be used to produce a rank ordering of a group of items. A method of geometrical representation developed by Kendall *et al.* [166, 204] was applied to the results summarised in Table 17a and 17b. The technique was developed for analysis of preference testing, and is designed for cases where all possible pair comparisons are made, as was the case in listening tests I and II, but usually assumes that all pairs have a defined outcome; as this condition was not always met in the listening test data considered here, the technique only permitted a partial ranking to be constructed.

The preference diagrams, or 'diagraphs', as they are sometimes called [205], are shown in Figure 73. The line between each pair of models shows that a comparison was completed. Dashed lines have been used here to show that no significant result was returned, i.e. neither A>B nor A<B could be concluded. A black solid line with an arrow shows that a significant result was returned, and the arrow indicates the direction, e.g. A→B indicates that A was selected over B. Note that the term diagraph is used in this discussion as the data was not derived from preference judgements, though they might be described as 'significance diagrams' in this context.



(a) Listening test I **(b)** Listening test II

Figure 73: Diagraphs for all pair results in listening tests I and II (post-screened data). Dashed line shows no significant result for that pair. Solid line shows a significant result, with arrow indicating the direction of the conclusion

Construction of the diagraphs permits simple investigation of the presence of circular triads; this is a form of inconsistency, known as intransivity, within the paired results that cannot be detected when performing direct rankings. This takes the the form:

$$X>Y, \quad Y>Z, \quad Z>X$$

Inspection of the diagrams in Figure 73 shows that, based on the significant pair results, no such inconsistencies were present within the listening test data for either set of experiments. This may be an indication that the judgements were not influenced by personal preference, i.e. participants responded correctly to the experimental question and based their comparisons only on perceived similarity to the reference model, but it is not possible to conclude this with certainty from the existing data.

Partial indirect rankings for each set of loudspeaker models were developed by counting the arrows in the diagraphs. An arrow pointing out from a model indicates that it was selected as sounding more like the reference than the model the arrow points towards. Therefore, counting the number of arrows pointing towards each system was used as a measure of where they should be ranked relative to other models in the group. Table 19 shows the results of this analysis. The partial rankings are shown in Table 20 beside the equivalent position based on mean algorithm score for each system; the exact values were presented in Table 17.

Inward arrow count	Models
0	R, F
1	C
2	G
3	-
4	E
5	D

(a) Listening test I (Group I virtual loudspeakers)

Inward arrow count	Models
0	R
1	D, E, F, G
2	-
3	-
4	-
5	C

(b) Listening test II (Group II virtual loudspeakers)

Table 19: Arrow counts derived from diagraphs

Rank	Subjective	Objective
1	R F	R
2		F
3	C	C
4	G	G
5	E	E
6	D	D

(a) Listening test I (Group I)

Rank	Subjective	Objective
1	R	R
2	D E F G	D
3		F
4		G
5		E
6	C	C

(b) Listening test II (Group II)

Table 20: Comparison of algorithm and listener-derived loudspeaker rankings

Clearly, the lack of significant results in listening test II produced a partial ranking where only the first and last positions are defined. The higher proportion of significant results returned by listening test I allowed a clearer ranking to be defined, with only one tied rank. It can be seen from Table 20 that where defined ranks exist in the ordering derived from subjective evaluation, the algorithm predicts them. The relatively large number of pairs failing to return a significant result in listening test II prevented development of a more conclusive ranking. Section 8.2 looks at these pairs in more detail to see whether the objective results were useful in understanding why a greater consensus between listeners could not be reached.

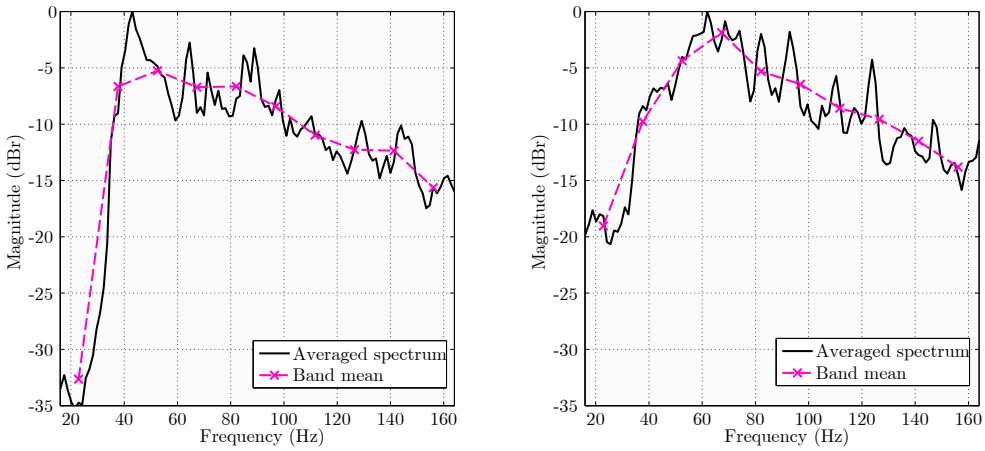
8.2 Comparison With Non-Significant Pair Results

Both listening tests I and II returned pairs where the null hypothesis was not rejected, i.e. the split of results in favour of one or other loudspeaker was insufficient to be confident that a consensus between listeners was present. Some of the results seemed surprising, as comparison of the mean algorithm scores had shown that models with smaller differences in $\overline{M}$ had returned a conclusive result.

The objective results were compared with subjective data for loudspeaker pairs where a significant result had not been reached; as for the analysis in section 8.1, results from the

post-screened data were used but that does not alter the key conclusions presented here. The numerical output of the MTF algorithm allowed all models within each experimental group to be ranked, even those that had alignments where the differences were not a simple overall reduction in output level at low frequencies. It seemed that some listeners found comparing the more complex alignments difficult, so it was necessary to inspect the detail within the results matrix to understand whether it showed contrasting behaviour between a given pair of models that might have led to the conflicting decisions across listeners.

The comparison was performed with consideration of signal content i.e. the relative spectral balance of the musical extracts used for evaluation. For each listening test, the power spectra for all extracts were averaged; the individual spectra are not shown here as they were presented in Figs. 51 and 52. The mean level was then calculated for each band defined in the algorithm. The results showed relative levels for signal content in each analysis band, as presented in Figure 74. This allowed investigation of whether the loudspeaker models in a given pair were showing contrasting behaviour in bands that contained a particular excess or absence of signal content.



(a) Listening test I. Averaged spectrum $L_{M1}(f)$ based on three extracts

(b) Listening test II. Averaged spectrum $L_{M2}(f)$ based on two extracts

Figure 74: Band-mean levels for averaged extract spectra used in listening tests I and II

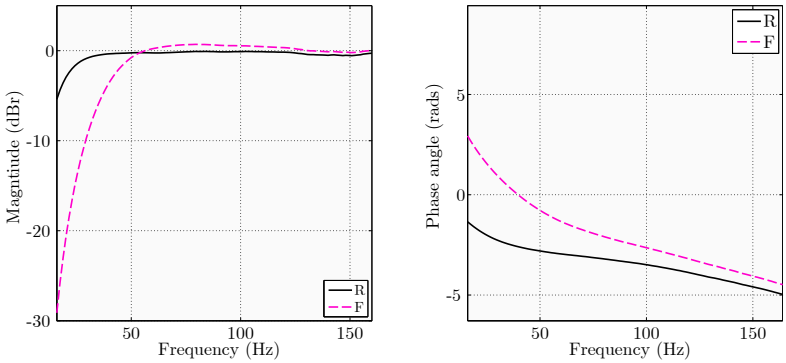
A summary of the analysis is presented in sections 8.2.1 and 8.2.2. Conclusions here were based initially on inspection of the intensity images, shown in Figs. 42 and 45, then confirmed using the numerical matrix scores, listed in Appendix G. Comparison of each pair was summarised after inspecting the subjective results, differences in algorithm output, and consideration of the relative signal level in frequency bands where differences occurred. Only two non-significant pairs for each listening test are shown here as they present interesting comparisons, and the general trends observed were consistent across the other pairs. Summaries for the remaining pairs are given in Appendix K (one for listening test I, four from listening test II).

8.2.1 Two Pairs from Listening Test I

Figures 75 and 76 show the steady state magnitude and phase responses for the pairs being compared. A linear frequency axis has been used to reflect the algorithm band definition, with

limits covering the extent of the MTF analysis range.

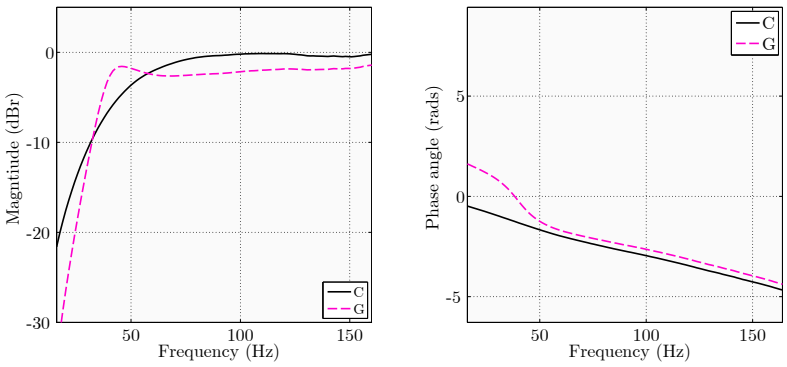
Figure 75: Group I: R vs F. Magnitude and phase



Pair	$\overline{M}$scores	Subjective result
R vs F	$\overline{M}_R > \overline{M}_F$ (0.939>0.893)	Tends towards R, selected in 28 of 45 trials, 62 %.
Summary:	$\overline{M}_\Delta$ reduces to 0.007 if band 1 is excluded; signal content here is approximately 17 dB lower than the mean level in any other band.	
Conclusion:	Very low signal content in the region where these two systems primarily differ made audible discrimination difficult.	

Pair RF was the only case from either listening test where primary differences between models coincided with an extreme lack of signal content, rather than a peak. The MTF indicates that R will reproduce low-frequency content more accurately than F, but the frequency bands where the greatest differences between them occurred did not contain enough signal content for the differences to be sufficiently compared by listeners. Therefore, this may be one pair where the majority of participants genuinely could not detect a difference between the loudspeakers they were evaluating.

Figure 76: Group I: C vs G. Magnitude and phase



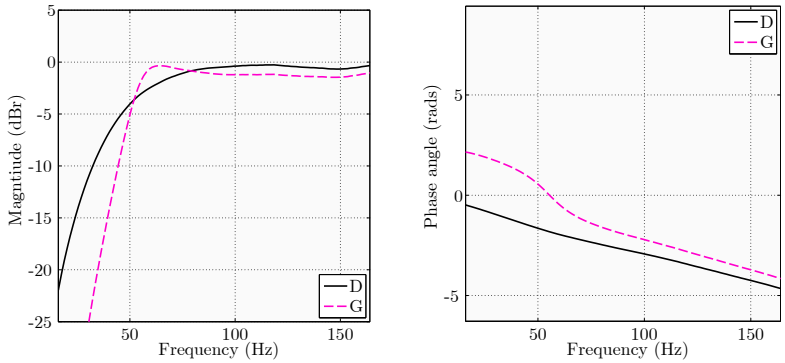
Pair	$\overline{M}$scores	Subjective result
C vs G	$\overline{M}_C > \overline{M}_G$ (0.795>0.692)	Almost equal distribution; C selected in 23 of 45 trials, 51 %.
Summary:	C returns higher m values in all matrix locations except for bands 2 and 3 ($\overline{M}_{C,2:3} = 0.576$; $\overline{M}_{G,2:3} = 0.654$). Music spectrum contains most energy in bands 2 and 3, with a peak at 43 Hz that is 2.8 dB higher than any other band.	
Conclusion:	Very strong signal content over two frequency bands divided participants between G, the loudspeaker with greater output due to a resonance in this range, and C, the one with MTF performance most like the reference elsewhere.	

For pair CG, it is difficult to judge which of these systems reproduce bass content most accurately based on the magnitude response alone; the phase response shows a large shift in G below 50 Hz, due to the system resonance around 45 Hz. The increased magnitude of model G between 32 and 58 Hz might be interpreted as an advantage over the lower output of C in this range, and this increase at lower frequencies will have been especially obvious to listeners as it coincides with frequency bands in the signal where there is lots of bass to reproduce. Without a thorough understanding of the different alignment shapes and their impact on transient response, it would not be understood that this bass boost is provided by an underdamped resonant element that is likely to distort the balance of the rhythm section instruments that have their fundamental frequencies in this range. The detail of the MTF matrix reflects the greater output level in G due to the resonance but returns a mean score showing it to be less accurate in reproducing bass content overall than loudspeaker C.

8.2.2 Two Pairs from Listening Test II

Figures 77 and 78 shows the steady state responses for the pairs being compared.

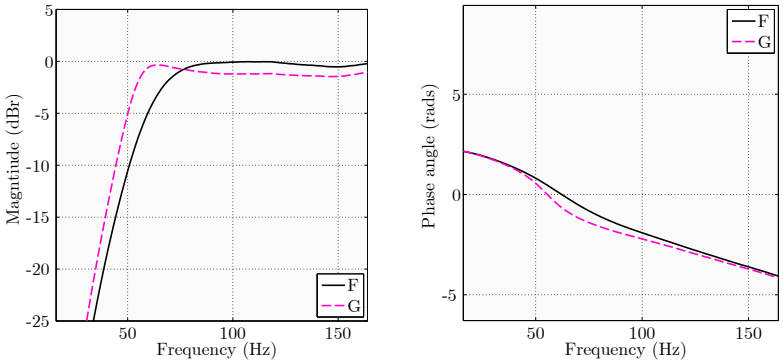
Figure 77: Group II: D vs G. Magnitude and phase



Pair	$\overline{M}$scores	Subjective result
D vs G	$\overline{M}_D > \overline{M}_G$ (0.775>0.696)	Equal split; D selected in 12 of 24 trials, 50 %.
Summary:	D outperforms G everywhere except a region in bands 3 and 4. Music spectrum contains peak energy in bands 3 and 4.	
Conclusion:	Listener votes are exactly divided between otherwise superior performance of loudspeaker D, and G which shows an isolated region of elevated m scores in the MTF matrix in bands coinciding with peak in signal content.	

For pair DG, detail in the results matrix of G shows an increase in m in bands 2 and 3 due to the elevated output level from the underdamped resonance. The increase in m within a band reduces with increasing modulation frequency, i.e. the apparent increase in performance is not consistent throughout a given band; the effect is therefore localised to a small range of elements within the matrix and the overall mean score indicates that G will be less accurate in reproducing bass content overall than loudspeaker D. The same effect, of a localised increase in m that does not extend across the whole band for all modulation frequencies, was also observed in model G discussed in section 8.2.1.

Figure 78: Group II: F vs G. Magnitude and phase



Pair	$\overline{M}$scores	Subjective result
F vs G	$\overline{M}_F > \overline{M}_G$ (0.706>0.696)	Split tends towards G, selected in 19 of 24 trials, 79 %
Summary:	Closest $\overline{M}$ scores of any pair featured in tests I and II. Only pair in any test where subjective results show a trend in opposite direction to that predicted by $\overline{M}$. Loudspeaker F is closer to reference system in bands 5 to 10; G is closer in bands 1 to 4. Music peaks in bands 3-4.	
Conclusion:	Listeners were influenced by increased performance in bands where peak signal content occurs, but algorithm reflects increased performance elsewhere.	

The examination of pair FG is presented here because it was of interest to understand why these two systems returned such close $\overline{M}$ scores; the magnitude response indicates that they should sound sufficiently different to be detected by listeners, with G having an extended low-frequency response due to an underdamped resonance at 63 Hz. Presence of a resonance can be seen in the phase response plots, where the two systems differ most at the frequency where the resonance in G occurs. As in the examination of other model pairs discussed so far, it seems that the underdamped resonance in G elevates the m scores over a small region of elements in the MTF matrix, decreasing with increasing modulation frequency; the effect is not sufficient to dominate the mean score, so $\overline{M}$ indicates that F will be more accurate at reproducing bass content overall, despite G having an apparently more extended response. This pair is also of interest because it seems to contradict the trend in responses from the group of more acute listeners, and the unscreened data for this pair produced a clear outcome where G was deemed as sounding more like the reference by the majority of participants. This suggests that many listeners in the experiment were responding to an apparent increase in loudness of the bass, compared to F, over a small range of frequencies where it will have been very obvious due to strong signal content. As discussed in section 1.1.3.3, professional mix engineers are unlikely to view a slight deficiency in overall bass level as a serious problem as long as a monitor does not

distort the balance of individual instruments. It is therefore considered possible that although the objective result for pair FG did not agree with the subjective judgement of the untrained listeners in this study, it would better reflect the perceived accuracy of bass reproduction when judged by the target population of listeners in the intended application.

8.2.3 Conclusions Following Investigation of Non-Significant Results

The following conclusions were made after closer inspection of objective results in comparison to subjective judgements from pairs in listening tests I and II where the null hypothesis was not rejected:

- The intensity images were found to be the most useful method for initial comparison of each model pair and understanding the contrasting behaviour in alignments over different regions, both frequency band and modulation frequency. Inspection of these plots allowed hypotheses to be formed that were then confirmed through inspection of the full numerical matrix values.

- It was observed that low-frequency alignments with underdamped resonances give an increase in m score over an isolated range of elements in the MTF matrix, seen both in the numerical results and the intensity images. These do not affect the whole band as the increase in m reduces for higher modulation frequencies. The boost in a loudspeaker's output level over a narrow frequency range due to an underdamped resonance is therefore not sufficient to bias the mean score; $\overline{M}$ still indicates that a loudspeaker with this type of alignment is less accurate in reproducing content overall than one that may be perceived as moderately lacking in bass but gives a more even reproduction throughout the entire low-frequency region.

- With consideration of the musical content used for assessment in this project, it was consistently found that differences in MTF performance between loudspeakers coincided with regions of peak or very strong musical content; an otherwise inferior system showed increased values of m for the lower modulation frequencies in bands where differences were likely to be more obvious due to the elevated signal level. The subjective data indicates that assessing a loudspeaker's ability to maintain overall accuracy of a musical presentation was very difficult for many participants in this study; it is suspected that many listeners were equating a perceived increase in bass level over a narrow frequency range with greater reproduction accuracy overall. Comparison of similarly contrasting alignments would benefit most from evaluation by the target population of listeners, professional mix engineers; they are expected to be more skilled in comparisons that require more than detection of large differences in overall bass level.

- The pairs of models considered in this section did not show sufficient consensus amongst listeners to be confident that they were really voting in favour of a particular model. However, where a tendency towards one or other loudspeaker existed, it was in the direction predicted by the $\overline{M}$ scores in all but one case, pair FG in Group II. It was concluded that the judgements of many participants were influenced towards the model with the more extended response arising from an underdamped resonance, occurring in a frequency range where the music used for evaluation had strong content. The algorithm

opposed this result, showing a higher overall score for the model with a less extended but better-aligned response. This result is not taken as evidence that the method needs immediate adjustment. As described in section 1.1.3.3, it is known that professional mix engineers do not always view low-frequency extension as a priority when listening critically in their work; the apparently anomalous result is therefore taken as evidence that the method requires subjective assessment by the target users before a decision can be made about whether to modify the algorithm.

8.3 Subjective Modification of Band Scores

The analysis in section 8.2 indicated that it might be useful to apply a form of subjective modification to the objective results. The first method is an adjustment to account for the increase in the hearing threshold at low frequencies; this allowed assessment of whether reproduction level during the listening tests was high enough to enable participants to hear all the available musical content. The second adjustment is a weighting according to programme content, as it had been concluded that this may have influenced listener judgements of some loudspeaker models.

8.3.1 Adjustment for Hearing Threshold

Holland *et al.* [64] applied a noise floor to the MTF matrix, based on the minimum audible field (MAF) [206, 207]:

$$\tilde{m}_q = m_q \cdot \frac{1}{1 + 10^{\frac{L_N - L_S}{10}}} \tag{8.1}$$

where: $\tilde{m}_q$ is the MAF-corrected version of modulation index m_q in the q^{th} frequency band, L_N is the MAF SPL at centre frequency of band q, and L_S is the mean SPL in that band. Note that this correction is independent of modulation frequency.

This form of correction reduces the MTF contribution of bands where a loudspeaker's output falls below the minimum audible field, or lies very close to it. A replay SPL relative to the MAF has to be assumed to calculate the correction scores. For the listening experiments in this study, reproduction levels were measured, as listed in Table 10; the mean SPL across all extracts within a given listening test were used to normalise the passband level of each virtual loudspeaker's magnitude response before calculation. A spline interpolation was then used to calculate correction factors at the centre frequency of each analysis band as the standardised threshold values are only listed at $1/3^{\text{rd}}$-octave values from 20 Hz upwards. Table 21 shows the $\overline{M}$ scores for each of the virtual loudspeakers in listening tests I and II before and after the MAF-correction. An alternative replay level of 70 dB SPL was also tested to see the effect of drastically reducing the playback SPL. The symbol $\widetilde{M}$ has been used to denote that the MTF matrix mean score has been calculated from the MAF-corrected m values.

Loudspeaker	$\overline{M}$	$\widetilde{M}$ 86.26 dB SPL	$\widetilde{M}$ 70.00 dB SPL
R	0.939	0.936	0.865
C	0.795	0.786	0.773
D	0.419	0.411	0.407
E	0.656	0.647	0.635
F	0.893	0.883	0.871
G	0.692	0.679	0.678

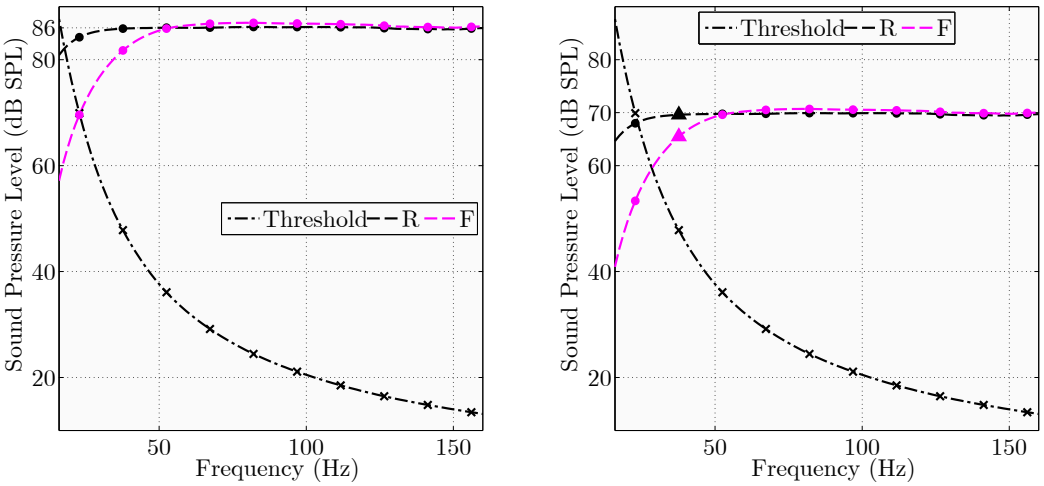
(a) Group I, evaluated in listening test I

Loudspeaker	$\overline{M}$	$\widetilde{M}$ 86.97 dB SPL	$\widetilde{M}$ 70.00 dB SPL
R	0.845	0.839	0.809
C	0.576	0.569	0.565
D	0.775	0.767	0.754
E	0.645	0.644	0.638
F	0.706	0.704	0.699
G	0.696	0.693	0.689

(b) Group II, evaluated in listening test II

Table 21: MAF-corrected $\overline{M}$ scores for Group I and II virtual loudspeakers. The listed SPLs are assumed replay level; the alternative notation $\widetilde{M}$ is used to show that these values are modified mean scores

It can be seen in Table 21 that at the playback levels used for experimentation, the MAF adjustment makes at most a reduction of 0.013 to the $\overline{M}$ scores (model R in Group I); this indicates that the replay SPL was not quite high enough to ensure that all low-frequency content in the musical signals was presented above the threshold of audibility, but the adjustments do not affect the objective rankings based on $\overline{M}$ scores; it therefore does not affect the conclusions based on comparison with subjective data presented in section 8.1 and 8.2. However, it can be seen that the $\overline{M}$ scores change by up to 0.074 from the uncorrected value if the level had been reduced to 70 dB SPL; this reduction changes the objective ranking for Group I, placing F above R. The difference between these models decreased from $\overline{M}_{\Delta,RF} = 0.046$ without correction to $\widetilde{M}_{\Delta,FR} = 0.006$; the order of the subscript here denotes that F was ranked above R. It was not initially clear whether this change in ranking was appropriate. As this correction was based on reproduction levels, the magnitude responses of these models were directly compared to the threshold of hearing, assuming the defined replay levels. Figure 79 plots the results.



(a) Assumed reproduction level of 86 dB SPL

(b) Assumed reproduction level of 70 dB SPL

Figure 79: Comparing output levels of loudspeaker models R and F (Group I) against the minimum audible field for two assumed reproduction levels. Markers show the centre frequencies of the MTF analysis bands

In Fig. 79a it can be seen why the correction slightly increased the difference in score between R and F when assuming experimental playback levels ($\overline{M}_\Delta = 0.046$, $\widetilde{M}_\Delta = 0.053$): the first point in F lies on the threshold; results in this band are penalised by the correction and the mean score reduces. This appears to be appropriate. In Fig. 79b, it is seen that from the third band upward, model F has a slightly increased sensitivity compared to R; this gave it slightly higher m scores compared to R. The second band, marked by a triangle in the plot, is the only point where R significantly outperforms F; results from the first frequency band in both models contribute nothing to the overall score because they lie below the hearing threshold if reproduction is performed at this SPL. Table 22 summarises the differences for this correction:

Model	F_{cor}	$\widetilde{M}$, Band 1 (below triangle)	F_{cor}	$\widetilde{M}$, Band 2 (triangle)	F_{cor}	$\widetilde{M}$, Bands 3–10 (above triangle)
R	0	0	0.99	0.92	1	0.97
F	0	0	0.98	0.60	1	1.01

Table 22: Comparison of corrected values assuming a reproduction level of 70 dB SPL: Group I models R and F. The MAF correction factors F_{cor} are shown next to the mean MTF scores for the corresponding bands. The band descriptions relate to the markers shown in Fig. 79b

The results in Table 22 show that after discounting (maximally penalising) the bands below the threshold, it is less easy to distinguish between the models; slightly better performance of loudspeaker F in all but one band is traded off against notably worse performance in a single

band. This is sufficient to bring its mean score closer to that of loudspeaker R. This appeared to be reasonable: returning approximately the same score for a loudspeaker that performs considerably better in a single band as one that returns slightly higher scores across all others. Given the behaviour observed across the other models, where mean scores were reduced if the loudspeaker could not audibly reproduce signal content at the specified replay level, the correction was considered to be a suitable method for implementing this form of subjective adjustment.

Based on this analysis, it was concluded that the MAF-correction is a useful modification if considering reproduction levels that are low for typical mix monitoring situations. However, it might limit general applicability of the results as they will only be valid for one assumed SPL. One option is to present results at several assumed replay levels, similar to the common practice in distortion measurements.

8.3.2 Adjustment for Programme Balance

An additional correction was attempted which adjusted for the differing relative level of signal content across bands. A set of weighting factors was calculated using the mean-band levels of the averaged music spectra shown in Figure 74. The peak mean-band level was used as the reference point to scale weighting factors for each band:

$$w_q = \frac{1}{1 + 10^{\frac{S_{max} - S_q}{10}}} \tag{8.2}$$

where: w_q is the weight for the q^{th} frequency band, S_{max} is the peak mean SPL across all bands, and S_q is the SPL in the q^{th} band.

The programme weighting was applied to the MAF-corrected algorithm results, adjusting for the weighting during calculation of the mean scores $\overline{M}_{wtd}$:

$$\overline{M}_{wtd} = \frac{\Sigma_{q=1}^{10} \tilde{m}_q w_q}{\Sigma_{q=1}^{10} w_q} \tag{8.3}$$

where $\tilde{m}_q$ is the MAF-corrected mean score in the q^{th} band. Note that there is one weight for each of the ten bands, independent of modulation frequency.

Table 23 lists the overall mean scores before and after this modification for the systems compared in listening tests I and II.

Speaker	*Rank*	$\overline{M}$	$\overline{M}_{wtd}$	*Rank*
R	1	0.939	0.963	1
C	3	0.795	0.813	3
D	6	0.419	0.389	6
E	5	0.656	0.646	5
F	2	0.893	0.951	2
G	4	0.692	0.731	4

(a) Group I

Speaker	*Rank*	$\overline{M}$	$\overline{M}_{wtd}$	*Rank*
R	1	0.845	0.898	1
C	6	0.576	0.551	6
D	2	0.775	0.821	2
E	5	0.645	0.706	5
F	3	0.706	0.748	4
G	4	0.696	0.784	3

(b) Group II

Table 23: Comparison of mean MTF scores before and after programme weighting

Inspection of the scores in Table 23 shows that the programme weighting modifies the mean scores but changes the ranking in only one case. A more sophisticated psychoacoustic model could correct for programme content more accurately; the simple level-weighting used here does not account for masking, peak level or equal loudness. However, two results from the correction are of particular interest:

- Group I, pair R vs F: In section 8.2.1 it was concluded that a lack of signal content in the frequency range covered by the lowest analysis band made differences between these two models very hard to detect. The programme weighting leads to $\overline{M}_{\Delta} = 0.012$ for these models; the mean score for R is still higher, as would be expected, but the equivalent result for F is much closer after correction; this better reflects the subjective result, where no directional outcome was concluded but there was a trend in listener votes towards R. This pair was highlighted as one case where differences between the models were believed to be genuinely inaudible for many listeners. It is suggested that programme weighting may be useful for modifying MTF results if it is suspected that severe lack of programme content has affected subjective evaluations.

- Group II, pair F vs G: In section 8.2.2 it was shown that this pair produced $\overline{M}_{\Delta} = 0.010$, and the trend in subjective data placed G above F, contradicting the objective result. The programme weighting changed the MTF rank to match the subjective result whilst still maintaining a small difference in mean scores, $\overline{M}_{\Delta} = 0.036$. It therefore seems that the correction may be useful if trying to correct the algorithm results to reflect subjective judgements that are largely based on perceived levels of bass, perhaps in relation to preference for loudspeakers intended for domestic reproduction. It is not recommended that it be routinely applied in the application of mix monitoring as it may distort the ability of the scores to summarise more subtle aspects of accurate bass reproduction, such as timbral fidelity and relative balance between rhythm section instruments; these will be modified by the presence of resonances in a loudspeaker's alignment that the programme weighting emphasises.

Results following the programme weighting indicate that adjustment of the algorithm scores can increase the correlation with subjective judgements if an extreme lack of signal content has affected evaluations, or listeners have been making judgements primarily on relative levels of bass in the reproduction. The rankings here were unchanged except for one pair, so do not affect the fundamental conclusions in this study where only ordinal data are considered. However, it is considered likely that the impact would be greater if trying to directly map numerical subjective ratings to the MTF scores. It should also be noted that this form of adjustment would be difficult to apply in a generalised application of the proposed method; an accurate weighting can only be derived using the exact extracts used for audition. This would be possible in experimental studies, but it would be difficult to extend any weightings to practical applications where a given loudspeaker will be used for mixing a variety of previously unknown musical signals. It is therefore suggested that such a form of correction should be viewed as a possible extension to the MTF algorithm, but not implemented routinely as part of the technique developed here.

8.4 Comparison With Other Methods

Preceding sections have established that the MTF-based method developed in this study appears to be a useful method for its intended application. This section compares features of the method with other techniques that might also be used to evaluate the reproduction accuracy of mix monitors at low frequencies. For this comparison it is helpful to summarise the desirable requirements that were defined when developing the current method:

- Focussed on low frequencies. This is the region containing the rhythm section instruments that are especially susceptible to being distorted due to the inherent roll-off of loudspeakers at the lower limits of their response.
- Describes a monitor's ability to accurately reproduce musical content. This means faithfully transmitting the constituent components in time and amplitude.
- Presents intuitive results. Widespread adoption of the method is more likely if results can be feasibly presented on a product datasheet and interpreted without extensive theoretical understanding of the method or electroacoustics.
- Allows easy comparison with other systems.
- Shows evidence of ability to predict subjective judgements.

8.4.1 Possible Alternatives

It was described in section 1.1.4.1 that there are many different objective measures which may be used to assess the performance of a loudspeaker. Considering the application in question, a small number of methods are discussed here as they are potential alternatives when evaluating the reproduction accuracy of monitors at low frequencies.

i) Frequency response: a) Magnitude and b) Phase.

ii) Impulse response.

iii) Waterfall plots.

iv) Wigner distribution.

The magnitude of the complex frequency response is a very commonly used measure of loudspeaker performance, and information relating to this parameter is a standard feature of product datasheets and technical specifications. It has been said that reasonable flatness of the steady state amplitude response, measured on-axis in free field, is a necessary but not sufficient requirement for design and selection of high-quality monitors [6]. As discussed in more detail in section 1.1.4.1, it is not the most revealing measure when assessing reproduction accuracy but has been shown to be a useful indicator of various aspects of subjective impression, including listener preference [61, 105, 165, 171].

As described in section 1.1.3.1, the phase response is the natural counterpart to the magnitude response, but is very rarely seen on monitor datasheets. This is presumably because it is less intuitive, and less well understood in terms of how it relates to subjective impression. With knowledge of how to interpret the information, the phase response can be very useful in

making an inference about a loudspeaker's behaviour in the time domain. Greater deviation from a constant gradient of the phase with frequency indicates that constituent components of a signal will be delayed by different amounts; the reproduced envelope will therefore not be simply a delayed copy of that which entered the system. This is why phase distortion is also referred to as envelope delay, group delay, or time-delay distortion [24, 29, 208]. However, even with an understanding of this behaviour, it is still difficult to relate it to subjective impression. Experiments have been done to evaluate the threshold of audibility of group delay [27, 28, 209] but in relation to the performance of loudspeakers at low frequencies, neither this nor the phase response is intuitive in the same way as plots of the magnitude, where it is generally understood that a lower –3 dB frequency means that lower notes will be reproduced.

Although the impulse response (IR) is a complete description of all the linear properties of a loudspeaker [210], it is not necessarily the most helpful way to represent the information. It was shown in section 4.5.1 that the IRs of several loudspeakers can be compared but important differences may appear to be subtle without closer scrutiny of the data. The duration of ringing in the impulse responses can be used as a basis for quantifiable comparison, but this alone is not a reliable measure for predicting audible differences [211]; Deer *et al.* [209] and Preis *et al.* [212], investigated the audibility of differences between filters with fixed magnitude but varying amounts of phase distortion, producing ringing in the corresponding impulse responses; it was seen that audible differences were detectable when the duration of ringing in the time domain was approximately the same but the envelope, or shape, of the IRs was different. As explained in section 3.2.2, it can also be difficult to 'isolate' the low frequency characteristics of a loudspeaker in the presence of mid- and high-frequency behaviour, as is the case when inspecting the impulse responses of real monitors. However, this can be an informative representation if the user has a thorough understanding of how to interpret the information; Preis [213] produced a comprehensive summary illustrating how changes in a system's magnitude and/ or phase affects the time-domain response.

Waterfall plots, or more formally, cumulative spectral decay (CSD) plots, are a time-frequency representation showing how the steady-state magnitude response of a loudspeaker decays after system excitation ceases [210, 214]. Ideally, a loudspeaker should 'stop sounding when the music stops' [215], but this is impossible in an electro-mechanical resonant system such as a loudspeaker, especially those which use bass reflex cabinets to extend the low frequency output (as discussed in section 1.1.3). However, the relative performance of different monitors may be compared by inspection of their decays at low frequencies; it is often seen that long resonant tails are present, showing that the system decay is not constant with frequency. The usefulness of waterfall plots for comparing mix monitors at low frequencies was clearly demonstrated in the study conducted by Newell *et al.* [34]. This is a visually appealing and intuitive format, but it has been noted by several authors that care must be taken when directly comparing the plots of different loudspeakers. The processing parameters for this representation are not standardised but greatly influence the appearance of the results; these include the time increment (spacing between 'slices') and the method used for windowing of each segment [63, 210, 215].

The final method considered here is the Wigner distribution. This has its origins in the field of quantum theory, but was suggested as a useful joint-domain method for audio analysis by Gerzon in 1974 [216]. The method is still not commonly used in evaluation of loudspeakers, but the idea of its use has been explored by several authors in the context of audio applications.

Janse and Kaizer [217] provided a comprehensive theoretical review of the method; Preis and Georgopoulos [218] took a more practical viewpoint, presenting clearer illustrations of the how the results of the Wigner distribution may be interpreted with respect to the more typical representations of audio measurements. An interesting comparison with other time-frequency measures for audio analysis was also presented by Brunet *et al.* [219]. If implemented and interpreted correctly, this representation summarises a lot of information about a system as it contains four main properties of interest: frequency response, group delay, instantaneous power, and instantaneous frequency [218]. Unfortunately, the method has a number of drawbacks which have limited its widespread adoption for the evaluation of loudspeaker performance. First, although it can be derived from either the impulse response or frequency-domain data, it is difficult to implement, especially so for discrete-time signals (as is typically the case with audio measurements). It is also prone to aliasing effects as it requires a sampling frequency of four times the highest frequency component of the signal [217]. The results are also difficult to interpret without a thorough understanding of the technique, and there is little evidence demonstrating that it is useful in predicting subjective impression [6, 69, 212]. However, if these limitations are overcome, it is potentially a very powerful method for evaluation. In relation to the Wigner distribution, Preis and Georgopoulos [218] stated that 'the problem of time-frequency analysis is to develop a two-dimensional display of one-dimensional data that can be computed efficiently, that clearly reveals those essential features to be extracted from the data, that has no distracting artifacts, and that is readily interpreted physically'. It can be said that this description is an elegant summary of the primary aims that were defined when developing the MTF-based method which is the focus of this study.

8.4.2 Demonstrating Alternatives

To illustrate the differences between the methods described in section 8.4.1, they were implemented for two of the virtual loudspeakers and results compared. The only alternative method not demonstrated here is the Wigner distribution. With reference to the drawbacks described in the previous section, there was uncertainty about whether this complicated method had been implemented correctly, partly due to the high sample rate requirements; the results were therefore considered unreliable and not included in the comparison of methods presented in this section.

Models D and G from Group II were chosen for this investigation because, as explained in section 4.1.3, they present an interesting juxtaposition in the context of the intended application; they represent a 'classic' sealed- vs ported-cabinet monitor comparison. Listeners were not in agreement about this pair, and no directional outcome was returned. However, based on objective data, it was considered unlikely that they could not be audibly discriminated; it was concluded to be more likely that loudspeakers D and G were audibly different but listeners could not agree about which sounded more like the 'ideal' reference.

The differences between the chosen loudspeakers in terms of the alternative analysis methods considered are demonstrated and considered in the following subsections. Note that magnitude and phase of this pair were presented in section 8.2.2 when discussing the evidence for the conclusion that listeners found this a difficult judgement to agree on; they are presented in the following sections as separate plots for the purposes of evaluating each method in isolation.

8.4.2.1 Frequency Response: Magnitude The magnitude responses are presented in Figure 80. As is typical in this type of plot, a logarithmic frequency axis has been used. The plots for a full-range loudspeaker can be expected to cover the nominal limits of the audio spectrum, 20–20 kHz; for this application where only the low frequency are of interest, the frequency range has been limited to the analysis range of the algorithm, 16–164 Hz.

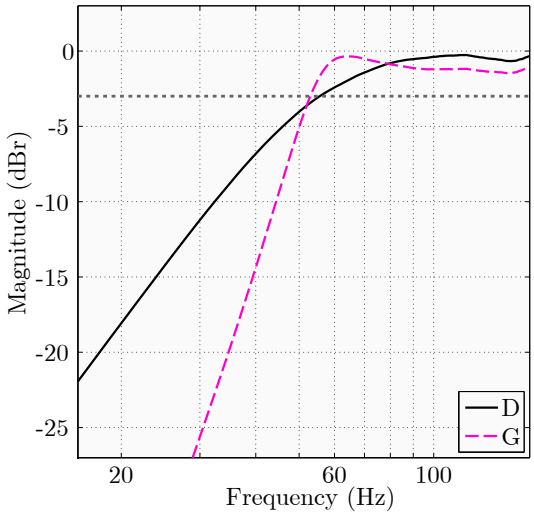


Figure 80: Loudspeaker D and G (Group II): Comparison of magnitude responses (low frequencies only). The grey dotted horizontal line marks –3 dB

This representation of the models is intuitive, and focussing of the plot on low frequencies allows easier inspection of the range of interest. The resonance of the ported model G increases its output relative to D between approximately 52 and 78 Hz, but it can be seen that both loudspeakers have approximately the same nominal cut-off frequency; close inspection shows that G is slightly more extended: $f_{cD} = 56$ Hz; $f_{cG} = 53$ Hz. Based on this criteria as a quantifiable measure of performance, G is ranked above D. However, this is only a reliable measure if interested in the relative extension of the loudspeakers.

8.4.2.2 Frequency Response: Phase The phase responses are compared in Figure 81. It is seen that the gradient of G is slightly steeper than that of D, but unlike D, it has a marked deviation from a linear characteristic below approximately 60 Hz.

Combined with the magnitude response, it is concluded that this provides a lot of information when comparing loudspeakers if the 'user' understands the implications of non-linear phase characteristic with frequency. However, based on the phase response in isolation it cannot be understood how the relative output levels of the two loudspeakers compare.

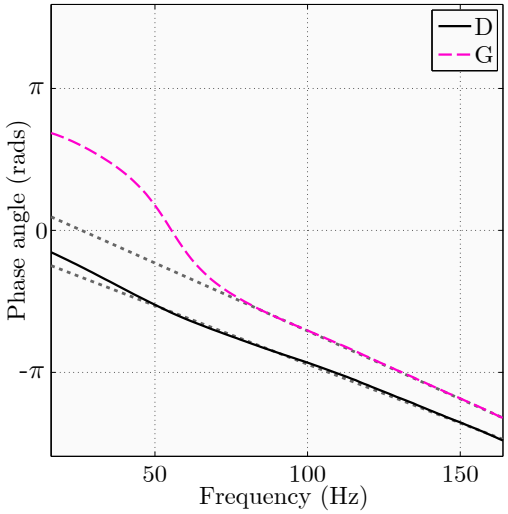
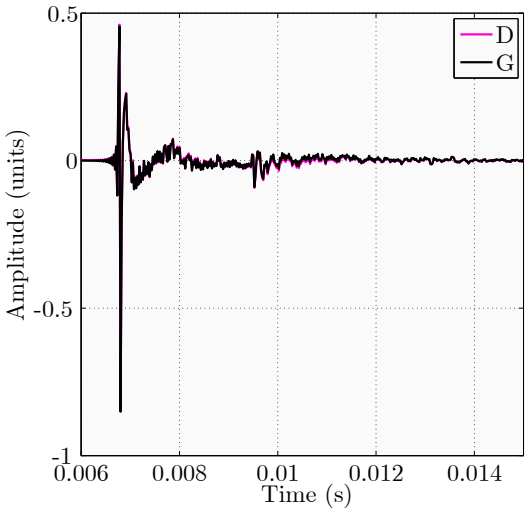


Figure 81: Loudspeaker D and G (Group II): Comparison of phase responses (low frequencies only). The unwrapped phase is shown; grey dotted lines show an equivalent linear characteristic for each loudspeaker in this frequency range

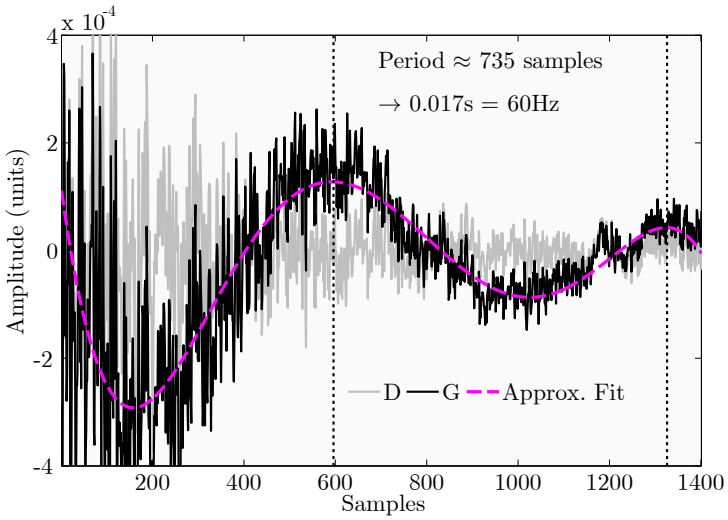
8.4.2.3 Impulse Response The impulse responses of loudspeakers D and G are compared in Figure 82. It was stated in section 3.2.2 that impulse responses were not the preferred method of representation in this study because comparison of systems was difficult in the presence of a loudspeaker's mid- and high-frequency behaviour; as demonstrated in Figure 82a, differences in the low frequencies at first appear to be negligible. This does not mean the IR representation is not informative; it contains useful information if the user knows how to interpret the features of the data. Figure 82b shows the same information but with restricted x and y axes; differences between the two loudspeakers then become more obvious.

It can be seen that G has ringing in the impulse response due to the port resonance that increases its low frequency extension. With some extra calculation, approximate curve fitting, and calculation of the period of the fluctuations, the corresponding frequency was found to be 60 Hz; comparison with Fig. 80 shows that this is the approximate frequency of the system resonance. Inspection of the results for loudspeaker D in Fig. 82b shows that it does not exhibit this type of ringing, indicating that it has a superior transient response compared to model G.

From this analysis it is seen that the IR can be useful in comparing low frequency behaviour of different monitors, but requires some extra work, understanding of how to interpret the behaviour and calculate the required information, and the ability to zoom into the data to examine the features of interest.



(a) The time and amplitude axes have been restricted to allow some difference between the two loudspeakers to be seen; full duration of the impulse response given $f_s = 44.1\,\text{kHz}$ and $N = 2^{15}$ points was 0.743 s.



(b) Closer inspection of the IRs allows greater differences to be seen. Vertical dotted lines mark the approximate period of the ringing in loudspeaker G

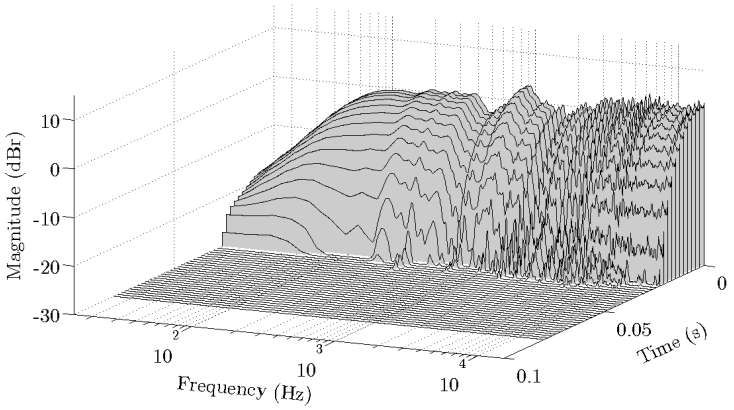
Figure 82: Loudspeaker D and G (Group II): Comparison of impulse responses

8.4.2.4 Waterfall Plots Waterfall plots for D and G are presented in Figure 83. The data was generated by taking windowed segments of the impulse response at increments of 2 ms, each 100 ms long. The FFT length for each segment was $2^{12} = 4096$ points long, giving a frequency resolution of approximately 11 Hz, but interpolation and five-point moving-average smoothing have been used to improve the visual appearance of the final plots.

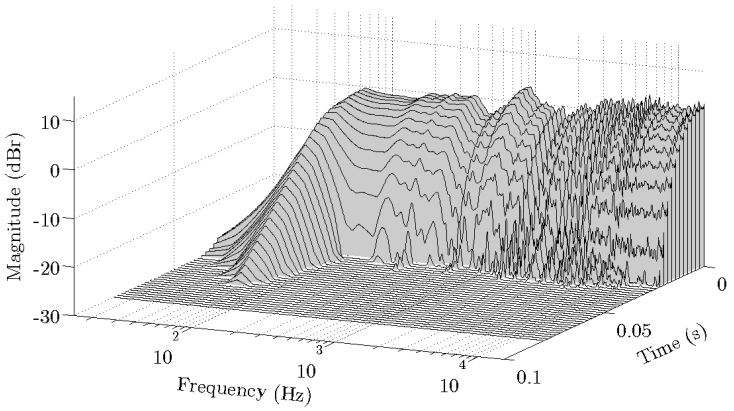
It is less easy to directly compare the extension of the loudspeakers in this representation, especially at the lowest frequencies, but this time-frequency visualisation provides useful information compared to inspecting only the magnitude or phase plots. The virtual loudspeakers used in this comparison have identical mid and high frequency responses so the results in Fig. 83 appear to be similar overall; however, the 'resonant tail', or ringing, of G around 60 Hz is clearly

visible, especially when compared to D which decays more quickly and is approximately constant for all frequencies.

It is concluded that this is an informative and intuitive method for comparing the performance of two loudspeakers in time and frequency, provided that the plots have been produced with exactly the same processing parameters.



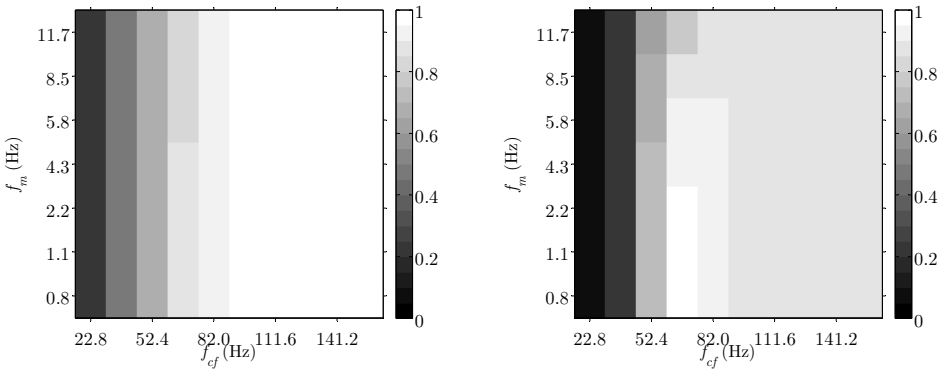
(a) D



(b) G

Figure 83: Loudspeaker D and G (Group II): Comparison of waterfall plots. Loudspeaker G (ported cabinet) shows a resonant tail around 60 Hz, causing musical content in this frequency range to 'hang on' relative to the rest of the spectrum; in contrast, the waterfall plot for D (sealed cabinet) shows a decay with approximately equal duration at all frequencies.

8.4.2.5 MTF-Based Method To conclude this demonstration of methods, the results from the MTF-based algorithm for virtual loudspeakers D and G are shown again in Figure 84:

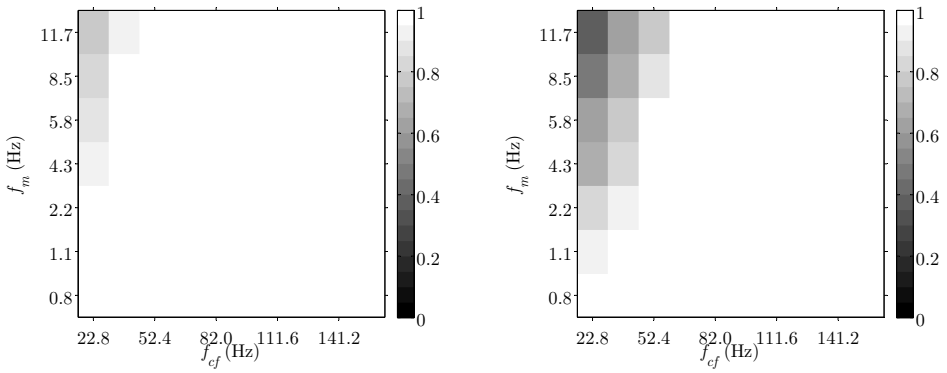


(a) D; $\overline{M} = 0.775$ **(b)** G; $\overline{M} = 0.696$

Figure 84: Loudspeaker D and G (Group II): Comparison of intensity images. The mean scores, $\overline{M}_{\Delta}$, are given in the individual figure captions

It is seen that D has an overall higher sensitivity (a lighter image) except over a restricted frequency range where G has an underdamped resonance centred on band 4 (60–75 Hz), identified by the 'bright spot' in an otherwise dark region. This is confirmed by looking at the magnitude plot. Loudspeaker D has virtually no variation in the vertical direction, indicating that it reproduces temporal fluctuations in a musical signal's envelope accurately. The increased 'ripple' behaviour of G reflects the fact that its performance in the time domain is compromised compared to D. This was inferred from inspection of the phase responses, and confirmed by inspecting the impulse responses and waterfall plots. Therefore, it is seen that these representations reveal useful information about which is the most accurate reproducer of bass content, assuming that the user understands how to interpret their features, and that manufacturers of different monitors will use the same parameters for analysis and presentation of the data. However, none of the alternative methods allows a direct quantifiable comparison between D and G that summarises their overall reproduction accuracy throughout the whole bass region. It is seen that D returns a higher mean MTF score than G: the numerical output of the algorithm signifies that D is an overall more accurate reproducer of bass content; the visual output (intensity image) provides more information about how the loudspeakers differ and allows them to be easily compared in different parts of the bass region.

Also presented here for interest are the results of these two virtual loudspeakers using an alternative MTF-based method. This was the method described but eventually rejected in chapter 2, using the Schroeder equation for computation with perfect-system normalisation. It is included in the comparison here because the Schroeder method was previously implemented in the application of the MTF for analysis of mix monitors at low frequencies [64], and has the significant advantage of extremely fast computation time. The results are shown in Figure 85. Identical parameters bands and modulation frequencies have been used in both methods. The results for the rest of the Group II models are presented in Appendix L.



(a) D; $\overline{M}_{\Delta} = 0.984$ **(b)** G: $\overline{M}_{\Delta} = 0.941$

Figure 85: Loudspeakers D and G (Group II): Comparison of intensity images derived from the normalised Schroeder method. The mean scores, $\overline{M}_{\Delta}$, are given in the individual figure captions

Notable features illustrated in Fig. 85 are:

- Smaller mean score variation - Across all loudspeaker models in Group II the mean score variation with the normalised Schroeder method was $\overline{M}_{\Delta} = 0.080$ compared to $\overline{M}_{\Delta} = 0.269$ with the method finally developed. For pair DG specifically, $\overline{M}_{\Delta,DG} = 0.043$ compared to $\overline{M}_{\Delta,DG} = 0.079$ for the proposed method.
- Lack of characteristic behaviour - There appears to be little value in showing the intensity images with the normalised Schroeder method as, unlike the chosen method, they don't display as much useful information about the loudspeakers' alignments.
- High scores when output is negligible - The mean scores are misleadingly high. It is known (e.g. as shown in Fig. 80) that these loudspeakers are not particularly extended, but the method returns scores suggesting that reproduction is close to perfect.

None of these observations are true for the method which was finally developed and used in this study. The last of these points was one of the main reasons for rejecting the normalised Schroeder method (as detailed in section 2.4); the method shows close to perfect scores in bands where the loudspeaker outputs negligible musical content. The other conclusions after analysing the experimental loudspeakers are taken as further confirmation that rejection of this approach was justified. It can also be observed that the effect in the intensity images is similar to that observed for the Group III models where the magnitude was fixed and only the phase altered. Therefore, if the magnitude response was of no interest, perhaps due to investigation of all-pass filters, this method could be considered as it is a much faster way to compute results. However, it can be seen that the scores will be misleadingly high when evaluating real monitors. Although the results presented for the normalised Schroeder method still rank D above G, it must also be noted that the mean-score ranking returned through this method is not the same as for the final algorithm (the mean scores are shown with the data in Appendix L); as such, the objective scores do not predict the subjective results as successfully as those shown earlier in section 8.1.

8.4.3 Summary of Alternative Methods

Table 24 summarises the conclusions regarding each method after the review in section 8.4.1 and examples in section 8.4.2. These are presented in the context of evaluating loudspeaker reproduction accuracy at low frequencies. The method developed in this study is referred to as LF MTF for brevity.

Method	Conclusions
Magnitude plot	Very well established, intuitive, large body of existing subjective validation; *but* provides limited information about reproduction accuracy, doesn't allow quantitative comparison of overall performance.
Phase plot	Useful for making inference about system behaviour in time domain; *but* hard to relate to subjective impression, not very intuitive, doesn't allow quantitative comparison of overall performance.
Impulse response	Behaviour viewed directly in time domain, intuitive way to compare duration of ringing; *but* difficult to examine only the low-frequency behaviour, relevant information can be hard to see in a 'full-scale' plot, doesn't allow quantitative comparison of overall performance.
Waterfall plot	Revealing of behaviour in time and frequency, intuitive, well established, visually appealing; *but* difficult to directly compare different monitors due to impact of processing parameters on appearance, doesn't allow quantitative comparison of overall performance.
Wigner Distribution	Comprehensive summary of important time and frequency behaviour; *but* hard to implement, hard to interpret, relation to subjective impression not well validated, doesn't allow quantitative comparison of overall performance.
LF MTF	Allows quantitative comparison of overall performance, low-frequency specific, summarises extension and flatness of magnitude response and how faithfully constituent components of a musical signal's envelope are reproduced, relatively simple to implement, easy to directly compare multiple monitors due to consistency of processing and presentation parameters, evidence that it can predict subjective judgements; *but* needs further subjective experimentation before it can be reliably used for predicting listener impression.

Table 24: Summary of method comparisons

It is seen that each method has its own merits and disadvantages. Based on the criteria defined for the intended application it is concluded that the MTF-based method developed in this study is the most appropriate of the alternatives considered. This is based on the assumption that, as intended, the method for calculating and presenting results is standardised. It was explained that waterfall plots are highly sensitive to the processing parameters used to generate results, making comparisons difficult unless identical processing parameters have been used; a similar problem could affect the MTF intensity images. As described in section 2.9.1, presentation of the intensity images was developed according to a specific regime; this must be strictly followed if the images are to be reliable indicators of performance. Visual results could be manipulated by adjusting the colour scaling parameters to make the intensity images appear

whiter overall or have less variation in shading across modulation frequencies; both of these would imply that a monitor has a greater reproduction accuracy than the numerical results actually suggest. This type of manipulation must be prevented as it would make direct comparison across different monitors unreliable, and therefore compromise the usefulness of the method.

8.5 Summary

Objective analysis of the experimental models showed that the algorithm was effective in revealing and summarising different aspects of a loudspeaker's low-frequency alignment; it was therefore shown to be a useful method for comparing pairs or groups of systems when deciding which would most accurately reproduce bass content in musical signals. To validate the method as a truly useful technique, some correlation with subjective impression had to be demonstrated. This chapter focussed on comparing objective results for the experimental loudspeakers with data obtained from listener responses of the same systems evaluated in pairs.

Based on post-screened subjective data, the direction of all significant results was predicted by the matrix mean scores. Partial rankings were indirectly developed from the subjective paired comparison data; it was seen that where defined ranks existed, they were identical to the ranks based on $\overline{M}$. The subjective data did not allow direct assessment of whether an audible difference between loudspeakers in a given pair existed; a perceptible difference was assumed in pairs where the null hypothesis was rejected. From these pair results, the smallest difference in mean MTF score which corresponded with a conclusion of perceptible difference was found to be $\overline{M}_{\Delta} = 0.003$. It is not known whether smaller differences between loudspeaker low-frequency alignments are detectable, but this was the smallest change between any systems compared in this study that coincided with a clear enough consensus amongst listeners to be confident that an audible effect existed.

The null hypothesis was not rejected in nine pairs across listening tests I and II based on post-screened data. After closer inspection of these pair results it was inferred that in the majority of these pairs, participants could detect a difference but could not reach consensus regarding the experimental question: which model sounded more like the reference. Inspection of the corresponding intensity images and numerical data described in detail how the loudspeaker alignments differed in behaviour throughout the bass region. Initial comparison of this information with the subjective split of results for a given pair suggested why listeners might be divided, depending on which aspect of bass reproduction they were focussing on. Viewing this data with consideration of the averaged music power spectrum for each set of experiments supported this hypothesis; there was a consistent pattern where the otherwise inferior system in a pair showed higher m scores in analysis bands that coincided with peak or very strong musical content. This was most obvious when considering pairs featuring a model with a well-controlled alignment, but moderately lacking in output throughout the low-frequency region, against one that had an underdamped system resonance, giving the impression of a more extended response but inconsistent reproduction overall. Even the more accurate listeners in this study were divided when comparing these types of alignment. Closer inspection of the visual and numerical matrix results showed that the apparent increase in performance due to underdamped resonances was only seen at lower modulation frequencies in bands where the resonance occurred; therefore, the mean MTF scores did not indicate that a loudspeaker with this type of alignment was more

accurate at reproducing bass overall compared to one that had a less extended but well-damped response. As a result, the algorithm contradicted the trend in subjective data for one pair. This was not considered a detrimental outcome as it was suspected that some participants in this study, who were not professional mix engineers, based their judgements on the perceived level of bass reproduction. It was concluded that auditory evaluation of loudspeaker alignments presenting this type of contrasting behaviour was very difficult for many participants in these experiments. It is believed that the target population of listeners for this MTF method would produce more conclusive outcomes for the comparison of alignments that participants in this study found difficult; it is also believed that the mean algorithm scores would better reflect the assessments of these experienced professionals, who would consider more than relative level at low frequencies when making judgements about accurate bass reproduction.

Two forms of adjustment were applied to the MTF scores to see whether they affected the comparisons with subjective results. The first of these was a correction to compensate for the increase in hearing threshold at low frequencies. This gave a maximum reduction in $\overline{M}$ of 0.013 at the reproduction levels used for experimental playback, and did not affect the conclusions regarding comparison of objective with subjective outcomes. However, simulating a reproduction level of 70 dB SPL showed a reduction in $\overline{M}$ of up to 0.074. It is therefore recommended that the correction be applied if replay SPL is low compared to usual mix monitoring levels, below approximately 75 dB SPL. For general use, the corrected results might be presented for a number of assumed playback levels, similar to the procedure used for distortion measurements. The second method of adjustment was a weighting according to the relative programme level in bands corresponding to the MTF algorithm. It was concluded that a much more sophisticated psychoacoustic model would be required to accurately adjust the scores in this way, but results suggested that this form of adjustment might be useful in some circumstances when comparing results with subjective data: if a severe lack of programme content is thought to have affected subjective judgements, or if it is suspected that listener evaluations have been primarily based on perceived levels of bass reproduction.

The findings from both objective and subjective assessment of the MTF-based method developed in this project indicate that it is well suited to the application of assessing low-frequency reproduction accuracy in professional mix monitors. The method's main features were compared against alternative objective measures that might be used in this application. It was seen that whilst it is not the most comprehensive descriptor of loudspeaker behaviour, it is the only one that met all of the criteria defined at the start of the project. It also provided further validation for the fundamental change in approach compared to the preceding work which was based on the Schroeder method of implementation; even with the modified frequency bands and modulation frequencies, the method developed in this study was seen to be more informative and did not return misleadingly high scores for monitors known to be lacking in output at low frequencies. The visual and numerical results reveal characteristic behaviour in a loudspeaker's alignment, and the mean $\overline{M}$ scores allow simple comparison and direct ranking of multiple systems. This was shown to be especially useful in comparing loudspeakers with contrasting types of alignment. The mean scores predicted all subjective outcomes in this study when listeners were in sufficient consensus about their judgements. Participants were in good agreement with each other when one of the models was severely deficient in output level throughout the entire low-frequency region; they were more divided in their responses when comparing models that

presented less simplistic differences between alignments. Therefore, obtaining judgements about these more challenging comparisons from the intended population of listeners, professional mix engineers, would allow further validation of the predictive capability of the MTF technique.

9 Conclusions and Suggestions for Further Work

This project has developed and evaluated a method for the objective assessment of loudspeaker reproduction accuracy at low frequencies; the method is based on the Modulation Transfer Function (MTF) and is specifically intended for the type of loudspeakers, or monitors, used by professional sound engineers when mixing recorded music. If the monitors do not accurately reproduce a musical waveform, the relative levels of rhythm section instruments may be incorrectly balanced. A mix affected in this way will not sound well balanced when replayed on other systems. The use of inaccurate mix monitors can therefore lead to very expensive remixing of recordings, or delivery of a degraded musical product to customers, potentially compromising crucial elements of the performance such as impact, involvement, and therefore enjoyment.

Previous studies had suggested that the MTF is a useful measure in this application, but the technique needed further refinement and validation. The primary aim of this project was therefore to investigate whether an MTF-based technique may be a useful descriptive and predictive measure for professional engineers when selecting the most appropriate monitors for their work. A summary of key findings from this study and conclusions arising from the work are presented in section 9.1. Suggestions on how the method may be developed based on these conclusions are described in section 9.2.

9.1 Conclusions

This project has reviewed the method used in preceding studies to evaluate the reproduction accuracy of mix monitors at low frequencies. A detailed investigation of the procedure and principles it is based on led to a different approach being adopted. Three key aspects of the technique were modified: MTF computation method, arrangement of frequency bands, and selection of modulation frequencies. It was concluded that the combination of these refinements produced an MTF-based algorithm that was more revealing and better tailored to the intended application than the method that previously existed. A new format for presenting results has also been established: numerical data in the MTF results matrix is shown as a greyscale 'intensity image'. This information is not presented in existing MTF-based applications because inspection of the data in numerical form is cumbersome and makes trends hard to identify, especially when comparing across systems, and it is typically assumed that only the overall (average) result is important. It has been shown in this study that characteristic patterns are produced within the MTF matrix that are very useful in understanding the low frequency behaviour of a loudspeaker; the visualisation of numerical data as a shaded image means that results can be easily interpreted, and several monitors can be quickly compared in a way that is simple to comprehend.

Earlier studies had not compared objective findings with subjective data; it was therefore not known whether an MTF-based method for evaluating loudspeakers at low frequencies could also be used to predict listener judgements. The problem was addressed in this study by conducting three sets of listening tests where participants evaluated loudspeaker models with different low-frequency alignments, auditioned using a range of musical extracts. Through comparison with objective results, it was concluded that the new algorithm can correctly predict listener judgements, and unlike commonly-used existing methods, allows a group of loudspeakers to be directly ranked according to their bass reproduction accuracy. The subjective experiments also

led to two further conclusions that are relevant to this research topic. Firstly, phase distortion, of the type introduced at low frequencies by certain types of loudspeaker, is audible to the majority of non-expert listeners when listening to music. There is very little evidence from formal experiments supporting this conclusion, and it is commonly assumed that the effects studied here are extremely subtle or even imperceptible, especially when using loudspeakers and music for reproduction (instead of headphones and non-musical stimuli). The findings therefore contribute towards the justification of designing and using mix monitors that reproduce low frequencies accurately. The effect was also seen to be strongly programme-dependant, i.e. low-frequency phase distortion in loudspeakers is audible with some types of music but not with others. This finding supports a small amount of existing evidence, and provides further justification for research to establish the signal properties that lead to this effect; it will then be possible to understand which types of music are most susceptible, and must therefore be mixed on the most accurate monitors. The final conclusion from the listening tests is that inexperienced listeners are not suitable participants for further subjective validation of the MTF algorithm. Although it is well established that training participants can improve their consistency and accuracy in a listening task, the results gathered in this study support the existing view that critical listening in this context requires more than the ability to detect audible differences between two stimuli; the work performed by professional mix engineers relies on being able to make comparative judgements about subtle effects such as timbre and instrumental balance, a skill that can take years of practice and experience to acquire.

Sections 9.1.1 and 9.1.2 summarise the key findings that support these contributions and discuss them in relation to existing work.

9.1.1 Algorithm Development and Objective Validation

Chapters 2 to 4 focussed on developing the algorithm and objective evaluation of experimental loudspeaker models. A distinction was made here to differentiate between the formal definition of the MTF, and an MTF-based method, which was the focus of this study. The fundamental principle of the MTF in acoustic evaluation is that it reflects the extent to which a signal's envelope in the time domain is distorted as it passes through the system under test; it is a measure of how well the depth of amplitude modulations are transferred. It was seen that 'the MTF' typically refers to only the modulus of the Complex MTF, and is traditionally calculated in two ways: direct measurement of acoustically-transmitted amplitude-modulated test signals, or computation from a measurement of the system's impulse response. The formal definition of the MTF describes a full-bandwidth continuous function of modulation frequency. An MTF-based method may be described as an application-specific version of the true MTF, band-limiting the test system and computed for a limited number of modulation frequencies shown to be relevant and revealing in the application of interest; results may be averaged or weighted and presented in a simplified format to make them more intuitive and easier to compare with those from other systems. The most prominent example of an MTF-based used in acoustic evaluation is the Speech Transmission Index (STI), a standardised indicator of speech intelligibility inside listening spaces.

Selection of a method for MTF computation was the first stage of development. Four methods were considered, forming two fundamental types, termed here BLIR (band-limited impulse response) – filtering the test *system*, and BLIP (band-limited input) – filtering the test *signal*. A review of MTF-related literature showed that both approaches are valid [68, 72, 74, 75].

It was demonstrated that selection of a BLIR or BLIP method is a fundamental issue in low frequency applications; band-limiting the impulse response means that the filter becomes part of the system being evaluated. This has been noted by other authors [78, 94, 95]. What was not found in the literature was an explanation as to why these errors are not routinely corrected for in the STI method. It was concluded that this is because the analysis is performed in a higher frequency range, relevant to speech; therefore the use of octave bands at mid to high frequencies produces band-limiting errors that range from negligible to observable (in the order of 10 % of the total score), and are reduced in the subsequent averaging procedures. It was shown that a similar approach at low frequencies, where octave-wide bands are much narrower in absolute terms, meant these errors were always large enough a critical problem, exceeding 50 % of the total score when the lowest bands are modulated by the highest frequencies. A form of correction was attempted, using normalisation by perfect system scores; this type of correction was alluded to by Linkwitz [94] and implemented by Fazenda *et al.* [95], who explained that the impact of system band-limiting may dominate results if not corrected. The main BLIR method considered here was the Schroeder equation, computing the MTF from the system's impulse response. This was the method used in preceding studies for evaluation of monitors at low frequencies [64, 87], and has also been used when applying the MTF to listening spaces at low frequencies [93, 95, 120, 220]. This computation method is simple to implement and extremely fast, but was eventually rejected for use in this study; the normalisation correction appeared to remove the errors due to system band-limiting but another inherent limitation remained: the method was seen to be insensitive to a loudspeaker's output level (changes in magnitude response). Apart from failing to reflect an important aspect of monitor reproduction accuracy, this leads to misleadingly high scores; a loudspeaker can still return near-perfect results if it has minimal impact on a signal's temporal structure, but is so lacking in low-frequency extension that it cannot reproduce audible signal content across the required frequency range. This insensitivity to magnitude response is acknowledged as a limitation of the STI method when implemented with the Schroeder equation [65, sec. 4.5.8], leading some authors to question the validity of results in relation to subjective impression [82, 84, 85].

The chosen computation method was of the BLIP type; it used band-limited white noise, amplitude modulated before being convolved with the full-bandwidth test system. This can be regarded as a simulated version of the direct-measurement method originally proposed when STI was developed in the 1970s [72]. This method is free from the inherent band-limiting errors that affect BLIR methods, so requires no correction. It also reflects changes in the system's magnitude response. The intention is that the system is excited over the specified bandwidth, β. Through a review of modulation theory, it was shown that the chosen method did have an inherent limitation: amplitude modulation of a band-limited signal 'spreads' the test bandwidth by $2f_m$ (lower and upper band limits, f_L and f_U, become $f_L - f_m$ and $f_U + f_m$ respectively). This was termed here the 'effective bandwidth', β_e. It was concluded to be an acceptable limitation; it introduced some redundancy to results where adjacent analysis bands partially overlap. The consequences were therefore preferable to the serious limitations already described for Schroeder implementation. A second issue with the chosen method was identified: the use of a pseudo-random test signal led to inconsistent results for repeated evaluation of a given test system. This was found to be another known limitation of the STI method, this time relating to the direct-measurement method [65, sec. 5.4]; as with the band-limiting errors inherent in

Schroeder computation, this problem is not corrected for in STI. Being unable to produce consistent results for a given loudspeaker was considered to be unacceptable in a method evaluating professional monitors which will be used by the most critical of listeners. An investigation was performed to determine how much averaging was needed to reduce variation in the mean score to no more than 1 % of the maximum theoretical score. The standard deviation of mean score, $\sigma_{\overline{M}}$, from 30 repeated evaluations of the same monitor was compared for different numbers of iterations between 10 and 1500; it was concluded that 100 gave a good compromise between computation time and acceptable stability of results, although as few as 10 may be used for a quick approximation ($\sigma_{\overline{M}} < 0.002$ and $\sigma_{\overline{M}} < 0.004$ respectively). It should be noted that the averaging procedure might be an area for improvement; the optimum number of iterations should be reviewed after analysing variations for a large sample of real monitors with different alignments. Also, the test signal envelopes were averaged across multiple periods of the modulation frequency as well as 'down' the ensemble of different iterations. For computational simplicity, the duration was always fixed to twice the period of the lowest modulation frequency; results at higher modulation frequencies therefore received more period averages, leading to smoother envelopes from which results were calculated. This increased variation for lower modulation frequencies appears to have been described by Rife [78] in relation to STI measurements, where it was said that the subsequent averaging and weighting was sufficient to minimise any errors. Varying the length of the test signal to ensure an equal number of period averages per modulation frequency is suggested as a possible solution. It is expected that this would make envelopes at the lowest values converge more quickly so that fewer iterations would be required; computation time would therefore decrease.

The second stage of algorithm development was selection of application-specific processing parameters. The test bands and modulation frequencies were different from those used in preceding work by Holland *et al.* [64]. That work, like the STI, defined bands logarithmically in both width and spacing. Six alternative types of arrangement were considered in this study. Selection was based on comparing results from different test systems; the chosen arrangement used ten linear-width contiguous bands, covering the range 16 to 164 Hz. It is known that the chosen arrangement produces some overlap between bands. Band overlap was found to be a source of redundancy in STI, reducing the correlation of mean scores with subjective results; corrective weightings were developed empirically from a large body of word score data [81]. There is currently no equivalent data from which such weights could be developed in the present application; there is also no evidence from this study that redundancy is a limitation that needs to be corrected. More extensive subjective validation is required before this issue can be investigated further.

Suitable modulation frequencies were chosen through inspection of the envelope spectrum derived from 168 musical extracts across a range of genres; the common styles of pop and rock were the largest categories, forming approximately one third of the total sample. This selection through analysis of application-specific material has similarities with that used in development of the STI, where averaged speech envelope spectra were used to select appropriate values [72]. The distribution of modulation frequencies in the 'generic' musical envelope spectrum approximated a continuous distribution below 20 Hz, increasing with decreasing frequency except for a small number of more prominent peaks. This was similar to the results presented by Polack *et al.* [74], although they used a statistical approach to model the fluctuations present in music rather than

the empirical one demonstrated here. Seven components were selected from the final musical envelope spectrum for use in the algorithm, ranging from 0.8 to 11.7 Hz. The range covered by the chosen frequencies is similar to that in STI (14 f_m between 0.6 and 12.5 Hz). The number of chosen values is equal to the preceding studies evaluating mix monitors (7 f_m between 3.2 and 11.5 Hz); these had been based on the STI values but refined through inspection and comparison of results, as was performed here when selecting a lower modulation frequency limit. This was chosen through comparison of results for 25 real mix monitors. The variation produced by four values of f_m below 1 Hz was analysed; 0.8 Hz was found to account for a greater proportion of standard deviation between the mean scores of all monitors than 0.1, 0.2, and 0.4 Hz combined. It was therefore concluded that 0.8 Hz was the most revealing of the tested frequencies; there appeared to be no benefit in increasing redundancy and computation time by including the lower values which contributed little in discriminating between monitors. This finding is interesting because, as pointed out by Schroeder when formally defining the CMTF [68], $\lim_{\omega \to 0} m(\omega) = 1$, i.e. sufficiently slow modulations will be transferred without attenuation though any linear passive system. A value for this was not specified, but based on these findings it seems that 'sufficiently slow' in the context of the intended application lies somewhere between $\frac{1}{0.8}$ and $\frac{1}{0.4}$ seconds (1.25 and 2.5 s).

The range of chosen modulation frequencies was not limited to extremely low values even though the generic envelope spectrum suggested that these occurred most frequently in many types of music; a small-scale analysis of specific genres returned more distinct modulation profiles, i.e. much clearer peaks that ranged up to approximately 13 Hz. Selected values were therefore chosen to cover this range whilst coinciding with small peaks that were also observed in the generic spectrum. It must be noted that this analysis was a preliminary investigation using only 30 musical extracts that were broadly categorised into the genres of Rock, Rap, and Dance. It was simply assumed here that music of the same genre would be more likely to share spectral and temporal characteristics that would give them similar envelope spectra; the fact that the three genres produced different averaged spectra with more pronounced peaks than the generic spectrum supports this hypothesis. It is expected that more formal categorisation by specific temporal and spectral features, such as performed by Wilson and Fazenda [221] when relating different signal properties to perception of musical recording quality, would return even more distinct envelope profiles. The cited study is considered to be especially relevant as one of the features considered was dynamic range compression; this is considered to be an interesting area for further work as its influence on the generic musical envelope spectrum was not formally investigated in this study. As illustrated in section 5.3.3, music with excessive dynamic range compression appears to have a less distinct temporal envelope compared to material that is not processed in this way.

It was concluded that the chosen modulation frequencies satisfy the defined criteria: they are common to a range of musical styles, they reveal differences between systems, and they cover an appropriate range of values without unnecessarily increasing computation time. The number of values was concluded to be sufficient to produce revealing results, based on evaluation of both real measured monitors and loudspeaker simulations with a range of different alignments. The findings from this study do not suggest that increasing the number or range of modulation frequencies is necessary, but this should be verified through further objective and subjective evaluation with a much greater range of real monitors.

The final stage in algorithm development considered the presentation of results. The formats used in preceding studies were adopted: average m scores in each band, and the overall mean score from the results matrix; these were given the notation of $\overline{m}$ and $\overline{M}$ respectively. An additional format was also developed, referred to as intensity images. These compact grayscale images were developed to allow simple inspection of detail within the results matrix without needing to present numerical values. They were found to be extremely useful and informative, especially when comparing multiple loudspeakers. There were two main sources of evidence that justified this conclusion from a purely objective point of view. First, it was observed that two loudspeakers may have almost identical mean scores whilst their intensity images are distinctly different. This might be an indication that the mean score needs some adjustment, such as the band weighting that is performed in STI when calculating the final score between 0 and 1 [70, 75, 80]; it might also indicate that the mean score is useful as an overall indicator of accuracy, but the supplementary information provided by the intensity images is beneficial. Secondly, objective analysis of the experimental models (presented in chapter 4), showed that the intensity images did not just demonstrate differences between loudspeakers; they were seen to exhibit specific shading patterns with characteristics that are determined by the features of a loudspeaker's low-frequency alignment.

Objective validation of the algorithm was conducted on both simulated and real (measured) loudspeakers; three were real mix monitors, sixteen were the experimental loudspeaker models evaluated in listening tests. The main finding from objective analysis of the experimental models relates to the intensity images. Visual trends produced by the MTF matrix detail were compared against the fundamental design features of each model; it was concluded that the intensity images display characteristic behaviour for different types of loudspeaker alignment. The behaviour can be broadly summarised in three main points:

1) Shading consistency in the horizontal direction (frequency band axis) indicates consistent reproduction level;
2) Shading consistency in the vertical direction (modulation frequency axis) indicates faithful envelope transmission;
3) Lighter shading reflects higher scores.

More specifically, underdamped resonances are identified by mottled shading and 'bright spots'; these typically occurred in loudspeakers with a ported cabinet. Well-aligned sealed-cabinet loudspeakers showed consistent vertical shading; this indicates that fluctuations in a musical signal's envelope will be accurately reproduced. The bass extension and the rate of attenuation at low frequencies can be judged by how far white pixels extend in the horizontal direction, and how quickly they transition from light to dark. Therefore, these small grayscale images were seen to provide more useful information about a loudspeaker's reproduction accuracy than the mean-band plots used in previous work [64], and considerably more than the single numerical scores; however, these had the significant advantage of being very simple to understand and allowed multiple systems to be directly rated and ranked. The trend in mean scores appeared to be appropriate and consistent with the general behaviour described by Holland *et al.* [64]; most notably, the reference models produced the highest values in their respective groups, and in Group III where the magnitude responses were fixed, models with greater low-frequency phase distortion produced lower overall scores.

The method was compared against other objective measures that might be used to evaluate the reproduction accuracy of mix monitors at low frequencies. Alternatives considered were the frequency response, both magnitude and phase, the impulse response, waterfall plots, and the Wigner distribution. The advantages and limitations of each method were considered and demonstrated for two of the experimental models; these presented an interesting comparison in the context of the intended application, a compromise between greater temporal fidelity and an increase in output at lower frequencies. This is a fundamental compromise that loudspeaker designers must consider when deciding what alignment a medium sized monitor should have [21]. The MTF-based method developed in this study was concluded to be the most suitable for the intended application of the alternatives considered. The visual results showed characteristic behaviour depending on the type of alignment; they reflected the extension and smoothness of the magnitude response, as well as how much the presence of a resonance used to increase extension will affect temporally varying signals passing through the loudspeaker. The numerical results summarised both of these aspects and allowed simple direct comparison between the systems, even though they had fundamentally different design strategies.

9.1.2 Collection of Subjective Data and Comparison with Algorithm Results

Chapters 5 to 8 were concerned with collecting subjective data and comparing it with results from the MTF algorithm. Conclusions regarding subjective evaluation of the loudspeaker models (before comparison with objective results) are summarised in two main points:

1) The assumptions (described in section 4.1) that influenced development of experimental loudspeaker models for listening tests I and II were valid. In listening test I, the majority of listeners could differentiate between models with exaggerated differences in overall bass output level. In listening test II where the differences between models became more representative of real monitor responses, a question of comparing alignment shape and more subtle differences in bass extension, listeners found it harder to agree on their judgements. This implies that naïve participants, even if demonstrated to be accurate and consistent in detecting differences at low frequencies, do not possess the skills required for evaluating the types of alignment that are typical when comparing real mix monitors.

2) Low-frequency phase distortion is audible in the absence of magnitude variations and strongly material dependant. This was evaluated in listening test III. The models had fixed magnitude but even-order phase responses between 2^{nd} and 8^{th} order. The cut-off frequency was 60 Hz; distortion therefore occurred through a region where rhythm section instruments (typically kick drum and bass guitar) have fundamental frequencies. It is also a realistic cut-off for mid-sized mix monitors. It was not expected that the audible effects would be so clearly demonstrated; even before post-screening, the results showed that low-frequency phase distortion, approximating that observed in a loudspeaker's roll-off through the bass region, is audible in music reproduction. It was further concluded that different programme material influenced listener judgements. Based on the chosen extracts, it was concluded that the distortion was not audible with fast music having a complex arrangement, but definitely audible with slower, less complex arrangements with more resonant elements in the rhythm section. The findings suggest that this type of music should be mixed on the most accurate monitors because it is most susceptible to inaccurate reproduction by mix monitors. It should be noted that this group

contained only simple Butterworth filters, therefore having a nominal system quality factor, Q_{TS}, of 0.71 [30]. It is expected that alignments with different quality-factors would produce changes in $\overline{M}$ of a different value. Increasing the value of Q_{TS} above 0.50 increases oscillatory behaviour in a loudspeaker's impulse response [18]; it is therefore expected that the algorithm would show greater variations in $\overline{M}$ for systems with a higher Q.

There is a wealth of information relating to phase perception, though many studies have been somewhat 'clinical', using artificially created signals and with a focus on detecting the thresholds of audibility. Many authors have demonstrated that changes in the phase relationship between components of a signal are not only detectable but produce different perceived effects, such as changes in timbre and pitch [40–42, 222–228]. Experiments directly relevant to the application considered in this study are rare, i.e. those that investigate the audible effects of phase distortion in loudspeakers at low frequencies, auditioned with music. However, the findings of Deer *et al.* [209] and Preis *et al.* [212] are relevant here, as they demonstrated that group delay in all-pass filters is audible when either: i) the difference in duration of ringing in the impulse response exceeds 2 ms, or ii) the duration of the ringing is constant but the shape of the impulse response envelope differs. The Group III models were equivalent to a minimum-phase filter cascaded with all-pass filters of increasing group delay, and it was seen that the difference between their impulse responses met both of these criteria (demonstrated in Fig. 47). Therefore, the significant finding from listening test III is not that phase distortion is audible, but that phase distortion of the order introduced by loudspeakers at low frequencies is audible even to the majority of non-expert untrained listeners when using music for audition: most of the differences were observed even before post-screening, and the sample was small so the critical threshold was high, i.e. a strong consensus amongst participants was required to conclude a significant result. This is notable because several authors have concluded that phase distortion to the degree introduced by loudspeakers is either imperceptible or barely detectable, especially when using music for evaluation [26, 27]. The conclusion regarding the different musical characteristics that make differences audible at low frequencies is notable because it is consistent with the observations reported by Fazenda *et al.* [7]. It seems that there can be no dispute that the effect is strongly dependent on certain temporal features, but further work is needed to formally define the characteristics that make certain music more revealing.

The chosen test method was a simple comparative method featuring, unknown to participants, two types of trial. The first of these was regarded as true ABX following the constant-reference Duo-trio protocol [146]; listeners had to identify which of loudspeaker A or B was identical to known reference X. These 'hidden reference' trials were included so that the results could be used as a basis for post-screening; data from only the most acute and consistent listeners was extracted and analysed separately from the sample as a whole. The hypothesis was that listeners who correctly identified the hidden reference consistently would perform well in the other comparisons. These featured the second type of trial, formally described as paired comparison with reference (PCwR): neither A nor B was identical to the reference; listeners had to choose the one that sounded most like X. These PCwR trials were expected to be more difficult, because they required judgement as well as sufficient acuity to first detect differences between A and B. Both of these skills are required in mix monitoring but it was unknown prior to testing how inexperienced listeners would respond to the task. Findings from listening tests I and II in

relation to this strategy[‡‡] are summarised in four main points:

1) In listening test I the proportion of significant pair results was the same for the PCwR trials (judgement of difference) as for the hidden reference trials (true ABX, discrimination only): both were equal to 80 % of tested pairs (8/10 and 5/5 respectively, based on post-screened data). As explained in section 4.1.1, the dominant difference between these models was in overall output level; the difference may therefore be regarded as approximately unidimensional. In Group II, the proportion of significant results differed between the hidden reference and PCwR comparisons: 100 % vs 40 % respectively (5/5 and 4/10, based on post-screened data). Differences between these models were not as simple as an overall drop in low frequency output; they featured a range of alignments that deviated from the reference response in different ways throughout the low-frequency region. It is acknowledged that the available data is limited, but the results indicate that performance in discrimination-only and judgement tasks can be equivalent if the alignment differences are simple, allowing comparisons against the reference to be made based on a single perceptual criterion (overall level of bass); performance in judgement tasks drops considerably if differences between the loudspeakers involves a more complicated 'weighing up' of relative performance [174]. The results for Group II in particular support the assumption made whilst designing the tests that gathering data via a more complicated method, such as gradings, would have required extensive training; if the group could not agree on even the direction of response (A or B) in 40 % of pairs, they could not be expected to agree on overall numerical ratings.

2) Listeners who consistently identified the hidden reference (true ABX trials, discrimination-only), tended to respond more consistently as a group, even when judgement was also required (PCwR): the most accurate listeners in a direct comparison task were overall in better consensus with each other about judgements where there was no a-priori correct answer. When significant outcomes were returned, the A/B split was stronger for this group of listeners; of the 30 pairs evaluated across listening tests I and II, 25 showed this trend (83 %). This supports the hypothesis that hidden reference performance is a reliable indicator of inter- and intra-listener consistency in trials where no direct comparison is present. This strategy may be regarded as inefficient due to the introduction of many extra trials to the experiment, but their inclusion was concluded to be a very useful feature of the test design as the value of post-screening was clear in the vast majority of pairs. It was further concluded that this hidden reference method could be used before similar studies to select participants.

3) The trend across both experiments was to produce a clear outcome at the extremes of reproduction accuracy ('best and worst') but lack of clarity in the middle. It was inferred that when the loudspeakers being compared had less simplistic differences between their alignments it made the judgement about which was most like the reference more challenging. The binomial threshold technique chosen for analysis required a specified level of consensus amongst the group of listeners before a significant result was concluded; pairs with insufficient consensus therefore failed to produce a clear directional outcome e.g. A>B. Group I had 3 out of 15, Group II had 9 out of 15 (20 % and 40 % of pairs respectively, based on post-screened data). The main problem with the chosen strategy is considered to be the fact that it did not allow any conclusion about

[‡‡]Listening test III featured only true ABX trials.

audibility vs indecision; unlike the direct comparison (hidden reference) trials, it could not be concluded with certainty whether listeners were split in their opinions because they could not detect differences between A and B, or whether they concluded that A and B were both equally different from X. Either could be the cause of a lack of consensus for A or B. It is believed that the 2AFC approach was useful in forcing a judgement in these trials where the 'correct' decision was not a straightforward detection of differences, but it cannot be concluded with certainty based on the experimental data; this would have required additional response options such as: *equally different* and *identical*. It was suggested that a two-stage response method could be used in future: participants would first have to confirm that they can detect a difference between A and B. If they respond in the affirmative, they are allowed to perform the comparison (vote for A or B); otherwise, the response of 'no difference' is recorded and the next trial begins. This approach is similar to the combined PEST-ABX method developed by Stephenson [220] to increase accuracy in subjective data when evaluating the perception of low frequencies in critical listening spaces. Such a strategy would provide clarity about which alignments are audibly identical and which are just hard to compare. This would help inform future investigations: lack of decision would imply further training in the task is required; inaudibility would suggest selection of participants based on hearing acuity or a review of the reproduction conditions.

4) A limitation of the subjective data is that the sample of listening test participants was not representative of the target population: professional mix engineers. Before testing, it was considered unrealistic to expect professionals to attend unpaid experiments for a method that had at the time no evidence to suggest that it was useful. As such, it was decided that testing must be performed whilst accommodating the known limitations. The expected sample of listeners therefore strongly influenced the experimental design; the approach was to make the test accessible to participants of little or no experience, without needing extensive training and the significant commitment of time that this would require. In this way, the aim was to gather responses from many participants, creating a large sample upon which to draw conclusions. As already discussed, findings from listening test I and II indicate that this approach is not efficient for further investigation into this topic. Results at the extremes of reproduction accuracy within the tested alignments are already clear. Results from the more challenging alignment comparisons were inconclusive even after post-screening; listeners were not in consensus about these judgements so there was no clear directional result for direct comparison with the mean algorithm scores. This supported the proposition that detailed evaluation of a loudspeaker's bass reproduction accuracy requires experience and learned judgement; being able to detect subtle differences is not sufficient. It is therefore concluded that there is little benefit in performing further subjective validation of the method with untrained listeners.

The comparison between listening test data and algorithm results was presented in chapter 8. Clear directional outcomes were returned in 24 out of 33 pairs evaluated across listening tests I to III (73 %), based on post-screened data; this was used for the main comparison with MTF results for the reasons of increased acuity and consistency observed in this data set, as already described. The subjective results were ordinal, so only the direction of subjective results was clear e.g. A>B; the MTF mean scores for a model group were reduced to ordinal data for comparison. Interval scores may be derived from paired comparison data using Thurstone's Law of Comparative Judgement [189]; this has been demonstrated in relation to subjective impression at

low frequencies, under the Case V assumptions [7, 62]. It was not attempted here but it remains an interesting area for exploration; this might be useful strategy for initial experiments with the target population of listeners. Useful guidance on testing the assumptions required to apply this technique to a data set were provided by Mosteller [229] and Guilford [191]; this testing would need to be performed before applying the method to the data gathered in this study.

Conclusions from comparison of the objective and subjective results are summarised in three main points:

1) For all three listening tests the unweighted mean MTF score ($\overline{M}$) predicted all listener judgements where defined outcomes from ordinal pairwise evaluations and indirect rankings existed: $\overline{M}$ agreed with all significant results from the listening tests. The $\overline{M}$ scores also matched the indirect partial rankings derived from listener judgements of Group I and II models.

2) The changes in $\overline{M}$ between the Group III models were very small (a maximum of 0.005) compared to those observed in Groups I and II, but listener consensus was very strong given revealing material. This does not necessarily indicate that perceptible differences were very obvious, only that they were consistently detectable. Informal feedback from participants indicated that the nature of the differences was quite subtle, sometimes presenting as a difference in relative level between certain instruments, especially the bass guitar. This might indicate that more subtle audible effects correspond with smaller changes in MTF score, but it is not strictly possible to conclude this due to the nature of the subjective data and the way it was analysed. It might also be an indication that the algorithm does not sufficiently reflect the perceptual impact of phase distortion. This would be an interesting area for further investigation, but the results are not considered to indicate a serious limitation of the method; it was seen that variation with modulation frequency - the primary indicator of phase distortion in this method - was an important feature in forming the characteristic intensity image behaviour, so the algorithm is certainly not insensitive to this aspect of reproduction. Based on intensity image behaviour seen in other systems, such as the Group I and II sealed-cabinet models, the apparent insensitivity of the algorithm for Group III is likely to be a consequence of using an inherently well-controlled (Butterworth) alignment to model the phase distortion.

3) The MTF matrix detail, and therefore the intensity image, is helpful in understanding listener judgements. The STI averages the MTF results and outputs only a single numerical score; previous work in evaluation of monitors at low frequencies went as far as using overall and band-mean scores. The present study developed the intensity image format to allow simple inspection of trends in the numerical results matrix. As described in section 9.1.1, this was shown to be a very useful format for objective analysis, but it was not known whether the detail would be useful when considering listener judgements. Based on the results from this study, it is concluded that this detail is informative when interpreting subjective results. Considering pairs with clear directional outcomes, it has been described that the mean scores predicted the results; but looking at the intensity images (as shown in Figs. 42, 45, and 50) it was seen that the 'winner' selected by listeners in each of these pairs was the loudspeaker with an overall whiter intensity image, i.e. they had more individual modulation index scores closer to 1 (the theoretical maximum score possible), producing an intensity image with more lighter-shaded pixels compared to the other model being evaluated. This seems appropriate given that the judgements

were based on similarity to the reference models, all of which had the whitest overall intensity image within their respective groups.

Detail in the MTF results matrix was also found to be useful when considering why some pairs failed to return a consensus vote from listeners. Pair DG in Group II was highlighted as a good example of this, described as a 'classic' sealed vs ported-cabinet comparison. For such pairs, the algorithm results were compared in more detail, first through inspection of the intensity images, and then through confirmation with the numerical values upon which they are based. This investigation also included consideration of the relative spectral weighting of the extracts used for evaluation: it was of interest to see whether differences between the algorithm results for virtual loudspeakers in these pairs coincided with a particular increase or lack of bass content in the corresponding frequency ranges. From this comparison a trend was observed across the pairs in listening test I and II which did not reach a consensus vote; one model in a pair showed lower modulation index scores than the other, except in a certain region of the results matrix; these regions coincided with bands having peak or very strong signal content. Therefore, in these pairs it was not immediately clear what the 'correct' listener response should have been. It was seen that models using reflex loading (a resonance) to increase the low frequency output produced localised regions in the MTF matrix of higher m scores, but this apparent increase in performance did not extend to higher modulation frequencies. The effect in the intensity images was a localised bright spot with pronounced variation in shading in the vertical direction (the modulation frequency axis). This identifies the behaviour as a resonance and shows that performance in the time domain is degraded: the signal envelope is not transmitted faithfully for all rates of temporal fluctuation; however, the trade-off is an increase in output level, shown by the lighter pixels in an otherwise dark region. It was proposed that this behaviour might explain the lack of consensus in listener votes for such pairs. It seemed that some listeners were voting for the model with the 'bright spot'– making it closer to the reference in terms of output level in the affected band, a band which was seen to contain strong signal content and therefore presumed to be particularly audible; conversely, other participants appeared to have voted for the model which more closely matched the behaviour of the reference in terms of vertical shading, indicating that it was more accurate in terms of transient performance but less like the reference in terms of sensitivity. It must be stated that this is not a firm conclusion, but a hypothesis that would benefit from further investigation; the split of subjective results in these pairs was within the limits defined for sampling error, so it is possible that the lack of a consensus vote was due to inability to detect differences between A and B rather than a difficulty in selecting which one was most like X (a limitation of the testing method, as already discussed). In such pairs the algorithm favoured the models with the better temporal performance, seen as less variation in vertical shading; the overall $\overline{M}$ score for alignments with localised increase in m scores did not sufficiently bias the overall score. As such, the algorithm returned a lower mean score in comparison to a loudspeaker with slightly less bass output but having a more accurate alignment in terms of transient performance. Further subjective evaluation of such pairs by the intended audience would be particularly interesting as there is evidence that professional mix engineers favour temporal fidelity over increased output at low frequencies [8,36]; if this is true, it is expected that the direction of subjective results would be in agreement with that predicted by the algorithm.

The conclusions in point 3) prompted some investigation into possible adjustments of the MTF results to account for additional factors that might affect subjective judgements: the increase in hearing threshold at low frequencies, and the relative spectral balance of programme material. The first of these was based on the minimum audible field (MAF) correction developed by Holland *et al.* [64] when applying the MTF to monitors at low frequencies; they noted that a drop in reproduction level is equivalent to reducing the 'useful bandwidth' of a loudspeaker if part of the response falls below the threshold of hearing. The correction made a significant difference to the band-mean scores, up to approximately 60 % of the maximum theoretical value in the lowest bands. This is believed to be because the Schroeder equation was used for computation. As described in section 9.1.1, that method is insensitive to a loudspeaker's output level, so scores in the lowest bands, where reproduction is most likely to naturally fall below the threshold of hearing, made an appreciable contribution to the unadjusted scores. The MAF correction therefore introduced a penalty for drop in output that was not inherent to the method. For the method developed in this project, which does account for changes in the magnitude response, the MAF correction was seen to give a maximum reduction in $\overline{M}$ of 0.013; this indicated that the reproduction level used for the experiments was sufficient for almost all content to be above the threshold of audibility. However, it was concluded that the correction should be applied if mixing is to be performed at untypically low levels; a mean score reduction of up to 0.074 was observed when a playback SPL of 70 dB was assumed, swapping the rank positions for two of the models in Group I.

The second subjective adjustment developed a weighting for the MTF band-mean scores based on relative programme spectrum level. It was concluded that a weighting scheme of this type had little practical use. It could only be applied post-hoc, as was attempted here, because the exact programme spectrum must be known to develop the weights; this makes the correction difficult to implement in a generalised method. However, it was concluded that this adjustment might be useful if it is believed that participants have made judgements based on perceived level of bass reproduction alone; this was suspected to have been the case in some of the judgements in listening test II. It was also suggested that it might be a useful adjustment in other applications, such as evaluating loudspeakers for domestic use, as higher levels of bass have been shown to increase listener preference [61].

9.2 Suggestions for Further Work

The recommendations for further work focus on two areas: increasing the predictive power of the method through further subjective validation, and improving relevance of the technique to the intended application through modifications to the algorithm. The points considered to be most beneficial from each are discussed below.

i) **Target population.** As described in section 9.1.2, participants in this study generally found it hard to reach a consensus about judgements when comparing loudspeakers with the more complex alignment differences. As such, it is believed that further subjective validation must be conducted with professional mix engineers, experienced in making this kind of critical judgement on a regular basis. The outcome of such evaluations would ideally be numerical ratings that can be directly compared with algorithm scores, or at least transformed reliably to values between 0 and 1. Construction of such a scale requires careful development, for the

reasons described in section 5.2.2, but it is recommended that the terms developed by Stephenson [220] be used as a basis for the descriptors and anchor points. These terms succinctly describe the key elements of bass reproduction accuracy; although they were defined in relation to perception of low frequencies inside critical listening spaces, they are suitable for use in further subjective investigations of the MTF algorithm and application addressed in this study. In this way, repeated subjective assessment might allow construction of a reliable subjective grading system for the mean algorithm scores. This is a similar approach to the scale developed for STI, where the numerical scores are compared with a standardised category scale that gives the method a useful diagnostic quality [65, Annex F, G].

ii) **Experimental loudspeakers.** It was explained in section 3.2.2 that the loudspeaker simulations used in this study are believed to have been sufficiently realistic to characterise the key alignment differences in real mix monitors; due to modification of only the low-frequency response, they did not allow investigation of how the coexistence of mid- and high-frequency variations affects listener judgements. The use of a single 'real' loudspeaker in this study fixed a number of factors that affects behaviour in these regions; these include cabinet edge diffraction, number and arrangement of drive units, as well as their crossovers and individual responses. However, it is considered unlikely that such factors would dominate subjective impression in the intended application of this MTF method; professional engineers are used to listening critically to all aspects of the recorded music, and can focus in on particular features of interest [4, 6]. Perhaps more relevant is investigation of the influence of secondary design factors affecting the low frequency region; these are known to vary across real monitors, and include internal cabinet damping, panel resonances, and dimensions and location of the port, where applicable [6, 124, 125].

iii) **Listening Environment.** The MTF algorithm developed in this study is intended to describe free-field behaviour; this allows inspection and comparison of performance without confounding effects of the listening space. Subjective assessment was performed in an anechoic environment to improve accuracy of the simulations and ensure the most sensitive listening conditions for a potentially difficult task. The typical mixing control room will be well damped, but it will not be, and *should* not be, anechoic [230]. It cannot be ascertained from this study how the presence of reflections and reverberation affects the relationship between listener impression and algorithm results for a monitor. The combined monitor-plus-room response will also vary according to the mounting arrangement [2, 93, 95], likely to be on stands or the mixing console for the type of mid-sized monitors considered in this project. It would be useful to be able to reliably combine results from separate analysis of the monitors and the room/ mounting behaviour. Engineers could then see how the results of any monitor would vary when situated in a particular (their own) control room.

iv) **Test Material.** It was concluded from listening test III that certain types of music make it easier to hear aspects of a loudspeaker's low-frequency behaviour that the algorithm describes; such revealing material should therefore be used in further subjective validation of the algorithm. This requires extra work to develop a collection of specific test extracts. From this it might be established on a more formal basis which types of music require the most accurate monitors when mixing, that is, when inspection of a monitor's MTF results will be especially useful. It is

suggested that the type of models featured in Group III, though somewhat unrealistic, would be helpful in identifying such material.

It is possible that subjective assessment by the target listeners will lead to some refinement of the current algorithm to make it a more effective predictive measure. In that event, the following suggestions are recommended based on the findings from this study:

a) Repeat modulation frequency analysis with revealing music: Following recommendation iv), the analysis described in section 2.6 may be repeated using only this critical material. The algorithm in this project was developed with the intention that results should not be strongly biased towards any one type of music, thereby making it generally applicable. Based on the subjective findings from listening test III, there is evidence that it might be more appropriate to select modulation frequencies based only on music with specific temporal characteristics.

b) Penalty for reproduction where $m > 1$: In section 4.2.3 it was described that the algorithm does not properly account for resonances which produce elements in the MTF matrix greater than the theoretical maximum value of 1; as for any behaviour that produces $m < 1$, this should be treated as a distortion. It may be shown in further testing that this problem leads to overestimation of a monitor's reproduction accuracy; the mean score is inflated and intensity images show 'perfect' (white) behaviour since the colour scheme was developed to cover the maximum theoretical range of $0 \leqslant m \leqslant 1$. This problem was only observed in one loudspeaker evaluated in this study (F, Group I) and the effect did not affect the correlation of mean score with listener judgements, so there is insufficient corresponding subjective data to suggest an appropriate method of correcting or adjusting for such distortions. It is believed that the mean-band output plots, adopted from the preceding use of the MTF in this application [64], would no longer be needed if further work could inform a method for applying such a penalty; as described in section 4.6, it was concluded that the only advantage of the mean-band plot over the intensity images was that it showed bands where $m > 1$ distortion occurred.

Of the two types of recommendation suggested here, it is believed that the most useful work lies in further validation of the method through extended subjective evaluation because this will confirm whether modifications to the algorithm are required. It is expected that the judgements of professional mix engineers will give greater insight into the relationship between subjective impression of bass reproduction accuracy and objective results, in particular, the averaged numerical scores. These engineers are likely to be highly critical and experienced listeners due to the nature of their work, capable of extended listening to subtle differences in reproduction; this was shown to have been required for some types of loudspeaker comparison presented in this study. Following input from these listeners, the method may be fully validated and optimised for its intended application. It may then be put to use by engineers when selecting tools to perform their job more effectively and consistently, and ultimately, it can help them to deliver a more powerful and engaging musical experience to the consumers of the material they are creating.

Appendix A. Results from Comparison of MTF Methods

The results for System 1, $h_\delta(t)$, are shown in Table 25. Each element of the 4-by-4 matrix shows the calculated value of modulation ratio m to 2 d.p. for a specific $f_{\mathrm{m,cf}}$ combination. The band-mean scores, $\overline{m}_{\mathrm{cf}}$, are also shown. Methods 2–4 were calculated using 50 noise iterations. The methods are summarised as:

Bandlimited system (BLIR) Method 1: Schroeder expression; Method 2: Formal simulation (modulated noise).

Bandlimited signal (BLIP) Method 3: Band-limit noise before modulating; Method 4: Band-limit noise after modulating.

Table 26 shows the corresponding results for System 2, $h_{\mathrm{F}}(t)$.

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.90	0.77	0.53	0.07	0.57
50	0.90	0.77	0.53	0.07	0.57
70	0.89	0.75	0.50	0.01	0.54
500	0.90	0.77	0.53	0.07	0.57

(a) Method 1

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.85	0.75	0.60	0.09	0.57
50	0.87	0.76	0.60	0.11	0.58
70	0.85	0.76	0.61	0.05	0.57
500	0.87	0.77	0.60	0.12	0.59

(b) Method 2

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00
70	1.00	1.00	1.00	1.00	1.00
500	1.00	1.00	1.00	1.00	1.00

(c) Method 3

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	1.00	1.00	1.00	1.00	1.00
50	1.00	1.00	1.00	1.00	1.00
70	1.00	1.00	1.00	1.00	1.00
500	1.00	1.00	1.00	1.00	1.00

(d) Method 4

Table 25: Results for System 1, $h_\delta(t)$; $\beta = 20$ Hz, centred on f_{cf}

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.87	0.69	0.43	0.05	0.51
50	0.89	0.74	0.50	0.06	0.55
70	0.89	0.74	0.49	0.01	0.53
500	0.90	0.77	0.53	0.07	0.57

(a) Method 1

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.82	0.70	0.50	0.10	0.53
50	0.86	0.73	0.56	0.10	0.56
70	0.85	0.75	0.58	0.02	0.55
500	0.87	0.75	0.61	0.13	0.59

(b) Method 2

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.11	0.11	0.12	0.12	0.12
50	0.28	0.28	0.29	0.30	0.29
70	0.48	0.48	0.48	0.49	0.48
500	1.00	1.00	1.00	1.00	1.00

(c) Method 3

f_{cf} (Hz) \ f_{m} (Hz)	**2**	**5**	**10**	**20**	$\overline{m}_{cf}$
30	0.10	0.10	0.09	0.14	0.11
50	0.27	0.27	0.26	0.29	0.27
70	0.47	0.47	0.47	0.49	0.48
500	1.00	1.00	1.00	1.00	1.00

(d) Method 4

Table 26: Results for System 2, $h_{\mathrm{F}}(t)$; $\beta = 20$ Hz, centred on f_{cf}

Appendix B. Modulation Spectrum: Tracklist and Genre Classifications

#	Track Name	Artist/ Composer	Genre
1	Concerto for Violin and Oboe in C Minor, BWV 1060R: II (Adagio)	Bach	Classical (chamber orchestra)
2	Requiem Op.48, Pie Jesu	Faure	Classical (choral-orchestral)
3	Le Merle Noir	Messian	Classical (flute & piano)
4	Violin Concerto No.1 In G Minor Op.26: II (Adagio)	Bruch	Classical (orchestral)
5	Piano Concerto No.1: Rondo (Vivace)	Chopin	Classical (orchestral)
6	Masques Et Bergamasques, Op.112: IV (Pastorale)	Faure	Classical (orchestral)
7	Hungarian Rhapsody No.2	Liszt	Classical (orchestral)
8	Piano Concerto No.2 in A Major S.125: III (Allegro Moderato)	Liszt	Classical (orchestral)
9	Symphony No.5: IV (Adagietto)	Mahler	Classical (orchestral)
10	Flute Concerto No.1 In G Major, K313: III (Rondò Tempo Di Menuetto)	Mozart	Classical (orchestral)
11	VI. Menuetto - Trio I - Trio II	Mozart	Classical (orchestral)
12	Piano Concerto No.2 in F Major, Op.102: II (Andante)	Shostakovich	Classical (orchestral)
13	The Sleeping Beauty Op.66: II Pas d'action (Rose Adagio)	Tchaikovsky	Classical (orchestral)
14	Sonata No.1 In C Minor, Op.4 (Finale)	Chopin	Classical (solo piano)
15	Petite symphonie à cordes: Vif et agité	Ravel	Classical (String Quartet)
16	Dance Mania Mix (Intro)	DJ Funk	Dance
17	Everybody Needs A 303	Fatboy Slim	Dance
18	SOS (Digital Dog Remix)	Rhianna	Dance
19	Sureshot	Kapricorn	Dance/ Electronic
20	Thrills	LCD Soundsystem	Dance/ Electronic
21	Love Bug	Ramsey & Fen	Dance/ Electronic
22	Balanced Chaos	Roni Size & Reprazent	Dance/ Electronic
23	Safe Heroin	Tomandandy	Dance/ Electronic
24	Found A Cure	Ultra Nate	Dance/ Electronic
25	Little Wonder	David Bowie	Dance/ Electronic

Continued on next page...

#	Track Name	Artist/ Composer	Genre
26	So Nice	Bebel Gilberto (DJ Marky remix)	Dance/ Urban
27	Don't Wanna Know	Shy FX & T power feat. Di	Dance/ Urban
28	On And On And On	ABBA	Pop
29	Heroes And Villains	Brian Wilson	Pop
30	Overprotected	Britney Spears	Pop
31	Tubthumping	Chumbawamba	Pop
32	In The Heat Of The Morning	David Bowie	Pop
33	When I Live My Dream	David Bowie	Pop
34	Union Of The Snake	Duran Duran	Pop
35	Annie, Let's Not Wait	Guillemots	Pop
36	Losing My Way	Justin Timberlake	Pop
37	It Must Have Been Love	Roxette	Pop
38	Nothing Compares To You	Sinead O'Connor	Pop
39	Good Vibrations	The Beach Boys	Pop
40	Release	Timbaland feat. Justin Timberlake	Pop
41	I Talk Too Much	Just Jack	Pop/ Electronic
42	Birds	Bic Runga	Pop/ Folk Rock
43	21st Century Life	Sam Sparro	Pop/ Funk
44	Substitute	The Who	Pop/ Soft Rock
45	Never Knew Love Like This	Alexander O'Neal	Pop/ Soul
46	Anniversary	Lemar feat. Joss Stone	Pop/ Soul
47	Juicy	Candy Hill	Pop/ Urban
48	Kiss Kiss	Chris Brown	Pop/ Urban
49	Get It Shawty	Lloyd	Pop/ Urban
50	Elektro	Ryan Leslie	Pop/ Urban
51	Put The Needle On It	Danni Minogue	Pop/ Urban
52	Love Song	David Jordan	Pop/ Urban
53	Livin' It Up	Jah Rule	Pop/ Urban
54	We Belong Together (Remix)	Mariah Carey feat. Jadakiss	Pop/ Urban
55	Waterloo Sunset	Barb Jungr	Easy Listening
56	We Three Kings Of Orient Are	Beach Boys	Easy Listening
57	Dance With Me	Earl Klugh	Easy Listening
58	Green Sleeves	Nigel North	Easy Listening
59	When I Fall In Love	Claire Martin	Easy Listening/ Jazz
60	Black Coffee	Claire Martin	Easy Listening/ Jazz
61	El Condor Pasa (If I Could)	Simon & Garfunkel	Easy Listening/ Folk
62	Walking Man	James Taylor	Easy Listening/ Folk
63	Danny Boy	James Galway, The Chieftains	Easy Listening/ Folk
64	Here And Now	Luther Vandross	Easy Listening/ Soul

Continued on next page...

#	Track Name	Artist/ Composer	Genre
65	Lovin' You	Minnie Riperton	Easy Listening/ Soul
66	When The Faction's Factioned	Biffy Clyro	Indie Rock
67	Head Up High	Black Rebel Motorcycle Club	Indie Rock
68	The Wizard Turns On	Flaming Lips	Indie Rock
69	Public Pervert	Interpol	Indie Rock
70	So Real	Jeff Buckley	Indie Rock
71	Screenager	Muse	Indie Rock
72	Orson Vodafone Mixtape	Orson	Indie Rock
73	Ann Don't Cry	Pavement	Indie Rock
74	When You Were Young	The Killers	Indie Rock
75	X-Ray	The Maccabees	Indie Rock
76	That Joke Isn't Funny Anymore	The Smiths	Indie Rock
77	Confusion	The Zutons	Indie Rock
78	Havana Gang Brawl	The Zutons	Indie Rock
79	Discotech	Young Love	Indie Rock
80	Youth Of Today	Amy MacDonald	Indie Rock/ Folk
81	Eyeless In Holloway	Johnny Flynn	Indie Rock/ Folk
82	I Can't Quit You Baby (Live Version From BBC Sessions)	Led Zeppelin	Classic/ Blues Rock
83	If I Had a Reason	Rory Gallagher	Classic/ Blues Rock
84	Long Distance Runaround	Yes	Classic / Prog rock
85	In The Land Of Grey & Pink	Caravan	Classic/ Folk Rock
86	Down On Me	Big Brother & The Holding Company	Classic/ Blues Rock
87	Zig Zag Wanderer	Captain Beefheart & His Magic Band	Classic/ Blues Rock
88	Mouthful Of Grass	Free	Classic/ Blues Rock
89	Try (Just A Little Bit Harder)	Janis Joplin	Classic/ Blues Rock
90	Somebody To Love	Jefferson Airplane	Classic/ Blues Rock
91	Angel	Jimi Hendrix	Classic/ Blues Rock
92	Welcome To The Machine	Pink Floyd	Classic/ Blues Rock
93	Paint It, Black	Rolling Stones	Classic/ Blues Rock
94	The Changeling	The Doors	Classic/ Blues Rock
95	Total Eclipse Of The Heart	Bonnie Tyler	Soft Rock
96	Garden	Pearl Jam	Soft Rock
97	Can't Turn Back The Years	Phil Collins	Soft Rock
98	Poem To A Horse	Shakira	Soft Rock
99	Ordinary Morning	Sheryl Crow	Soft Rock
100	Rain	The Ben Taylor Band	Soft Rock/ Folk
101	The Man Who Sold The World	David Bowie	Soft Rock/ Pop

Continued on next page...

#	Track Name	Artist/ Composer	Genre
102	Femme Fatale	Velvet Underground & Nico	Soft Rock/ Pop
103	Man Or Animal	Audioslave	Rock/ Metal
104	Going Under	Evanescence	Rock/ Metal
105	Think About You	Guns N' Roses	Rock/ Metal
106	Take My Scars	Machine Head	Rock/ Metal
107	Naked Aggression	Marshall Law	Rock/ Metal
108	Wall Of Fire	Monster Magnet	Rock/ Metal
109	Crop Circle	Monster Magnet	Rock/ Metal
110	True Belief	Paradise Lost	Rock/ Metal
111	The Enemy	Paradise Lost	Rock/ Metal
112	March Of The Dogs	Sum 41	Rock/ Metal
113	Down In It (Shred)	Nine Inch Nails	Rock/ Electronic
114	Child Of The Night	Ludacris	Hip Hop
115	Hit Em Wit Da Hee	Missy Elliott feat. Lil' Kim	Hip Hop
116	Come And Get Me	Timbaland feat. 50 Cent & Tony Yayo	Hip Hop
117	California Love	2 Pac feat. Dr Dre	Hip Hop
118	I Want You	Keith Sweat feat. Royalty and Nasdaq	Hip Hop
119	Afro Puffs	The Lady Of Rage feat. Dr Dre & Snoop Doggy Dogg	Hip Hop
120	It's Like That	Run-DMC vs. Jason Nevins	Hip Hop/ Dance
121	Bad Boy	Kano	Hip Hop/ Pop
122	Sweet Mother	Skepta	Rap
123	So Rotten	Blak Twang	Rap
124	Bubbles	Dizzee Rascal	Rap
125	Where's Da G's	Dizzee Rascal	Rap
126	Fu-Gee-La	Fugees	Rap
127	Reggae Reggae Bump	R. Kelly feat. Elephant Man	Rap/ Dance
128	Doin' It Again	Skepta	Rap/ Electronic
129	Throw It On Me	Timbaland feat. The Hives	Rap/ Electronic
130	Bow E3	Wiley	Rap/ Electronic
131	Do Right Woman, Do Right Man	Aretha Franklin	Rhythm and Blues
132	The Thrill Is Gone	B. B. King	Rhythm and Blues
133	The Promised Land	Chuck Berry	Rhythm and Blues
134	I'd Rather Go Blind	Etta James	Rhythm and Blues
135	Need Your Love So Bad	Fleetwood Mac	Rhythm and Blues
136	One Bourbon, One Scotch, One Beer	John Lee Hooker	Rhythm and Blues
137	I've Been Hurt So Many Times	Larry Davis	Rhythm and Blues
138	The Right Time	Ray Charles	Rhythm and Blues

Continued on next page...

#	Track Name	Artist/ Composer	Genre
139	5-10-15 Hours	Ruth Brown	Rhythm and Blues
140	Come Fly With Me	Tina May	Jazz
141	Blue Rondo A La Turk	Dave Brubeck Quartet	Jazz
142	This Time Around	S.O.U.L.	Soul
143	I See You	Best Man	Soul
144	Got To Be With You Tonight	Bobby Womack	Soul
145	Summer In The City	The Drifters	Soul
146	Kiss My Love Goodbye	Bettye Swann	Soul/ Disco
147	Continental Square Dance	Joe Bataan	Soul/ Disco
148	Never Too Much	Luther Vandross	Soul/ Funk
149	Mister Magic	Grover Washington Jr.	Soul/ Funk
150	Minute By Minute	Larry Carlton	Soul/ Funk
151	Funkin' For Jamaica	Tom Browne	Soul/ Funk
152	Bnh	Brand New Heavies	Funk
153	Thank You (Falettinme Be Mice Elf Agin)	Sly & The Family Stone	Funk
154	Too High	Stevie Wonder	Funk
155	Don't You Worry 'Bout A Thing	Hank Crawford	Jazzfunk
156	Music Of The Mind	Jamiroquai	Jazzfunk
157	Expansions	Lonnie Liston Smith	Jazzfunk
158	Always There	Ronnie Laws & Pressure	Jazzfunk
159	Birdland	Weather Report	Jazzfunk
160	Black Is The Color	Wilbert Longmire	Jazzfunk
161	H.A.P.P.Y. Radio	Edwin Starr	Disco
162	Love Sensation	Loleatta Holloway	Disco
163	Baby Come Back	Pato Banton and the Reggae Revolution	Reggae
164	Punkie	Sean Paul	Reggae
165	Monkey Man	Toots & The Maytals	Reggae
166	Falling In Love With You (I Can't Help)	UB40	Reggae
167	Bamboleo	Gipsy Kings	Flamenco
168	Espiritu	The Guitar Trio	Flamenco

Total number of tracks = 168.

Table 27: Extracts used for analysis of musical modulation frequencies

Appendix C. Derivation of the Lumped Parameter Model

This derivation is based on the model provided by Holland [231].

x_d, x_p = Displacement of the diaphragm and mass of air in the port.

A_d, A_p = Effective piston areas of the diaphragm and port.

S_d, S_c = Stiffness of the diaphragm and air in the cabinet.

m_d, m_p = Mass of the diaphragm and air in the port.

R_d, R_p = Damping of the diaphragm and port.

Calculate S_c from:

$$S_c = \frac{\rho_0 c_0^2 A_d^2}{V_c} \tag{C.1}$$

where: ρ_0 is density of air, c_0^2 is the speed of sound in air, and V_c is the volume of air inside the loudspeaker cabinet.

Equate the forces acting on the two moving masses (diaphragm mass and mass of air in the port):

$$F = m_d\ddot{x}_d + R_d\dot{x}_d + S_d x_d + S_c\left(x_d + \left\{\frac{A_p}{A_d}\right\}x_p\right) \tag{C.2}$$

$$0 = m_p\ddot{x}_p + R_p\dot{x}_p + S_c\left\{\frac{A_p}{A_d}\right\}\left(x_d + \left\{\frac{A_p}{A_d}\right\}x_p\right) \tag{C.3}$$

Let: $F = Fe^{\mathrm{j}\omega t}$ and $x = xe^{\mathrm{j}\omega t}$, so that: $\dot{x}$ = derivative of x, i.e. $x = \mathrm{j}\omega x e^{\mathrm{j}\omega t}$, and $\ddot{x} = -\omega^2 x e^{\mathrm{j}\omega t}$.

If $u = \dot{x}$, then $ue^{\mathrm{j}\omega t} = \dot{x}e^{\mathrm{j}\omega t}$; therefore: $u = \dot{x} = \mathrm{j}\omega x$.

Divide by $\mathrm{j}\omega$: $\frac{\mathrm{j}\omega x}{\mathrm{j}\omega} = x = \frac{u}{\mathrm{j}\omega}$.

Then:

$$S_c x_d = \frac{S_c u}{\mathrm{j}\omega} \tag{C.4}$$

Now:

$$Fe^{\mathrm{j}\omega t} = \left(\mathrm{j}\omega m_d u_d + R_d u_d + \frac{S_c u_d}{\mathrm{j}\omega} + \frac{S_c u_d}{\mathrm{j}\omega} + \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\}u_p\right)e^{\mathrm{j}\omega t} \tag{C.5}$$

$$0e^{\mathrm{j}\omega t} = \left(\mathrm{j}\omega m_p u_p + R_p u_p + \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\}u_d + \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\}^2 u_p\right)e^{\mathrm{j}\omega t} \tag{C.6}$$

Note: u_d and u_p are diaphragm and port velocity respectively.

Cancel $e^{\mathrm{j}\omega t}$ to rewrite (C.5) and (C.6) as:

$$F = B_{11}u_d + B_{12}u_p \tag{C.7}$$

$$0 = B_{21}u_d + B_{22}u_p \tag{C.8}$$

where:

$$B_{11} = \mathrm{j}\omega m_d + R_d + \frac{S_d}{\mathrm{j}\omega} + \frac{S_c}{\mathrm{j}\omega} \tag{C.9}$$

$$B_{12} = \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\} = B_{21} = \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\} \tag{C.10}$$

$$B_{22} = \mathrm{j}\omega m_p + R_p + \frac{S_c}{\mathrm{j}\omega}\left\{\frac{A_p}{A_d}\right\}^2 \tag{C.11}$$

Rearrange (C.7) and (C.8) to get mechanical impedance, Z_m:

$$Z_m = \frac{F}{u_d} \rightarrow \frac{F}{u_d} = B_{11} + \frac{B_{12}u_p}{u_d} \tag{C.12}$$

And:

$$u_p = -\frac{B_{21}u_d}{B_{22}} \tag{C.13}$$

So:

$$\frac{F}{u_d} = B_{11} + \frac{B_{12}}{u_d}\left(-\frac{B_{21}u_d}{B_{22}}\right) = B_{11} - \frac{B_{12}B_{21}\cancel{u_d}}{\cancel{u_d}B_{22}} = B_{11} - \frac{B_{12}B_{21}}{B_{22}} \tag{C.14}$$

Also from (C.8):

$$\frac{u_p}{u_d} = -\frac{B_{21}}{B_{22}} \tag{C.15}$$

The force acting on the diaphragm, F, is the product of the electrical current through the driver voice coil, I, and the coefficient ϕ:

$$\phi = Bl \tag{C.16}$$

where: B is the magnetic field strength of the magnet and l is the length of the voice coil.

$$F = \phi I \tag{C.17}$$

The voltage across the voice coil, V, is:

$$V = Z_{\text{eb}} I + \phi u_d \tag{C.18}$$

Where: Z_{eb}, the blocked electrical impedance, is defined as:

$$Z_{\text{eb}} = \left.\frac{V}{I}\right|_{u=0} = R_{\text{eb}} + \text{j}\omega L_{\text{eb}} \tag{C.19}$$

Note: R_{eb} and L_{eb} are the electrical resistance and inductance respectively.

From (C.17), (C.18),and (C.19), the velocity of the driver moving mass can be written:

$$u_d = \frac{\phi V}{Z_{\text{eb}} Z_m + \phi^2} \tag{C.20}$$

The combined volume velocity output of the loudspeaker (from diaphragm and port) is given by the vector sum:

$$q = u_d A_d + u_p A_p \tag{C.21}$$

This is used to calculate the on-axis sound pressure, p:

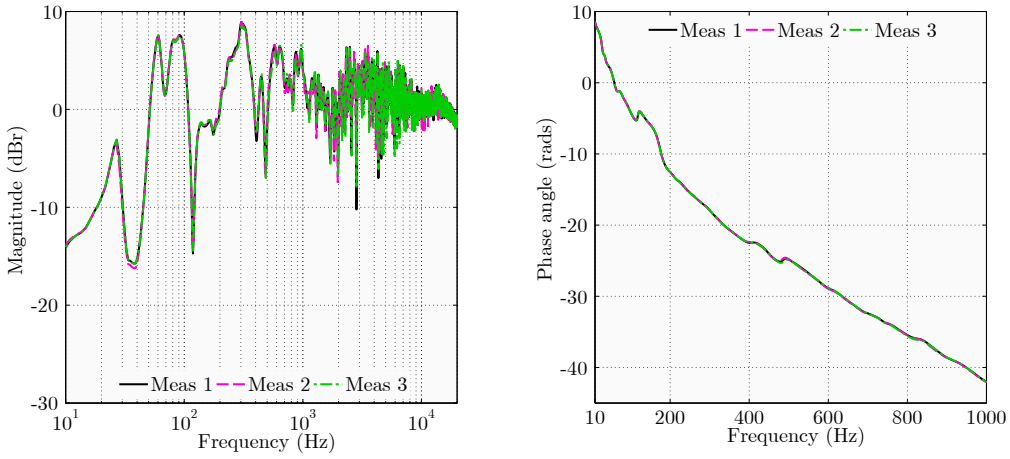
$$p = \text{j}\omega\rho_0 q \frac{e^{-\text{j}\omega\frac{r}{c_0}}}{2\pi r} \tag{C.22}$$

where: c_0 is the speed of sound in air, ρ_0 is the density of air, and r is the distance from the diaphragm.

Appendix D. Comparison of Large- and Small-Chamber Loudspeaker Measurements

This section compares the results from measurement of the experimental loudspeaker response in two locations considered for testing: a large anechoic chamber, and a smaller semi-anechoic chamber. The loudspeaker port was blocked in each case, as described in section 3.3.1.1.

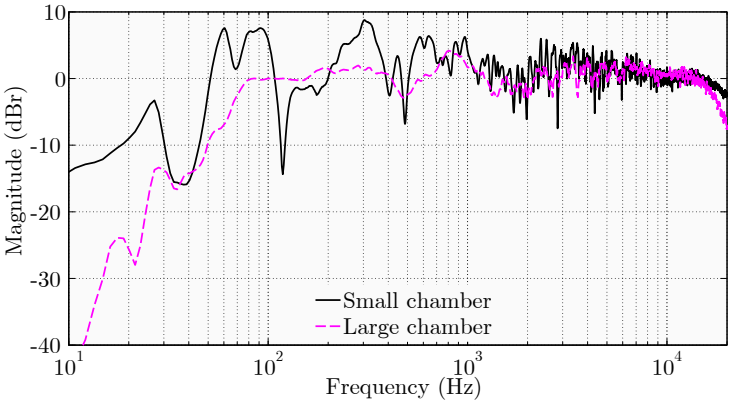
It was expected that the small-chamber measurement would be affected by reflections because the room had not been emptied of unnecessary equipment. However, the floor space between source and receiver was covered with acoustic wedges, approx. 30 cm deep made of foam, to prevent strong reflections from this hard surface. Mid- and high-frequency reflections were not of primary concern during these measurements; the focus was to assess the low-frequency response of the experimental loudspeaker in this space. Figure 86 shows three repeated measurements taken on the same day. These were generated using the same process described in section 3.3.1. It can be seen that the results are consistent at low frequencies across measurements.



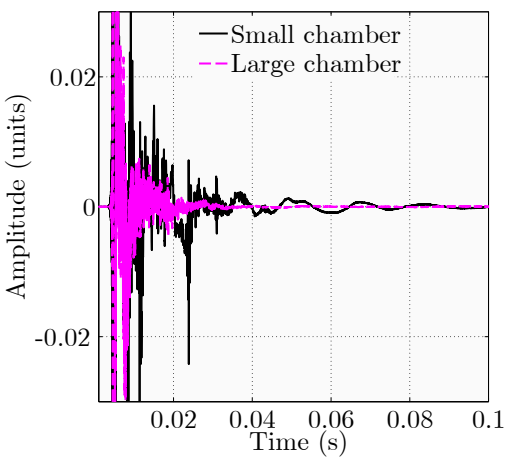
(a) Magnitude **(b)** Phase (unwrapped)

Figure 86: Three measurements of the experimental loudspeaker taken in the small chamber

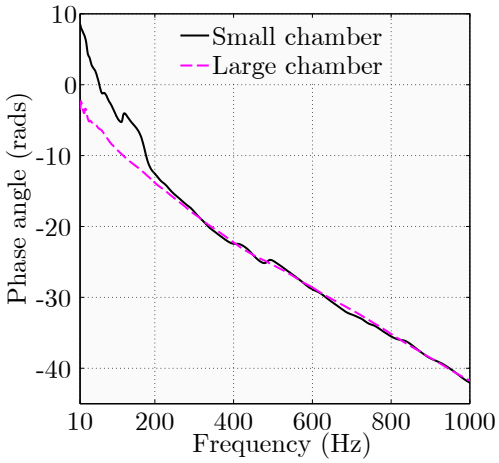
Figure 87 shows the average of these three measurements compared to the average of the three large-chamber anechoic measurements of the experimental loudspeaker shown in Fig. 32. The impulse responses have been aligned to correct for any slight difference in source-receiver separation between the two locations (by applying an integer sample delay, as described in section 3.2.2). The magnitude responses in Fig. 87a show large peaks and deep notches present in the small chamber measurement at low frequencies. The plot of the impulse responses in Fig. 87b has been cropped to make it easier to see that the small chamber measurement has long-term fluctuations. In Fig. 87c it can be seen that the phase responses differ considerably at low frequencies; in the small chamber the response has much greater non-linear phase shift, and is overall less smooth compared to the large-chamber equivalent.



(a) Magnitude



(b) Impulse responses (partial view)



(c) Phase (unwrapped)

Figure 87: Comparing large- and small-chamber behaviour: experimental loudspeaker measurements

Appendix E. List of Experimental Equipment

Table 28 lists the experimental equipment used during measurement and subjective experimentation.

Item	Serial number	Notes
JBL LSR32 reference monitor	0108639	Loudspeaker
Custom made wooden bung (covered in black tape) and adhesive tac	-	To seal port; covered to reduce visual distraction
B&K Falcon 1/2" pre-polarised 0V condenser mic, Type 4189	2285285	Measurement mic
B&K Nexus conditioning amplifier	2165583	
Microphone stand and clip	-	
RME ADI-8 DS ADAT AD/DA converter	ISVR SV7702 / SV9703	8-Channel ADAT®/TDIF AD/DA Converter
RME ADI-648 Multichannel audio digital interface	-	64-Channel 24 Bit/96 kHz
Dell Desktop PC (Windows 7, 64-bit, 4GB RAM, 3.3GHz processor) with RME Hammerfall MADI I/O card	2200282	128 Channel 96 kHz I/O Card
PC monitor	-	
MATLAB R2102a 32-bit	-	
Adobe Audition v.3.0	-	
Sound level meter	ISVR 3540	
Data transfer switch (2-VGA input switchbox)	-	Switches between large and touch screens
2off 69cm tall loudspeaker stands	-	For placememnt of touchscreen and switchbox
2off 61cm tall loudspeaker stands	-	For placement of JBL
4off rubber mounts	-	Interface between JBL and stands
56x56cm piece of plywood	-	Place between grids and JBL stands
Office chair and cable ties	-	Height adjustable. Cable ties for fixing
Acoustic foam wedges	-	To dampen sound of amp and PC fans
4off 5 m phono-phono cables	-	
5 m USB extension lead	-	For touchscreen to PC
6-socket extension leads	-	
Loudspeaker cable	-	Heavy gauge OFC
Wooden table	-	For placing amp and PC

Table 28: Equipment used for measurement and experimentation

Appendix F. Model Parameters

This shows the additional electroacoustic parameters used to create the virtual loudspeakers using the *WooferMaker* program. Names of the actual units that these parameters describe are also listed for reference under the heading *Driver*. Cabinet volume in all cases was fixed at 25 l. Note that the reference, R, responses are not shown; these were modelled as 2^{nd}-order Butterworth high-pass filters and therefore had no input parameters other than f_c. The design parameters for Group I and II models are listed in Tables 29 and 30 respectively.

Label	Driver	A_d (m^2)	V_{as} (m^3)	f_d (Hz)	Q_{ms} (-)	Bl (Tm)	R_e (Ω)	f_p (Hz)
C	JBL 116A	0.0180	0.0736	28	5.00	6.70	5.2	-
D	Bumper 648X	0.0100	0.0028	74	8.30	9.70	3.5	-
E	Volt LS201	0.0201	0.0250	40	3.90	10.80	5.4	10
F	Scan-Speak 18W/8535-00	0.0145	0.0720	26	2.50	5.70	5.8	26
G	Focal 6V 415	0.0125	0.0256	40	3.78	9.17	7.8	40

Table 29: Group I model input parameters

Label	Driver	A_d (m^2)	V_{as} (m^3)	f_d (Hz)	Q_{ms} (-)	Bl (Tm)	R_e (Ω)	f_p (Hz)
C	Davis 17MRP	0.0125	0.0145	77	2.64	5.07	6.2	-
D	Swan W6	0.0139	0.0251	40	3.04	6.7	6.5	-
E	Davis 16 GKLV6M	0.0141	0.0177	60	2.55	7.84	6.0	65
F	Davis 20 GKLV8 TDF	0.0208	0.0195	77	2.79	8.52	6.1	67
G	Eminence LA6-MB	0.0133	0.0125	73	6.6	8.34	7.4	58

Table 30: Group II model input parameters

Appendix G. MTF Results Listing

Full numeric results for all experimental systems are listed in Tables 31 to 33 for reference. The modulation frequencies and band centre frequencies are denoted by f_{cf} and f_m respectively.

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.63	0.86	0.94	0.97	0.98	0.97	0.96	0.95	0.94	0.94
8.5	0.69	0.90	0.96	0.97	0.98	0.98	0.97	0.96	0.94	0.95
5.8	0.73	0.93	0.96	0.97	0.99	0.98	0.98	0.96	0.94	0.95
4.3	0.75	0.94	0.97	0.98	0.99	0.98	0.98	0.96	0.94	0.95
2.2	0.77	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.95
1.1	0.77	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.95
0.8	0.78	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.96

(a) R (reference)

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.22	0.47	0.68	0.83	0.92	0.96	0.97	0.96	0.94	0.95
8.5	0.21	0.46	0.69	0.85	0.93	0.96	0.97	0.96	0.94	0.95
5.8	0.20	0.46	0.69	0.85	0.94	0.97	0.98	0.96	0.94	0.96
4.3	0.20	0.46	0.70	0.85	0.94	0.97	0.98	0.96	0.94	0.96
2.2	0.20	0.45	0.70	0.85	0.94	0.97	0.98	0.96	0.95	0.96
1.1	0.20	0.45	0.70	0.85	0.94	0.97	0.98	0.97	0.95	0.96
0.8	0.20	0.45	0.70	0.86	0.94	0.97	0.98	0.97	0.95	0.96

(b) C

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.09	0.18	0.26	0.35	0.42	0.49	0.54	0.58	0.62	0.66
8.5	0.09	0.18	0.26	0.35	0.42	0.49	0.55	0.58	0.62	0.66
5.8	0.09	0.17	0.26	0.35	0.42	0.49	0.55	0.59	0.62	0.66
4.3	0.08	0.17	0.26	0.35	0.42	0.49	0.55	0.59	0.62	0.66
2.2	0.08	0.17	0.26	0.34	0.43	0.49	0.55	0.59	0.62	0.66
1.1	0.08	0.17	0.26	0.35	0.43	0.49	0.55	0.59	0.62	0.66
0.8	0.08	0.17	0.26	0.34	0.43	0.49	0.55	0.59	0.62	0.67

(c) D

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.20	0.36	0.51	0.63	0.72	0.78	0.81	0.83	0.84	0.86
8.5	0.19	0.36	0.51	0.63	0.72	0.78	0.82	0.83	0.84	0.87
5.8	0.19	0.36	0.51	0.63	0.72	0.78	0.82	0.84	0.84	0.87
4.3	0.18	0.35	0.51	0.63	0.73	0.78	0.83	0.84	0.84	0.87
2.2	0.18	0.35	0.51	0.63	0.73	0.79	0.83	0.84	0.84	0.87
1.1	0.18	0.35	0.51	0.63	0.73	0.79	0.83	0.84	0.84	0.87
0.8	0.18	0.35	0.51	0.63	0.73	0.79	0.83	0.84	0.84	0.87

(d) E

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.21	0.61	0.84	0.99	1.05	1.06	1.04	1.00	0.98	0.98
8.5	0.22	0.61	0.89	1.02	1.06	1.06	1.04	1.01	0.98	0.98
5.8	0.21	0.62	0.92	1.04	1.07	1.06	1.04	1.01	0.98	0.98
4.3	0.21	0.62	0.93	1.05	1.07	1.06	1.05	1.01	0.98	0.99
2.2	0.21	0.62	0.94	1.05	1.08	1.06	1.05	1.01	0.98	0.99
1.1	0.21	0.62	0.95	1.06	1.08	1.06	1.05	1.01	0.98	0.99
0.8	0.21	0.61	0.95	1.06	1.08	1.06	1.05	1.01	0.98	0.99

(e) F

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.13	0.47	0.59	0.70	0.75	0.77	0.78	0.79	0.80	0.83
8.5	0.13	0.52	0.68	0.72	0.75	0.77	0.79	0.80	0.80	0.83
5.8	0.13	0.57	0.74	0.73	0.75	0.77	0.79	0.80	0.80	0.83
4.3	0.13	0.60	0.76	0.74	0.75	0.77	0.79	0.80	0.80	0.83
2.2	0.13	0.62	0.78	0.74	0.75	0.77	0.80	0.80	0.80	0.83
1.1	0.13	0.63	0.79	0.74	0.75	0.77	0.80	0.80	0.80	0.83
0.8	0.13	0.63	0.79	0.74	0.75	0.77	0.80	0.80	0.81	0.83

(f) G

Table 31: MTF matrix results for Group I models

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.63	0.86	0.94	0.97	0.98	0.97	0.96	0.95	0.94	0.94
8.5	0.69	0.90	0.96	0.97	0.98	0.98	0.97	0.96	0.94	0.95
5.8	0.73	0.93	0.96	0.97	0.99	0.98	0.98	0.96	0.94	0.95
4.3	0.75	0.94	0.97	0.98	0.99	0.98	0.98	0.96	0.94	0.95
2.2	0.77	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.95
1.1	0.77	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.95
0.8	0.78	0.95	0.97	0.98	0.99	0.99	0.98	0.96	0.94	0.96

(a) R (reference)

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.07	0.17	0.31	0.46	0.60	0.72	0.80	0.85	0.87	0.90
8.5	0.07	0.17	0.30	0.45	0.61	0.73	0.81	0.85	0.87	0.90
5.8	0.07	0.16	0.30	0.45	0.61	0.73	0.81	0.86	0.87	0.90
4.3	0.07	0.16	0.30	0.45	0.61	0.73	0.82	0.86	0.87	0.90
2.2	0.07	0.16	0.29	0.45	0.61	0.73	0.82	0.86	0.87	0.90
1.1	0.07	0.16	0.29	0.45	0.61	0.73	0.82	0.86	0.87	0.90
0.8	0.07	0.16	0.29	0.45	0.61	0.73	0.82	0.86	0.87	0.90

(b) C

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.21	0.45	0.65	0.81	0.90	0.94	0.95	0.95	0.93	0.93
8.5	0.20	0.44	0.66	0.82	0.91	0.95	0.96	0.95	0.93	0.93
5.8	0.20	0.44	0.66	0.82	0.91	0.95	0.96	0.95	0.93	0.94
4.3	0.19	0.44	0.67	0.83	0.91	0.95	0.96	0.95	0.93	0.94
2.2	0.19	0.43	0.67	0.83	0.92	0.95	0.96	0.95	0.93	0.94
1.1	0.19	0.43	0.67	0.83	0.92	0.95	0.96	0.95	0.93	0.94
0.8	0.19	0.43	0.67	0.83	0.92	0.95	0.96	0.95	0.93	0.94

(c) D

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.01	0.06	0.33	0.68	0.81	0.87	0.88	0.85	0.84	0.85
8.5	0.01	0.06	0.33	0.74	0.88	0.89	0.88	0.86	0.84	0.85
5.8	0.01	0.06	0.33	0.80	0.93	0.90	0.88	0.86	0.84	0.85
4.3	0.01	0.06	0.33	0.83	0.94	0.90	0.88	0.86	0.84	0.85
2.2	0.01	0.06	0.32	0.86	0.96	0.91	0.88	0.86	0.84	0.85
1.1	0.01	0.06	0.32	0.86	0.96	0.91	0.88	0.86	0.84	0.85
0.8	0.01	0.05	0.32	0.86	0.97	0.91	0.88	0.86	0.84	0.85

(d) E

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.02	0.12	0.42	0.71	0.88	0.96	0.98	0.97	0.94	0.95
8.5	0.02	0.12	0.42	0.74	0.91	0.97	0.99	0.97	0.94	0.95
5.8	0.02	0.12	0.41	0.76	0.93	0.98	0.99	0.97	0.94	0.95
4.3	0.02	0.12	0.41	0.76	0.94	0.98	0.99	0.97	0.95	0.95
2.2	0.02	0.12	0.40	0.77	0.95	0.98	0.99	0.97	0.95	0.95
1.1	0.02	0.12	0.40	0.77	0.95	0.98	0.99	0.97	0.95	0.95
0.8	0.02	0.12	0.40	0.77	0.95	0.99	0.99	0.97	0.95	0.95

(e) F

f_{cf} (Hz)

f_m (Hz)	22.8	37.6	52.4	67.2	82.0	96.8	111.6	126.4	141.2	156.0
11.7	0.03	0.20	0.59	0.76	0.85	0.86	0.86	0.85	0.84	0.86
8.5	0.03	0.21	0.63	0.83	0.87	0.86	0.86	0.86	0.84	0.86
5.8	0.03	0.20	0.67	0.88	0.88	0.87	0.87	0.86	0.84	0.86
4.3	0.03	0.20	0.68	0.91	0.89	0.87	0.87	0.86	0.84	0.86
2.2	0.03	0.20	0.70	0.93	0.89	0.87	0.87	0.86	0.84	0.86
1.1	0.03	0.20	0.70	0.94	0.89	0.87	0.87	0.86	0.84	0.86
0.8	0.03	0.20	0.70	0.94	0.89	0.87	0.87	0.86	0.85	0.86

(f) G

Table 32: MTF matrix results for Group II models

f_m (Hz) \ f_{cf} (Hz)	**22.8**	**37.6**	**52.4**	**67.2**	**82.0**	**96.8**	**111.6**	**126.4**	**141.2**	**156.0**
11.7	0.19	0.41	0.66	0.83	0.90	0.92	0.90	0.91	0.89	0.91
8.5	0.17	0.37	0.67	0.84	0.91	0.91	0.91	0.92	0.89	0.92
5.8	0.15	0.35	0.69	0.83	0.90	0.91	0.92	0.93	0.90	0.93
4.3	0.15	0.36	0.69	0.84	0.91	0.91	0.92	0.94	0.90	0.93
2.2	0.16	0.38	0.69	0.85	0.92	0.92	0.92	0.94	0.90	0.93
1.1	0.16	0.40	0.70	0.86	0.92	0.92	0.92	0.94	0.90	0.93
0.8	0.17	0.40	0.70	0.87	0.92	0.92	0.93	0.95	0.90	0.93

(a) z_2

f_m (Hz) \ f_{cf} (Hz)	**22.8**	**37.6**	**52.4**	**67.2**	**82.0**	**96.8**	**111.6**	**126.4**	**141.2**	**156.0**
11.7	0.19	0.41	0.65	0.80	0.87	0.90	0.90	0.91	0.89	0.91
8.5	0.17	0.37	0.67	0.83	0.89	0.91	0.91	0.92	0.89	0.92
5.8	0.15	0.35	0.69	0.83	0.90	0.91	0.92	0.93	0.90	0.93
4.3	0.15	0.36	0.69	0.84	0.90	0.91	0.92	0.94	0.90	0.93
2.2	0.16	0.38	0.69	0.85	0.91	0.92	0.92	0.94	0.90	0.93
1.1	0.16	0.39	0.70	0.86	0.92	0.92	0.92	0.94	0.90	0.93
0.8	0.17	0.40	0.70	0.87	0.92	0.92	0.93	0.95	0.90	0.93

(b) z_4

f_m (Hz) \ f_{cf} (Hz)	**22.8**	**37.6**	**52.4**	**67.2**	**82.0**	**96.8**	**111.6**	**126.4**	**141.2**	**156.0**
11.7	0.19	0.40	0.63	0.74	0.83	0.89	0.91	0.91	0.89	0.91
8.5	0.17	0.37	0.66	0.79	0.87	0.90	0.92	0.92	0.89	0.92
5.8	0.15	0.36	0.68	0.82	0.89	0.91	0.92	0.93	0.90	0.93
4.3	0.15	0.36	0.68	0.83	0.90	0.91	0.92	0.94	0.90	0.93
2.2	0.16	0.38	0.68	0.85	0.91	0.92	0.92	0.94	0.90	0.93
1.1	0.16	0.40	0.69	0.86	0.92	0.92	0.92	0.94	0.90	0.93
0.8	0.16	0.40	0.70	0.86	0.92	0.92	0.93	0.94	0.90	0.93

(c) z_6

f_m (Hz) \ f_{cf} (Hz)	**22.8**	**37.6**	**52.4**	**67.2**	**82.0**	**96.8**	**111.6**	**126.4**	**141.2**	**156.0**
11.7	0.19	0.39	0.61	0.68	0.80	0.88	0.91	0.90	0.89	0.91
8.5	0.17	0.36	0.65	0.75	0.85	0.90	0.92	0.92	0.89	0.92
5.8	0.15	0.35	0.67	0.80	0.88	0.91	0.92	0.93	0.90	0.93
4.3	0.15	0.36	0.68	0.82	0.90	0.91	0.92	0.94	0.90	0.93
2.2	0.15	0.38	0.68	0.85	0.91	0.92	0.92	0.94	0.90	0.93
1.1	0.16	0.39	0.69	0.85	0.91	0.92	0.92	0.94	0.90	0.93
0.8	0.17	0.40	0.69	0.86	0.92	0.92	0.93	0.94	0.90	0.93

(d) z_8

Table 33: MTF matrix results for Group III models

Appendix H. High-Pass Filter Transfer Functions

The high-pass filter transfer function equations used to generate the Group III loudspeaker models, described in section 4.5, are presented here. Even-orders 2 to 8 inclusive were used (12 dB to 48 dB/octave roll-off).

Time constant, τ, is determined by the filter's low-frequency cut-off, f_c:

$$\tau = \frac{1}{2\pi f_c} \tag{H.1}$$

Complex frequency variable, s, is defined throughout as:

$$s = \mathrm{j}2\pi f \tag{H.2}$$

where: f is the angular frequency in Hertz.

2nd order:

$a_1 = \sqrt{2}$;

$$H_2(f) = \frac{s^2\tau^2}{s^2\tau^2 + a_1 s\tau + 1} \tag{H.3}$$

4th order:

$a_1 = a_3 = 2.613$; $a_2 = 3.414$;

$$H_4(f) = \frac{s^4\tau^4}{s^4\tau^4 + a_3 s^3\tau^3 + a_2 s^2\tau^2 + a_1 s\tau + 1} \tag{H.4}$$

6th order:

$a_1 = a_5 = 3.864$; $a_2 = a_4 = 7.464$; $a_3 = 9.141$;

$$H_6(f) = \frac{s^6\tau^6}{s^6\tau^6 + a_5 s^5\tau^5 + a_4 s^4\tau^4 + a_3 s^3\tau^3 + a_2 s^2\tau^2 + a_1 s\tau + 1} \tag{H.5}$$

8th order:

$a_1 = a_7 = 5.126$; $a_2 = a_6 = 13.138$; $a_3 = a_5 = 21.848$; $a_4 = 25.691$;

$$H_8(f) = \frac{s^8\tau^8}{s^8\tau^8 + a_7 s^7\tau^7 + a_6 s^6\tau^6 + a_5 s^5\tau^5 + a_4 s^4\tau^4 + a_3 s^3\tau^3 + a_2 s^2\tau^2 + a_1 s\tau + 1} \tag{H.6}$$

Appendix I. Extract Spectrograms

Figures 88 to 90 contain spectrograms showing power distribution across the full duration of each listening test extract. A logarithmic y-axis has been used to allow easier inspection of signal content at low frequencies. The colours show power (dBr), relative to the peak value in each extract; darkest red indicates frequencies with the strongest power. The colours for all extracts have been mapped across the same scale (-120–0 dB) to make comparisons between them easier. With reference to the discussion in section 5.3.1, it can be seen that the extracts do not feature large deviations in their spectral content or rhythmic pattern. The tracks had a sampling frequency of 44.1 kHz, divided into Hamming-windowed segments 8192 samples long with 8192-point FFTs; segments overlapped by 50 %. The plots therefore show a frequency resolution and temporal spacing of approximately 5 Hz and 0.1 s respectively.

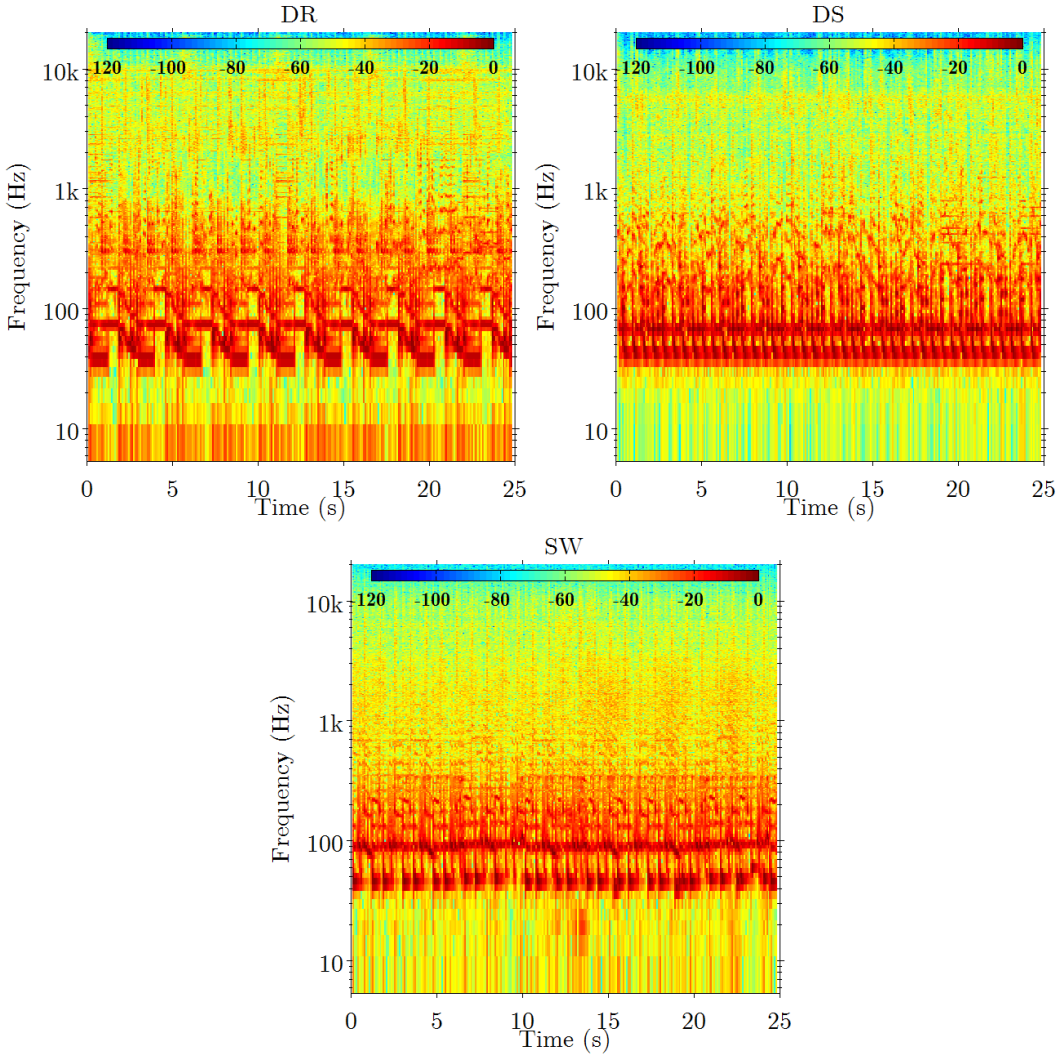


Figure 88: Spectrograms for listening test I extracts

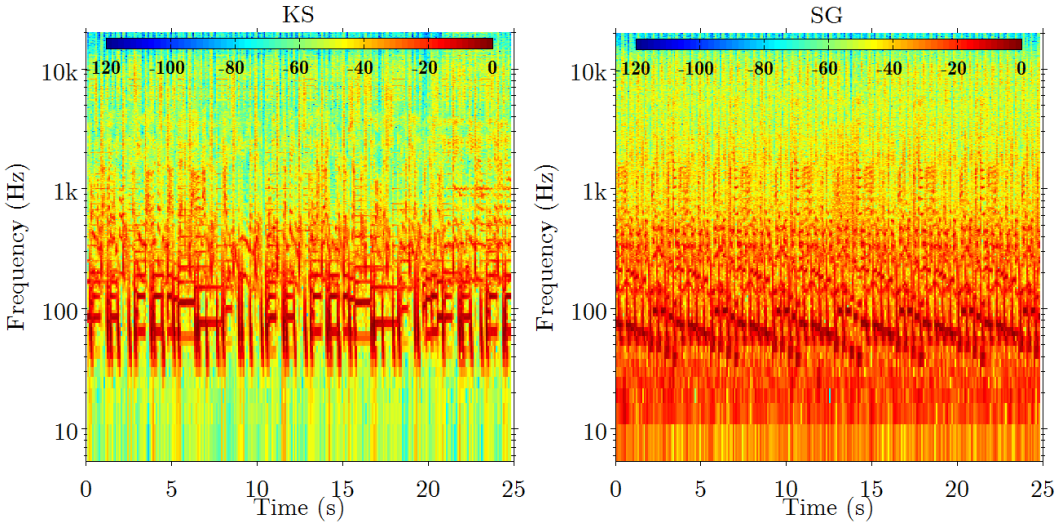


Figure 89: Spectrograms for listening test II extracts

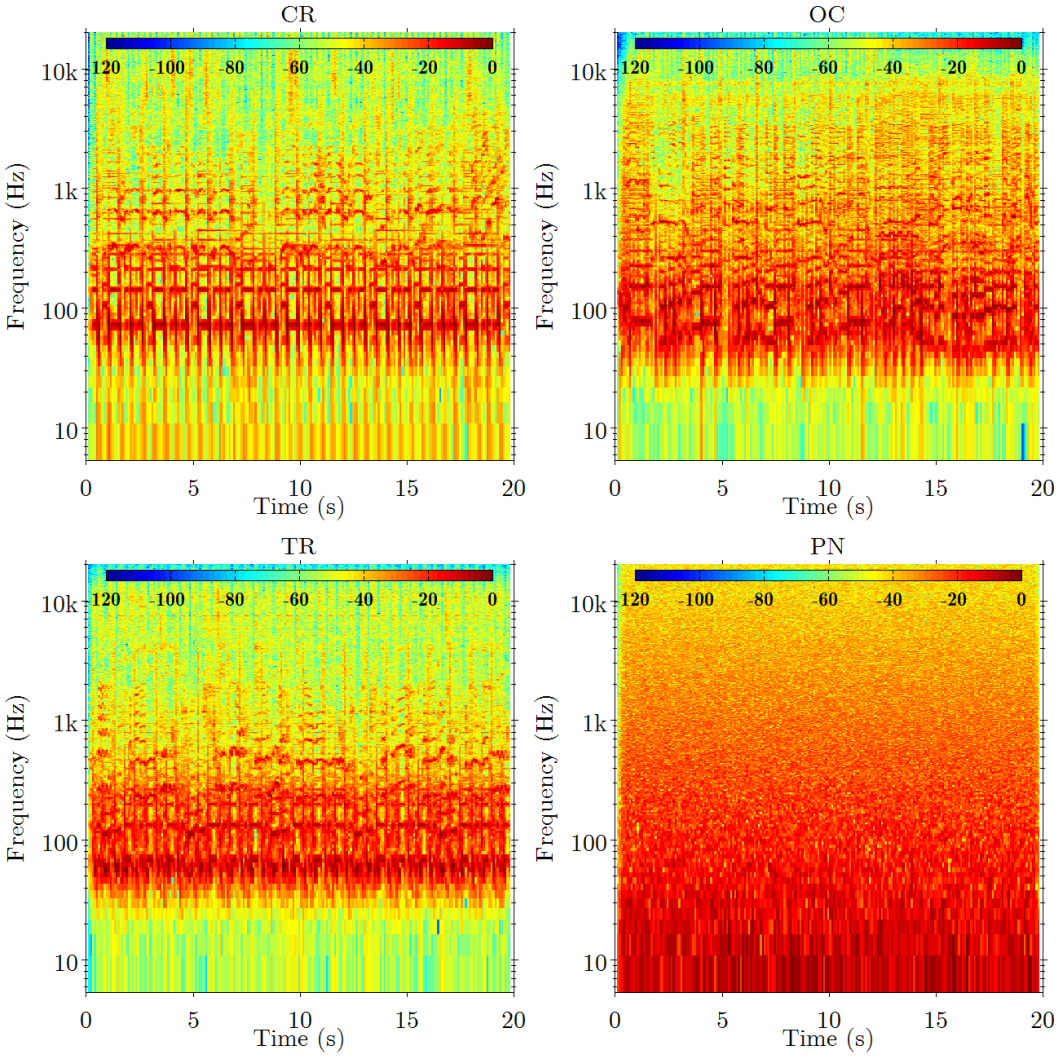


Figure 90: Spectrograms for listening test III extracts

Appendix J. Comparison of Listening Test Pair Results with MTF Scores: No Post-Screening

Table 34 shows the comparison of pair results from each listening test with the corresponding algorithm $\overline{M}$ scores; this is equivalent to the comparison in section 8.1 but the subjective pair results are shown before post-screening the data according to hidden reference trial performance.

A>B	$\overline{M}_A$	$\overline{M}_B$
R>C	0.939	0.795
R>D		0.419
R>E		0.656
R>G		0.692
C>D	0.795	0.419
C>E		0.656
E>D	0.656	0.419
F>D	0.893	0.419
F>E		0.656
F>G		0.692
G>D	0.692	0.419

(a) Listening test I (Group I models)

A>B	$\overline{M}_A$	$\overline{M}_B$
R>C	0.854	0.576
R>D		0.775
R>F		0.706
D>C	0.775	0.576
E>C	0.645	
F>C	0.706	
G>C	0.696	0.576
G>E		0.645
G>F		0.706

(b) Listening test II (Group II models)

A>B	$\overline{M}_A$	$\overline{M}_B$
$z_2 > z_6$	0.758	0.752
$z_2 > z_8$		0.747

(c) Listening test III (Group III models)

Table 34: Comparison of significant listening test pair results with corresponding mean algorithm scores. For pairs containing models R or z_2, $\alpha = 0.025$; otherwise, $\alpha = 0.05$

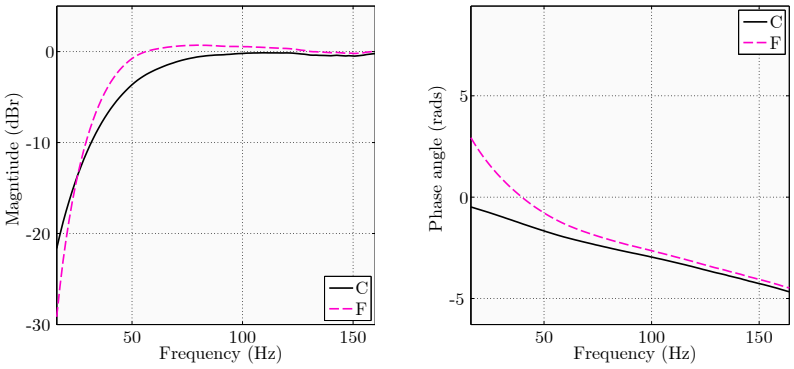
Appendix K. Additional Comparison of Non-Significant Pair Results

Listed here are comparisons of non-significant pair results in listening tests I and II that were not shown in section 8.2.

Listening Test I

Figure 91 shows the steady-state magnitude and phase for the pairs being compared. Frequency axis is limited to the range covered by the algorithm.

Figure 91: Listening test I responses: C vs F

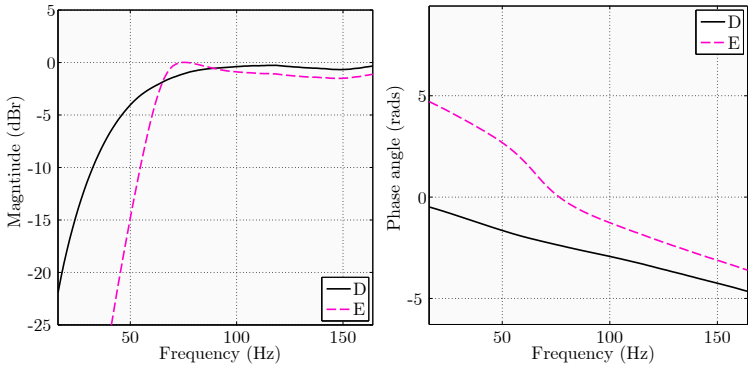


Pair	$\overline{M}$ scores	Subjective result
C vs F	$\overline{M}_F > \overline{M}_C$ (0.89>0.79)	Slight tendency towards F (selected in 26 of 45 trials, 58 %).
Summary:	Loudspeaker C is closer to the reference in bands 5 to 10; F is closer in bands 1 to 4. Bands 2 to 4 cover critical bass region and contain music spectrum peak values.	
Conclusion:	Participants struggled to choose between the loudspeaker more like the reference in lower frequencies (F), and the one more similar otherwise (C); location of peak signal content influenced the split of results slightly towards the former.	

Listening Test II

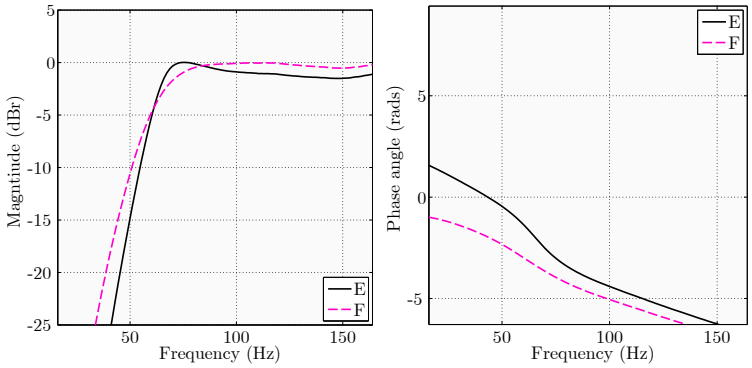
Figures 92 to 95 show the steady-state magnitude and phase for the pairs being compared.

Figure 92: Listening test II responses: D vs E



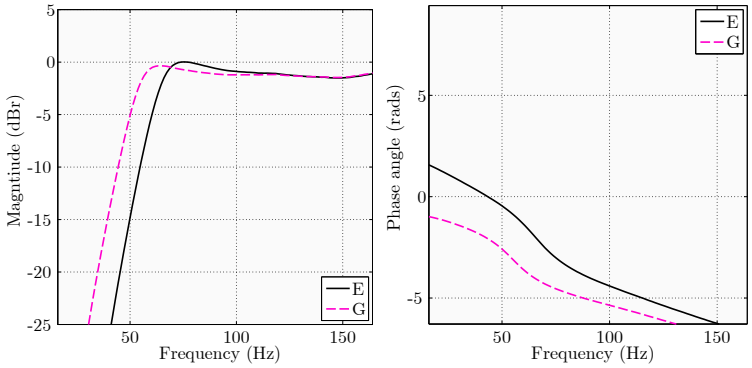
Pair	$\overline{M}$ scores	Subjective result
D vs E	$\overline{M}_D > \overline{M}_E$ (0.78>0.64)	Tends towards D (selected in 16 of 24 trials, 67 %)
Summary:	D is overall closer to the reference than E except for a region in bands 4 and 5. Music spectrum peaks in band 4 and has a strong peak at 82 Hz (in band 5).	
Conclusion:	Majority of participants identified D as closer to the reference but could not reach a clear consensus due to relative superiority of loudspeaker E in regions with peak music content.	

Figure 93: Listening test II responses: E vs F



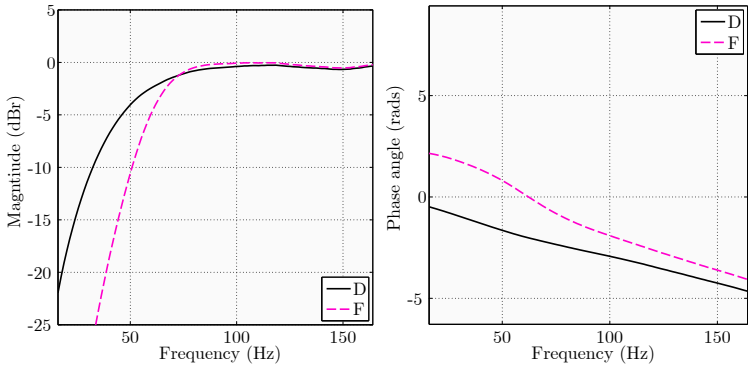
Pair	$\overline{M}$ scores	Subjective result
E vs F	$\overline{M}_F > \overline{M}_E$ (0.71>0.64)	Almost equally split (F selected in 13 of 24 trials, 54 %)
Summary:	F outperforms E overall except for band 4; music spectrum peaks in band 4.	
Conclusion:	Listener decisions split between overall better performance of F and increased performance of E in one band that coincides with music spectrum peak.	

Figure 94: Listening test II responses: E vs G



Pair	$\overline{M}$ scores	Subjective result
E vs G	$\overline{M}_G > \overline{M}_E$ (0.70>0.64)	Tend towards G (selected in 17 of 24 trials, 71 %).
Summary:	Band-mean and intensity images show similar behaviour within this pair, but neither appear more similar to the reference than the other. The systems return the same mean score for bands 8 to 10; G is superior in bands 1 to 4, E outperforms it in bands 5 to 7. Music spectrum is maximum in bands 3 and 4 but has prominent peaks in bands 5 and 6 (82 and 93 Hz respectively).	
Conclusion:	Majority of participants selected G as being superior but clear consensus not reached as E outperforms it in bands containing strongest bass content.	

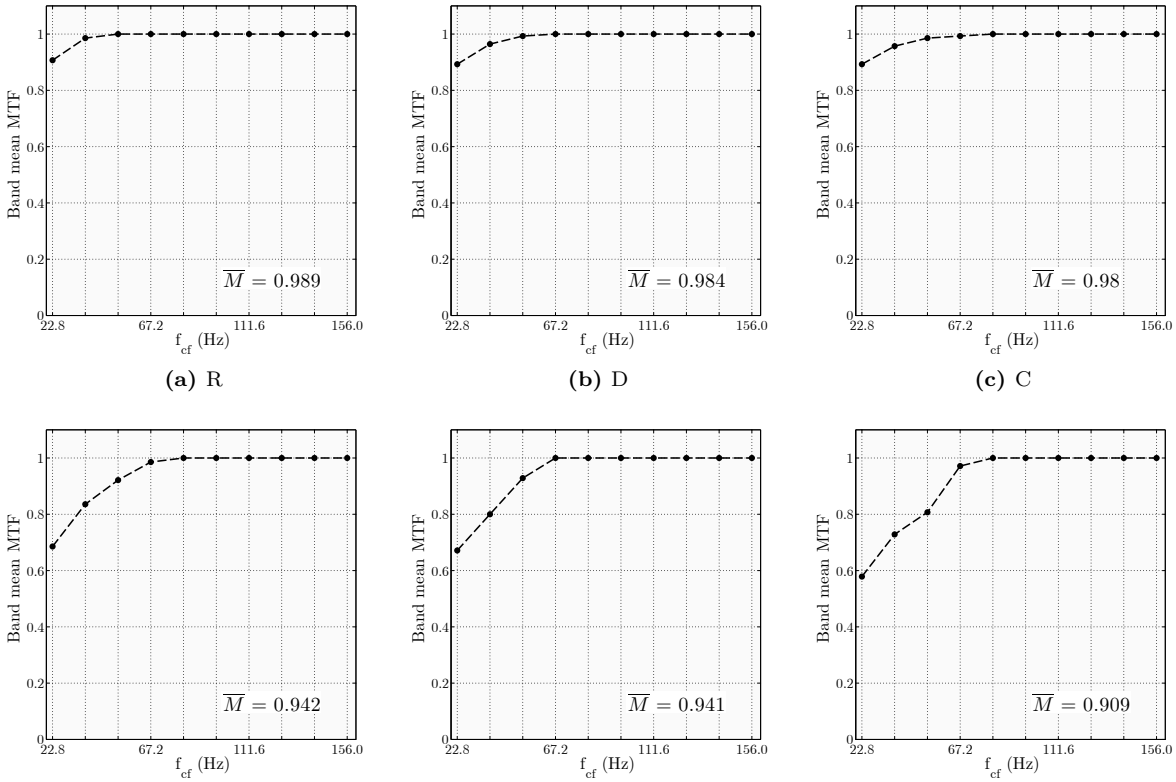
Figure 95: Listening test II responses: D vs F



Pair	$\overline{M}$ scores	Subjective result
D vs F	$\overline{M}_D > \overline{M}_F$ (0.78>0.71)	Almost equally split (D selected in 13 of 24 trials, 54 %)
Summary:	D overall closer to the reference but F shows increased performance in band 5 (contains strong peak around 82 Hz).	
Conclusion:	MTF results indicate that D should have been identified as closer to reference but increased similarity to reference of F in a band with prominent music content affected the comparison and divided opinions almost equally.	

Appendix L. Group II Normalised Schroeder Results

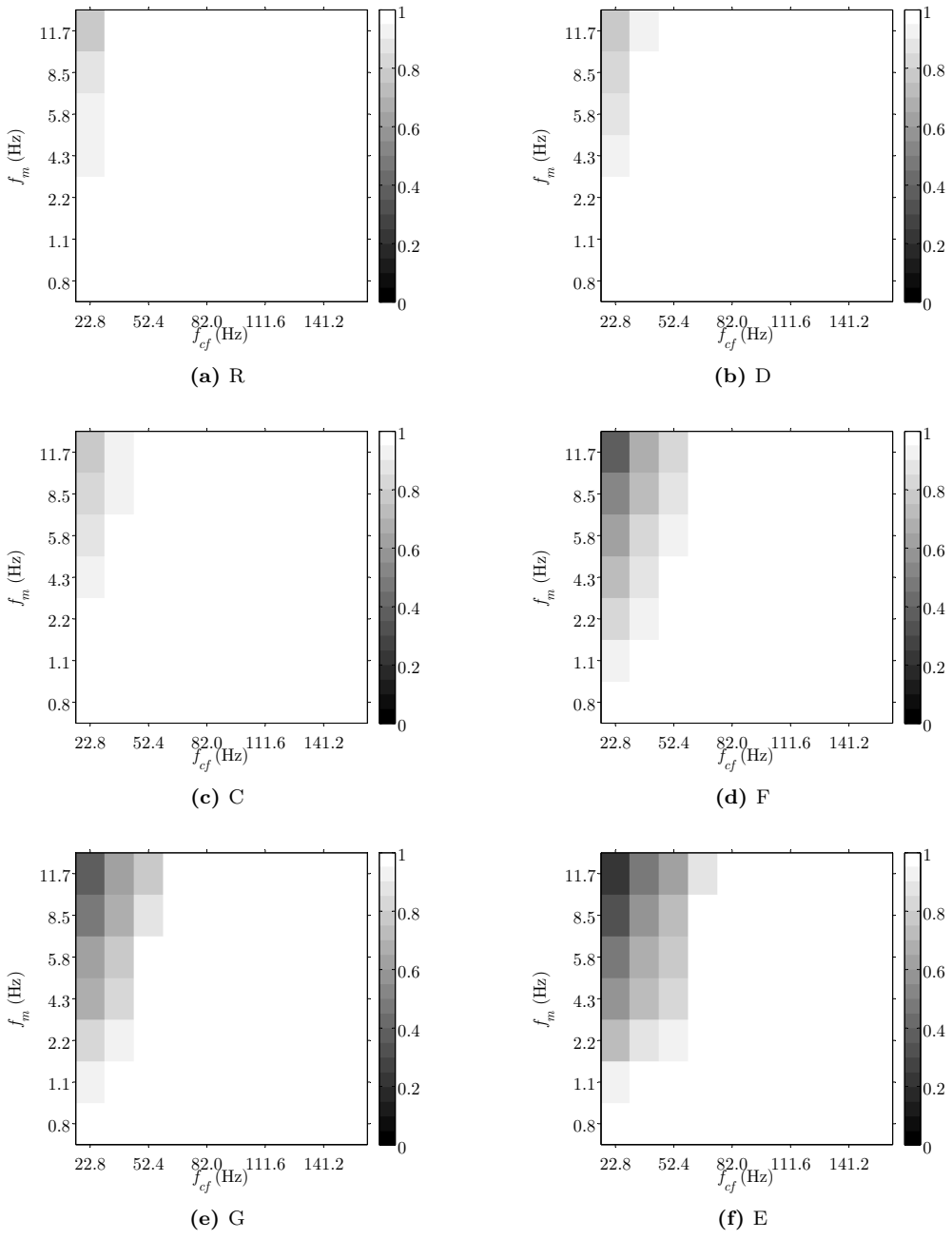
As in chapter 4, the results for each virtual loudspeaker are presented in order of decreasing mean score. The ranking is not the same as that for the final MTF-based method and therefore does not predict the listener judgements as successfully. The mean scores in this case predict subjective outcomes only in pair DC and pairs featuring the reference R.



(a) R **(b)** D **(c)** C

(d) F **(e)** G **(f)** E

Figure 96: Normalised Schroeder method: Group II MTF results– Mean-band plots (from simulated responses)



(a) R **(b)** D

(c) C **(d)** F

(e) G **(f)** E

Figure 97: Normalised Schroeder method: Group II MTF results– Intensity images (from simulated responses)

References

[1] Newell, Philip and Keith Holland. *Loudspeakers: For Music Recording and Reproduction*, Chapter 8: Form Follows Function. Focal Press, 2007.

[2] Goldberg, Andrew, Aki Makivirta, and Ari Varla. 'Compensating the Response of Near-Field Loudspeaker Monitors to Minimise the Effects of Desktop Acoustic Loading'. *Journal of the Audio Engineering Society*, **54**(5):401–411, May 2006.

[3] Toole, Floyd. 'Art and Science in the Control Room'. *Proceedings of the Institute of Acoustics*, **25**(8):243–252, 2003.

[4] Newell, Philip. 'Audiophile vs. Professional Monitoring Priorities'. *HIFICRITIC*, (7):pp.11–14, Jan/Feb. 2008.

[5] Horning, Susan Schmidt. 'Engineering the Performance: Recording Engineers, Tacit Knowledege and the Art of Controlling Sound'. *Social Studies Of Science*, **34**(5):703–731, October 2004.

[6] Mathers, C. 'On the Design of Loudspeakers for Broadcast Monitoring'. Technical Report BBC RD 1988/14, BBC Research Department, December 1988.

[7] Fazenda, Bruno, Matthew Wankling, Jonathan Hargreaves, Lucy Elmer, and Jonathan Hirst. 'Subjective Preference of Modal Control Methods in Listening Rooms'. *J. Audio Eng. Soc.*, **60**(5):338–348, May 2012.

[8] Newell, Philip, Keith Holland, and Julius Newell. 'The Yamaha NS10M: Twenty Years a Reference Monitor. Why?' *Proceedings of the Institute of Acoustics*, **23**(8):29–40, 2001.

[9] Colloms, Martin. 'Basso Profundo'. *Stereophile*, December 1991. URL `http://stereophile.com/reference/39/`. Available online; last accessed 18/02/2016.

[10] Harley, Robert. 'A Guide to Better Bass'. *The Absolute Sound*, **TAS 197**, November 2009. URL `http://www.theabsolutesound.com/articles/a-guide-to-better-bass-tas-197-1/`. Available online; last accessed 18/02/2016.

[11] Hove, Michael, Celine Marie, Ian Bruce, and Laurel Trainor. 'Superior Time Perception for Lower Musical Pitch Explains Why Bass-Ranged Instruments Lay Down Musical Rhythms'. *Proceedings of the National Academy of Sciences*, **111**(28):10,383–10,388, July 2014.

[12] Fielder, Louis and Eric Benjamin. 'Subwoofer Performance for Accurate Reproduction of Music'. *J. Audio Eng. Soc.*, **36**(6):443–456, June 1988.

[13] Andrews, Tony. 'Reality? Or Soft Focus?' *Presented at PLASA for the Gottelier Award Master Class*, 2008. URL `http://www.funktion-one.com/dl/files/1819.pdf`. Available online; last accessed 18/02/2016.

[14] Dyck, Edith Van, Dirk Moelants, Michiel Derney, Alexander Deweppe, Pieter Coussement, and Marc Leman. 'The Impact of the Bass Drum on Human Dance Movement'. *Music Perception*, **30**(4):349–359, April 2013.

[15] Burger, Birgitta, Marc Thompson, Geoff Luck, Suvi Saarikallio, and Petri Toiviainen. 'Influences of Rhythm- and Timbre-Related Musical Features on Characteristics of Music-Induced Movement'. *Frontiers in Psychology*, **4**(183), April 2013.

[16] Fincham, L. 'The Subjective Importance of Uniform Group Delay at Low Frequencies'. *J. Audio Eng. Soc.*, **33**(6):436–439, June 1985.

[17] Thiele, Neville. 'The Loudspeaker Parameters and Their Evolution'. *Proceedings of the Institute of Acoustics*, **31**(4):27–49, 2009.

[18] Small, Richard. 'Closed-Box Loudspeaker Systems– Part I: Analysis'. *J. Audio Eng. Soc.*, **20**(10):798–808, December 1972.

[19] Small, Richard. 'Vented-Box Loudspeaker Systems– Part I: Small-Signal Analysis'. *J. Audio Eng. Soc.*, **21**(5):363–372, June 1973.

[20] Newell, Philip and Keith Holland. *Loudspeakers: For Music Recording and Reproduction*, Chapter 11.2: Commercial Solutions. Focal Press, 2007.

[21] Holland, Keith, Philip Newell, and Peter Mapp. 'Steady State and Transient Loudspeaker Frequency Responses'. *Proceedings of the Institute of Acoustics*, **25**(8):105–118, 2003.

[22] Preis, D. 'Linear Distortion'. *J. Audio Eng. Soc.*, **24**(5):346–367, June 1976.

[23] Heyser, Richard. 'Loudspeaker Phase Characteristics and Time Delay Distortion: Part 1'. *J. Audio Eng. Soc.*, **17**(1):30–41, January 1969.

[24] Heyser, Richard. 'Loudspeaker Phase Characteristics and Time Delay Distortion: Part 2'. *J. Audio Eng. Soc*, **17**(2):130–137, April 1969.

[25] Lipshitz, Stanley and John Vanderkooy. 'The Great Debate: Subjective Evaluation'. *J. Audio Eng. Soc.*, **29**(7-8):482–491, August 1981.

[26] Lipshitz, Stanley, Mark Pocock, and John Vanderkooy. 'On the Audibility of Midrange Phase Distortion in Audio Systems'. *J. Audio Eng. Soc.*, **30**(9):580–595, September 1982.

[27] Blauert, J. and P. Laws. 'Group Delay Distortions in Electroacoustical Systems'. *J. Acoust. Soc. Am.*, **63**(5):1478–1483, May 1978.

[28] Moller, Henrik, Pauli Minnaar, Soren Krarup Olesen, Flemming Christensen, and Jan Plogsties. 'On the Audibility of All-Pass Phase in Electroacoustical Transfer Functions'. *J. Audio Eng. Soc.*, **55**(3):115–134, March 2007.

[29] Preis, D. 'Phase Distortion and Phase Equalization in Audio Signal Processing - A Tutorial Review'. *J. Audio Eng. Soc.*, **30**(11):774–794, November 1982.

[30] Small, Richard. 'Direct-Radiator Loudspeaker System Analysis'. *J. Audio Eng. Soc.*, **20**(5):383–395, June 1972.

[31] Leach, W. Marshall. 'The Differential Time-Delay Distortion and Differential Phase-Shift Distortion as Measures of Phase Linearity'. *J. Audio Eng. Soc.*, **37**(9):709–715, September 1989.

[32] Oppenheim, Alan, Alan Willsky, and Hamid Nawad. *Signals & Systems (2^{nd} edition)*, Chapter 6: Time and Frequency Characterisation of Signals and Systems. Prentice Hall, 1997.

[33] Lathi, B. P. *Modern Digital and Analogue Communication Systems (3^{rd} edition)*, Chapter 3: Analysis and Transmission of Signals. OUP USA, 1998.

[34] Newell, Philip, Keith Holland, and Peter Mapp. 'The Perception of the Reception of a Deception'. *Proceedings of the Institute of Acoustics*, **24**(8), 2002.

[35] Newell, Philip and Keith Holland. *Loudspeakers: For Music Recording and Reproduction*, Chapter 11: Low Frequency and Transient Response Dilemmas. Focal Press, 2007.

[36] Tervo, Sakari, Perttu Laukkanen, Jukka Patynen, and Tapio Lokki. 'Preferences of Critical Listening Environments Among Sound Engineers'. *J. Audio Eng. Soc.*, **62**(5):300–314, May 2014.

[37] Fazenda, Bruno, Matthew Stephenson, and Andrew Goldberg. 'Perceptual Thresholds for the Effects of Room Modes as a Function of Modal Decay'. *J. Acoust. Soc. Am.*, **137**(3):1088–1098, March 2015.

[38] Newell, Philip and Keith Holland. *Loudspeakers: For Music Recording and Reproduction*, Chapter 9: Subjective and Objective Assessment. Focal Press, 2007.

[39] Ewaskio, Charles and Osman Mawardi. 'Electroacoustic Phase Shift in Loudspeakers'. *J. Acoust. Soc. Am.*, **22**(4):444–448, July 1950.

[40] Mathes, R. and R. Miller. 'Phase Effects in Monaural Perception'. *J. Acoust. Soc. Am.*, **19**(5):780–797, September 1947.

[41] Laitinen, Mikko-Ville, Sascha Disch, and Ville Pulkki. 'Sensitivity of Human Hearing to Changes in Phase Spectrum'. *J. Audio Eng. Soc.*, **61**(11):860–877, November 2013.

[42] Plomp, R. and H. Steeneken. 'Effect of Phase on the Timbre of Complex Tones'. *J. Acoust. Soc. Am.*, **2**(2):409–421, 1969.

[43] Colloms, Martin. 'Archive III; Pace, Rhythm, & Dynamics'. *HIFICRITIC (reprinted from Stereophile November 1992)*. URL `http://www.hificritic.com/uploads/2/8/8/0/28808909/classic-sc7-pace_rhythm__dynamics.pdf`. Available online; last accessed 18/02/2016.

[44] King, R., D. Shorter, and T. Somerville. 'The Selection of a Wide-Range Loudspeaker for Monitoring Purposes'. Technical Report M.008, BBC Research Department, February 1948.

[45] Somerville, T. 'High Fidelity Sound Reproduction: A Brief Survey'. Technical Report A.024, BBC Research Department, March 1948.

[46] Presonus. 'Eris Series: TECH SPECS', accessed 23/06/2015. URL `http://www.presonus.com/products/Eris/techspecs`.

[47] Yamaha. 'HS8 Powered Studio Monitor: Specs', accessed 23/06/2015. URL `http://usa.yamaha.com/products/music-production/speakers/hs_series/hs8/`.

[48] KRK. 'ROKIT 5 G3: Technical Specifications', accessed 23/06/2015. URL http://www.krksys.com/krk-studio-monitor-speakers/rokit/rokit-5.html.

[49] equator. 'D5 STUDIO MONITORS: Technical Specs', accessed 23/06/2015. URL http://www.equatoraudio.com/D5-Studio-Monitors-with-DSP-Pair-p/d5.htm.

[50] ATC. 'SCM20ASL Pro: Specification', accessed 23/06/2015. URL http://www.atcloudspeakers.co.uk/professional/loudspeakers/scm20sl-pro/.

[51] ADAM. 'A7X: Technical Data', accessed 23/06/2015. URL http://www.adam-audio.com/en/pro-audio/products/a7x/technical-data.

[52] GENELEC. 'M040 Studio Monitor', Accessed 23/06/2015. URL http://www.genelec.com/studio-monitors/m-series-studio-monitors/m040-studio-monitor.

[53] ADAM. 'A7X Product Sheet', accessed 23/06/2015. URL http://www.adam-audio.com/files/downloads/A7X_engl150dpi.pdf.

[54] ATC. 'SCM20ASL Technical Data Sheet', accessed 23/06/2015. URL http://www.atcloudspeakers.co.uk/wp-content/uploads/2014/08/Technical-Datasheet-SCM20ASL-Pro-Mk2-revB-WEB.pdf.

[55] EVENT. 'Event Opal Specifications', accessed 23/06/2015. URL http://www.eventelectronics.com/downloads/opal/datasheet.pdf.

[56] Focal. 'CMS 65 Professional analog monitoring loudspeaker; Specification Sheet', accessed 23/06/2015. URL http://www.focal.com/en/cms/166-cms-65-3544052801309.html.

[57] KRK. 'ROCKIT 5 G3 Support / Documentation: Series Cutsheet (09/13 Version 1.1)', accessed 23/06/2015. URL http://www.krksys.com/manuals/rokit/Rokit-CutSheet-May14.pdf.

[58] Yamaha. 'HS Series Datasheet', accessed 23/06/2015. URL http://usa.yamaha.com/products/music-production/speakers/hs_series/hs8/.

[59] Neumann. 'KH 120 D - Active Studio Monitor with Digital Input and Delay: Measurements', accessed 23/06/2015. URL http://www.neumann-kh-line.com/neumann-kh/home_en.nsf/root/prof-monitoring_studio-monitors_nearfield-monitors_KH120D.

[60] Toole, Floyd. 'Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 1'. *J. Audio Eng. Soc.*, **34**(4):227–235, April 1986.

[61] Toole, Floyd. 'Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2'. *J. Audio Eng. Soc.*, **34**(5):323–348, May 1986.

[62] Wankling, Matthew, Bruno Fazenda, and William Davies. 'The Assessment of Low-Frequency Room Acoustic Parameters Using Descriptive Analysis'. *J. Audio Eng. Soc.*, **60**(5):325–337, May 2012.

[63] Colloms, Martin. *High Performance Loudspeakers (5^{th} edition)*, Chapter 9: Loudspeaker Assessment. Wiley, 1997.

[64] Holland, Keith, Philip Newell, and Peter Mapp. 'Modulation Depth as a Measure of Loudspeaker Low-Frequency Performance'. *Proceedings of the Institute of Acoustics*, **26**(8), 2004.

[65] 'BS EN 60268-16: Sound System Equipment: Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index', 2011.

[66] Baker, L. 'Status of OTF in 1970'. *Optica Acta: International Journal of Optics*, **18**(2):81–92, February 1971.

[67] Viemeister, Neal. 'Temporal Modulation Transfer Functions Based Upon Modulation Thresholds'. *J. Acoust. Soc. Am.*, **66**(5):1364–1380, November 1979.

[68] Schroeder, M. 'Modulation Transfer Functions: Definition and Measurement'. *Acustica*, **49**(3):179–182, November 1981.

[69] Backman, Juha. 'Complex Modulation Transfer Function and Its Applications in Transducer and Room Acoustics Measurements'. Paper 8172. AES 129th Convention, San Francisco, November 2010.

[70] Steeneken, H. and T. Houtgast. 'A Physical Method for Measuring Speech-Transmission Quality'. *J. Acoust. Soc. Am.*, **67**(1):318–326, January 1980.

[71] Jacob, Kenneth, Thomas Birkle, and Christopher Ickler. 'Accurate Prediction of Speech Intelligibility Without the Use of In-Room Measurements'. *J. Audio Eng. Soc*, **39**(4):232–242, April 1991.

[72] Houtgast, T. and H. Steeneken. 'The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility'. *Acustica*, **28**(1):66–73, January 1973.

[73] Houtgast, T., H. Steeneken, and R. Plomp. 'Predicting Speech Intelligibility in Rooms From the Modulation Transfer Function – 1. General Room Acoustics'. *Acta Acustica United with Acustica*, **46**(1):60–72, September 1980.

[74] Polack, J., H. Alrutz, and M. Schroeder. 'The Modulation Transfer Function of Music Signals and its Applications to Reverberation Measurement'. *Acustica*, **54**(5):257–265, March 1984.

[75] Houtgast, T. and H. Steeneken. 'A Review of the MTF Concept in Room Acoustics and its Use For Estimating Speech Intelligibility in Auditoria'. *J. Acoust. Soc. Am.*, **77**(3):1069–1077, March 1985.

[76] Preis, Douglas, Vo Van Toi, and H. Sobie. 'The Complex Modulation Transfer Function of Duffing's Equation'. Paper 4075. AES 99th Convention, New York, October 1995.

[77] Steeneken, Herman, Sander van Wijngaarden, and Jan Verhave. 'The Evolution of the Speech Transmission Index'. Paper 8315. AES 130th Convention, London, May 2011.

[78] Rife, Douglas. 'Modulation Transfer Function Measurement with Maximum-Length Sequences'. *J. Audio Eng. Soc.*, **40**(10):779–790, October 1992.

[79] 'IEC 60268-16: Sound System Equipment. Part 16– The Objective Rating of Speech Intelligibility in Auditoria by the RASTI Method', 1988. Withdrawn.

[80] Keele, Don, Jr. 'Evaluation of Room Speech Transmission Index and Modulation Transfer Function by the Use of Time Delay Spectrometry'. Paper 6-006. AES 6th International Conference, Nashville, May 1988.

[81] Steeneken, Herman and Tammo Houtgast. 'Mutual Dependence of the Octave-Band Weights in Predicting Speech Intelligibility'. *Speech Communication*, **28**(2):109–123, June 1999.

[82] Leembruggen, Glenn, Marco Hippler, and Peter Mapp. 'Further Investigations Into Improving STI's Recognition of the Effects of Poor Frequency Response on Subjective Intelligibility'. Paper 8051. AES 128th Convention, London, May 2010.

[83] Houtgast, Tammo, Herman Steeneken, Wolfgang Ahnert, Loius Braida, Rob Drullman, Joost Festen, Kenneth Jacob, Peter Mapp, Steve McManus, Karen Payton, Reiner Plomp, Jan Verhave, and Sander van Wijngaarden. *Past, Present and Future of the Speech Transmission Index*. TNO Human Factors, Soesterberg, 2002.

[84] Mapp, Peter. 'Modifying STI to Better Reflect Subjective Impression'. Paper 000094. AES 21st International Conference, St. Petersburg, June 2002.

[85] Leembruggen, G. and A. Stacey. 'Should the Matrix be Reloaded?' *Proceedings of the Institute of Acoustics*, **25**(8):10–23, November 2003.

[86] Leembruggen, Glenn, Marco Hippler, and Peter Mapp. 'Exploring Ways to Improve STI's Recognition of the Effects of Poor Spectral Balance on Subjective Intelligibility'. *Proceedings of the Institute of Acoustics*, **31**(4):133–169, November 2009.

[87] Harris, Lara. *Can the Modulation Transfer Function be Used to Predict the Low Frequency Reproduction Quality of Musical Signals?* Master's thesis, ISVR, University of Southampton, MSc Dissertation, 2007.

[88] Stremler, F. G. *Introduction to Communication Systems (3^{rd} edition)*, Chapter 5: Amplitude Modulation. Addison-Wesley, 1990.

[89] Lathi, B. P. *Modern Digital and Analogue Communication Systems (3^{rd} edition)*, Chapter 4: Amplitude (Linear) Modulation. OUP USA, 1998.

[90] Oppenheim, Alan, Alan Willsky, and Hamid Nawad. *Signals & Systems (2^{nd} edition)*, Chapter 8: Communication Systems. Prentice Hall, 1997.

[91] Otung, Ifiok. *Communication Engineering Principles*, Chapter 3.2.3: Modulation Factor. Palgrave, 2001.

[92] Schroeder, M. 'New Method of Measuring Reverberation Time'. *J. Acoust. Soc. Am.*, **37**(6):1187–1188, June 1965.

[93] Holland, Keith, Philip Newell, Sergio Castro, and Bruno Fazenda. 'Excess Phase Effects and Modulation Transfer Function Degradation in Relation to Loudspeakers and Rooms Intended for the Quality Control Monitoring of Music'. *Proceedings of the Institute of Acoustics*, **27**(8):1–8, 2005.

[94] Linkwitz, Siegfried. 'Investigation of Sound Quality Differences between Monopolar and Dipolar Woofers in Small Rooms'. Paper 4786. AES 105th Convention, San Francisco, September 1998.

[95] Fazenda, Bruno, Keith Holland, and Phillip Newell. 'Modulation Transfer Function as a Measure of Room Low Frequency Performance'. *Proceedings of the Institute of Acoustics*, **28**(8):187–194, 2006.

[96] Moller, H. and C. Pedersen. 'Hearing at Low and Infrasonic Frequencies'. *Noise & Health*, **6**(23):37–57, April/June 2004.

[97] Glasberg, Brian and Brian Moore. 'Derivation of Auditory Filter Shapes from Notched-Noise Data'. *Hearing Research*, **47**(1/2):103–138, August 1990.

[98] Brookes, Mike. 'VOICEBOX (frq2erb.m).', Imperial College London. Last updated 2015. URL `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/frq2erb.html`.

[99] Thrane, N., J. Wismer, H. Konstantin-Hansen, and S. Gade. 'Practical Use of the Hilbert Transform'. Technical report, Bruel & Kjaer: BO 0173-11. Available online. URL `http://www.bksv.co.uk/doc/bo0437.pdf`.

[100] Kellaris, James and Robert Kent. 'Exploring Tempo and Modality Effects, On Consumer Responses to Music.' *Advances in Consumer Research*, **18**(1):243 – 248, January 1991.

[101] McKinney, Martin and Dirk Moelants. 'Ambiguity in Tempo Perception: What Draws Listeners to Different Metrical Levels?' *Music Perception: An Interdisciplinary Journal*, **24**(2):155–166, December 2006.

[102] McKinney, M., D. Moelants, M. Davies, and A. Klapuri. 'Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms'. *Journal of New Music Research*, **36**(1):1–16, January 2007.

[103] Muller, M., D. Ellis, A. Klapuri, and G. Richard. 'Signal Processing for Music Analysis'. *IEEE Journal of Selected Topics in Signal Processing*, **5**(6):1088–1110, October 2011.

[104] Benjamin, Eric and Benjamin Gannon. 'The Effect of Room Acoustics on Subwoofer Performance and Level Setting'. Paper 5232. AES 109th Convention, Los Angeles, September 2000.

[105] Bech, Soren. 'Requirements for Low-Frequency Sound Reproduction, Part I: The Audibility of Changes in Passband Amplitude Ripple and Lower System Cutoff Frequency and Slope'. *J. Audio Eng. Soc.*, **50**(7-8):564–580, July/August 2002.

[106] Choisel, Sylvain and Geoff Martin. 'Audibility of Phase Response Differences in a Stereo Playback System. Part 2: Narrow-Band Stimuli in Headphones and Loudspeakers'. Paper 7559. AES 125th Convention, San Francisco, October 2008.

[107] Hiekkanen, Timo, Aki Makivirta, and Matti Karjalainen. 'Virtualized Listening Tests for Loudspeakers'. *J. Audio Eng. Soc.*, **57**(4):237–250, April 2009.

[108] Olive, Sean, Peter Schuck, Sharon Sally, and Marc Bonneville. 'The Effects of Loudspeaker Placement on Listener Preference Ratings'. *J. Audio Eng. Soc.*, **42**(9):651–669, September 1994.

[109] Bech, Soren. 'The Influence of the Room and of Loudspeaker Position on the Timbre of Reproduced Sound in Domestic Rooms'. Paper 12-007. AES 12th International Conference, Copenhagen, June 1993.

[110] Toole, Floyd. *Sound Reproduction: Loudspeakers and Rooms*, Chapter 17.7: Creating a Listening Environment for Loudspeaker Evaluations. Focal press, 2013.

[111] Toole, Floyd. *Sound Reproduction: Loudspeakers and Rooms*, Chapter 17.5: Bias from Nonauditory Factors. Focal press, 2013.

[112] Toole, Floyd. 'Listening Tests- Turning Opinion Into Fact'. *J. Audio Eng. Soc.*, **30**(6):431–445, June 1982.

[113] Pedersen, J. and A. Makivirta. 'Requirements for Low-Frequency Sound Reproduction, Part II. Generation of Stimuli and Listening System Equalization'. *J. Audio Eng. Soc.*, **50**(7-8):581–593, July/August 2001.

[114] Nelson, P., H. Hamada, and S. Elliott. 'Adaptive Inverse Filters for Stereophonic Sound Reproduction'. *IEEE Transactions on Signal Processing*, **40**(7):1621–1632, July 1992.

[115] Nelson, P., F. Orduiia-Bustamante, and H. Hamada. 'Inverse Filter Design and Equalization Zones in Multichannel Sound Reproduction'. *IEEE Transactions on Signal Processing*, **3**(3):185–192, May 1995.

[116] Fielder, Louis. 'Analysis of Traditional and Reverberation-Reducing Methods of Room Equalization'. *J. Audio Eng. Soc.*, **51**(1/2):3–26, January/February 2003.

[117] Kirkeby, O., F. Orduna, P. Nelson, and H. Hameda. 'Inverse Filtering in Sound Reproduction'. *Measurement and Control*, **26**(9):261–266, November 1993.

[118] Schmitt, Regina. 'Audibility of Nonlinear Loudspeaker Distortions'. Paper 4016. AES 98th Convention, Paris, February 1995.

[119] BBC Research Department. 'Loudspeaker Distortion Associated with Low-Frequency Signals (report 1972/25)'. Technical report, 1972.

[120] Karlsson, Robert. *Loudspeaker-Room Equalisation.* Master's thesis, ISVR, University of Southampton, 2011.

[121] Norcross, S., G. Soulodre, and M. Lavoie. 'Subjective Investigations of Inverse Filtering'. *J. Audio Eng. Soc.*, **52**(10):1003–1028, October 2004.

[122] Small, Richard. 'Vented-Box Loudspeaker Systems Part II: Large-Signal Analysis'. *J. Audio Eng. Soc.*, **21**(6):438–444, August 1973.

[123] Small, Richard. 'Vented-Box Loudspeaker Systems Part III: Synthesis'. *J. Audio Eng. Soc.*, **21**(7):549–554, September 1973.

[124] Colloms, Martin. *High Performance Loudspeakers (5^{th} edition)*, Chapter 4: Low Frequency System Analysis. Wiley, 1997.

[125] Geddes, Earl. 'On Sound Radiation From Ported Enclosures'. *J. Audio Eng. Soc.*, **49**(3):117–124, March 2001.

[126] Holland, Keith. 'HHb Circle 3P: Review'. *Studio Sound*, December 2000.

[127] Holland, Keith. 'Hafler TRM8: Review'. *Studio Sound*, May 1998.

[128] JBL. 'LSR32 Linear Spatial Reference Studio Monitor System Datasheet. Version: 3/98.', accessed: 29/02/2016. URL `https://www.jblpro.com/pub/recording/lsr32.pdf`.

[129] Holland, Keith. 'JBL LSR32: Bench Test'. *Studio Sound*, October 1998.

[130] 'Large anechoic chamber; ISVR Consulting website.', Accessed 08/06/2015. URL `http://www.isvr.co.uk/faciliti/lg_anech.htm`.

[131] Herlufsen, H. 'Dual Channel FFT Analysis (Part I)'. *Bruel & Kjaer Technical Review*, No.1 1984.

[132] Randall, R. B. *Frequency Analysis (3^{rd} edition)*, Chapter 7: Dual Channel Analysis. Bruel & Kjaer, 1987.

[133] Kinsler, Lawrence, Austin Frey, Alan Coppens, and James Sanders. *Fundamentals Of Acoustics (4^{th} edition)*, Chapter 1.2: Noise, Spectrum Level, and Band Level. Wiley, 2000.

[134] Greenfield, Richard and Malcom Hawksford. 'The Audibility of Loudspeaker Phase Distortion'. Paper 2927. AES 88th Convention, Montreux, March 1990.

[135] Mackie. 'MR Mk3 Series: Powered Studio Monitors', Accessed 09/06/2015. URL `http://www.mackie.com/products/mrmk3-studio-monitors/specs`.

[136] Genelec. '8320A Bi-Amplified Smart Active Monitor: Specifications', Accessed 09/06/2015. URL `http://www.genelec.com/products/8320a/`.

[137] Yamaha. 'HS7 Powered Studio Monitor (specs)', Accessed 09/06/2015. URL `http://uk.yamaha.com/en/products/music-production/speakers/hs_series/hs7_white`.

[138] 'BS EN ISO 5492: Sensory Analysis - Vocabulary', 2009.

[139] Toole, Floyd. 'Subjective Measurements of Loudspeaker Sound Quality and Listener Performance'. *J. Audio Eng. Soc*, **33**(1-2):2–32, February 1985.

[140] Kemp, Sarah, Tracey Hollowood, and Joanne Hort. *Sensory Evaluation: A Practical Handbook*, Chapter 5: Sensory Test Methods. Wiley-Blackwell, 2009.

[141] Munson, W. and M. Gardner. 'Loudness Patterns - A New Approach'. *J. Acoust. Soc. Am.*, **22**(2):177–190, March 1950.

[142] Clark, David. 'High-Resolution Subjective Testing Using a Double-Blind Comparator'. *J. Audio Eng. Soc*, **30**(5):330–338, May 1982.

[143] 'BS 6840-13: Sound System Equipment Part 13– Listening Tests on Loudspeakers', 1998.

[144] 'ITU-R BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems', 2015.

[145] 'BS EN ISO 4120: Sensory Analysis - Methodology - Triangle Test', 2007.

[146] 'BS EN ISO 10399: Sensory Analysis - Methodology - Duo-Trio Test', 2010.

[147] 'BS EN ISO 5495: Sensory Analysis - Methodology - Paired Comparison Test', 2007.

[148] 'BS ISO 6658: Sensory Analysis - Methodology - General Guidance', 2005.

[149] Punch, Jerry and Cheryl Parker. 'Pairwise Listener Preferences in Hearing-Aid Evaluation'. *Journal Of Speech And Hearing Research*, **24**(3):366–374, September 1981.

[150] Ulrich, R. and J. Miller. 'Threshold Estimation in Two-Alternative Forced-Choice (2AFC) Tasks: The Spearman-Karber Method'. *Perception & Psychophysics*, **66**(3):517–533, April 2004.

[151] Jason, M. 'Design Considerations for Loudspeaker Preference Experiments'. *J. Audio Eng. Soc*, **40**(12):979–996, December 1992.

[152] Zielinski, S., F. Rumsey, and S. Bech. 'On Some Biases Encountered in Modern Audio Quality Listening Tests - A Review'. *J. Audio Eng. Soc*, **56**(6):427–451, June 2008.

[153] Gabrielsson, A. and B. Lindstrom. 'Perceived Sound Quality of High-Fidelity Loudspeakers'. *J. Audio Eng. Soc*, **33**(1-2):33–53, February 1985.

[154] 'BS ISO 4121: Sensory Analysis - Guidelines for the Use of Quantitative Response Scales', 2003.

[155] David, Herbert A. *The Method of Paired Comparisons*, Chapter 1: Probability Models. Hodder Arnold, London, 1988.

[156] Rumsey, F., S. Zielinski, R. Kassier, and S. Bech. 'On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality'. *J. Acoust. Soc. Am.*, **118**(2):968–976, August 2005.

[157] Kendall, Maurice. *Rank Correlation Methods (4^{th} edition)*, Chapter 11: Paired Comparions. Griffin, London, 1970.

[158] 'ITU-R BS.1534-2: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems', 2014.

[159] Zimmer, K. and W. Ellermeier. 'Deriving Ratio-Scale Measures of Sound Quality from Preference Judgments'. *Noise Control Engineering Journal*, **51**(4):210–215, Jul/Aug 2003.

[160] Zimmer, K., W. Ellermeier, and C. Schmid. 'Using Probabilistic Choice Models to Investigate Auditory Unpleasantness'. *Acta Acustica United with Acustica*, **90**(6):1019–1028, November 2004.

[161] Kemp, Sarah, Tracey Hollowood, and Joanne Hort. *Sensory Evaluation: A Practical Handbook*, Chapter 2: Sensory Evaluation. Wiley-Blackwell, 2009.

[162] Olive, S. 'Differences in Performance and Preference of Trained Versus Untrained Listeners in Loudspeaker Tests: A Case Study'. *J. Audio Eng. Soc.*, **51**(9):806–825, September 2003.

[163] Kemp, Sarah, Tracey Hollowood, and Joanne Hort. *Sensory Evaluation: A Practical Handbook*, Chapter 4: Requirements For Sensory Testing. Wiley-Blackwell, 2009.

[164] Hautus, M. and X. Meng. 'Decision Strategies in the ABX (Matching-To-Sample) Psychophysical Task'. *Perception & Psychophysics*, **64**(1):89–106, January 2002.

[165] Gabrielsson, A., B. Hagerman, T. Beck-Kristensen, and G. Lundberg. 'Perceived Sound Quality of Reproductions With Different Frequency Responses and Sound Levels'. *J. Acoust. Soc. Am.*, **88**(3):1359–1366, September 1990.

[166] Kendall, M. 'Further Contributions to the Theory of Paired Comparisons'. *Biometrics*, **11**(1):43–62, March 1955.

[167] Kemp, Sarah, Tracey Hollowood, and Joanne Hort. *Sensory Evaluation: A Practical Handbook*, Chapter 3: Planning Your Sensory Project. Wiley-Blackwell, 2009.

[168] Ryden, Thomas. 'Using Listening Tests to Assess Audio Codecs'. *Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 115–125, May 1996.

[169] 'ITU-R BS.1116-1: Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems', 1994.

[170] Montgomery, Douglas. *Design and Analysis of Experiments (6^{th} edition)*, Chapter 4: Randomized Blocks, Latin Squares, and Related Designs. Wiley, Hoboken, N.J., 2005.

[171] Gabrielsson, A., B. Lindstrom, and O. Till. 'Loudspeaker Frequency Response and Perceived Sound Quality'. *J. Acoust. Soc. Am.*, **90**(2):707–719, August 1991.

[172] Chapman, Peter. 'Programme Material Analysis'. Paper 4277. AES 100th Convention, Copenhagen, May 1996.

[173] Francombe, Jon, Russell Mason, Martin Dewhirst, and Soren Bech. 'Investigation of a Random Radio Sampling Method for Selecting Ecologically Valid Music Program Material'. Paper 9029. AES 136th Convention, Berlin, April 2014.

[174] Neuendorf, Max and Frederik Nagel. 'Exploratory Studies on Perceptual Stationarity in Listening Test - Part I: Real World Signals from Custom Listening Tests'. Paper 8562. AES 131st Convention, New York, October 2011.

[175] Nagel, Frederik and Max Neuendorf. 'Exploratory Studies on Perceptual Stationarity in Listening Test - Part II: Synthetic Signals with Time Varying Artifacts'. Paper 8563. AES 131st Convention, New York, October 2011.

[176] Deruty, Emmanuel and Damien Tardieu. 'About Dynamic Processing in Mainstream Music'. *J. Audio Eng. Soc*, **62**(1/2):42–55, January 2014.

[177] Hjortkjær, Jens and Mads Walther-Hansen. 'Perceptual Effects of Dynamic Range Compression in Popular Music Recordings'. *J. Audio Eng. Soc*, **62**(1/2):37–41, February 2014.

[178] Human Experimentation Safety and Ethics Committee. 'Guide to Experimentation Involving Human Subjects; ISVR Technical Memorandum No. 808', October 1996.

[179] 'BS EN 61672-1: Electroacoustics. Sound level meters. Specifications', 2013.

[180] Conover, W. J. *Practical Nonparametric Statistics*, Chapter 2.5: Nonparametric Statistics. Wiley, New York, 1971.

[181] Field, Andy. *Discovering Statistics Using SPSS (2^{nd} edition)*, Chapter , pp. 132–133. SAGE Publications, London, 2005.

[182] Raffin, M. and D. Schafer. 'Application of a Probability Model Based on the Binomial-Distribution to Speech-Discrimination Scores'. *Journal Of Speech And Hearing Research*, **23**(3):570–575, September 1980.

[183] Bech, Soren. 'Listening Tests on Loudspeakers: A Discussion of Experimental Procedures and Evaluation of the Response Data'. Paper 8-014. AES 8th International Conference, Washington D.C., May 1990.

[184] Lantz, Bjorn. 'The Large Sample Size Fallacy'. *Scandinavian Journal of Caring Sciences*, **27**(2):487–492, June 2013.

[185] Argyrous, George. *Statistics for Research; With a Guide to SPSS (2^{nd} edition)*, Chapter 23: The Chi-square Test for Independence. SAGE, 2005.

[186] Snedcor, George and William Cochran. *Statistical Methods (6^{th} edition)*, Chapter 9.13: The RxC table, pp. 250–251. Iowa State University Press, Ames, 1967.

[187] Srednicki, M. 'A Bayesian Analysis of A-B Listening Tests'. *J. Audio Eng. Soc*, **36**(3):143–146, March 1988.

[188] Hsu, Poh Ser and Tar Su Hsu. 'Statistical Analysis of Double-Blind Tests for Multiple Audiences'. Paper 2515. AES 83rd Convention, New York, October 1987.

[189] Thurstone, L. 'A Law Of Comparative Judgment'. *Psychological Review*, **101**(2):266–270, April 1994. (Reprinted From Psychological Review, Vol. 34, 1927).

[190] Bradley, Ralph and Milton Terry. 'Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons'. *Biometrika*, **39**(3/4):324–345, 1952.

[191] Guilford, J. P. *Psychometric Methods (2^{nd} edition)*, Chapter 7: The Method of Pair Comparisons. McGraw Hill, New York, 1954.

[192] Staffeldt, Henrik. 'Correlation Between Subjective and Objective Data for Quality Loudspeakers'. *J. Audio Eng. Soc*, **22**(6):402–415, March 1974.

[193] Bech, Soren and Nick Zacharov. *Perceptual Audio Evaluation: Theory, Method and Application*, Chapter 4.2.2: Indirect Scaling. John Wiley, 2006.

[194] Leventhal, Les. 'Type 1 and Type 2 Errors in the Statistical Analysis of Listening Tests'. *J. Audio Eng. Soc*, **34**(6):437–453, June 1986.

[195] Leventhal, L. and C. Huynh. 'Analyzing Listening Tests with the Directional Two-Tailed Test'. *J. Audio Eng. Soc*, **44**(10):850–863, October 1996.

[196] Burstein, Herman. 'Approximation Formulas for Error Risk and Sample Size in ABX Testing'. *J. Audio Eng. Soc*, **36**(11):879–883, November 1988.

[197] Conover, W. J. *Practical Nonparametric Statistics*, Chapter 3: Some Tests Based on the Binomial Distribution. Wiley, 1971.

[198] Conover, W. J. *Practical Nonparametric Statistics*, Chapter 5: The Use of Ranks. Wiley, 1971.

[199] Argyrous, George. *Statistics For Research: With a Guide to SPSS (2^{nd} edition)*, Chapter 21: One Sample Tests for a Binomial Distribution. SAGE, 2005.

[200] Field, Andy. *Discovering Statistics Using SPSS (2^{nd} edition)*, Chapter , pp. 372–373. SAGE, 2005.

[201] Field, Andy. *Discovering Statistics Using SPSS (2^{nd} edition)*, Chapter , p. 402. SAGE, 2005.

[202] Argyrous, George. *Statistics For Research: With a Guide to SPSS (2^{nd} edition)*, Chapter , pp. 227–229. SAGE, 2005.

[203] Leventhal, Les. 'Statistically Significant Poor Performance in Listening Tests'. *J. Audio Eng. Soc.*, **42**(7/8):585–587, July/August 1994.

[204] Kendall, M. and B. Babington-Smith. 'On the Method of Paired Comparisons'. *Biometrika*, **31**(3-4):324–345, March 1940.

[205] David, Herbert A. *The Method of Paired Comparisons*, Chapter 2.1: Scores and Circular Triads. Hodder Arnold, London, 1988.

[206] 'BS ISO 226: Acoustics. Normal Equal-Loudness-Level Contours', 2003.

[207] 'BS EN ISO 389-7: Acoustics. Reference Zero for the Calibration of Audiometric Equipment. Part 7: Reference Threshold of Hearing Under Free-field and Diffuse-field Listening Conditions', 2005.

[208] Heyser, Richard. 'The Delay Plane, Objective Analysis of Subjective Properties: Part 1'. *J. Audio Eng. Soc*, **21**(9):690–701, November 1973.

[209] Deer, J., P. Bloom, and D. Preis. 'Perception of Phase Distortion in All-Pass Filters'. *J. Audio Eng. Soc.*, **33**(10):782–786, October 1985.

[210] Bunton, John and Richard Small. 'Cumulative Spectra, Tone Bursts, and Apodization'. *J. Audio Eng. Soc*, **30**(6):386–395, June 1982.

[211] Toole, Floyd and Sean Olive. 'The Modification of Timbre by Resonances: Perception and Measurement'. *J. Audio Eng. Soc*, **36**(3):122–142, March 1988.

[212] Preis, D., F. Hlawatsch, P. Bloom, and J. Deer. 'Wigner Distribution Analysis of Filters with Perceptible Phase Distortion'. *J. Audio Eng. Soc*, **35**(12):1004–1012, December 1987.

[213] Preis, Douglas. 'A Catalog of Frequency and Transient Responses'. *J. Audio Eng. Soc*, **25**(12):990–1007, December 1977.

[214] Shorter, D. 'Recent Investigations Into Methods of Measuring the Transient Response of Loudspeakers'. Technical Report M.004, BBC Research Department, 1944.

[215] Murray, Fancher. 'MTF as a Tool in Transducer Selection'. Paper 6-031. AES 6th International Conference, Nashville, May 1988.

[216] Gerzon, Michael. 'Comments on The Delay Plane, Objective Analysis of Subjective Properties: Part 1'. *J. Audio Eng. Soc*, **22**(2):104–106, March 1974.

[217] Janse, Cornelis and Arie Kaizer. 'Time-Frequency Distributions of Loudspeakers: The Application of the Wigner Distribution'. *J. Audio Eng. Soc*, **31**(4):198–223, April 1983.

[218] Preis, Douglas and Voula Georgopoulos. 'Wigner Distribution Representation and Analysis of Audio Signals: An Illustrated Tutorial Review'. *J. Audio Eng. Soc*, **47**(12):1043–1053, December 1999.

[219] Brunet, Pascal, Zachary Rimkunas, and Steve Temme. 'Evaluation of Time-Frequency Analysis Methods and Their Practical Applications'. Paper 7203. AES 123rd Convention, New York, October 2007.

[220] Stephenson, Matthew. *Assessing the Quality of Low Frequency Audio Reproduction in Critical Listening Spaces*. Ph.D. thesis, School of Computing, Science, & Engineering; University of Salford, 2012.

[221] Wilson, Alex and Bruno Fazenda. 'Perception of Audio Quality in Productions of Popular Music'. *J. Audio Eng. Soc*, **64**(1/2):23–34, January 2016.

[222] Craig, James and Lloyd Jeffress. 'Effect of Phase on the Quality of a Two-Component Tone'. *J. Acoust. Soc. Am.*, **34**(11):1752–1760, November 1962.

[223] Stodolsky, D. 'The Standardization of Monaural Phase'. *IEEE Transactions on Audio and Electroacoustics*, **18**(3):288–299, September 1970.

[224] Patterson, James and David Green. 'Discrimination of Transient Signals Having Identical Energy Spectra'. *J. Acoust. Soc. Am.*, **48**(4B):894–905, October 1970.

[225] Green, David. 'Temporal Acuity as a Function of Frequency'. *J. Acoust. Soc. Am.*, **54**(2):373–379, August 1973.

[226] Bilsen, F. 'On the Influence of the Number and Phase of Harmonics on the Perceptibility of the Pitch of Complex Signals'. *Acustica*, **28**(1):60–65, January 1973.

[227] Hansen, Villy and Erik Madsen. 'On Aural Phase Detection: Part 1'. *J. Audio Eng. Soc*, **22**(1):10–14, February 1974.

[228] Hansen, Villy and Erik Madsen. 'On Aural Phase Detection: Part 2'. *J. Audio Eng. Soc*, **22**(10):783–788, December 1974.

[229] Mosteller, Frederick. 'Remarks on the Method of Paired Comparisons: III. A Test of Significance for Paired Comparisons When Equal Standard Deviations and Equal Correlations are Assumed'. *Psychometrika*, **16**(2):207–218, June 1951.

[230] Fazenda, Bruno and W. Davies. 'The Views of Recording Studio Control Room Users'. *Proceedings of the Institute of Acoustics*, **23**(23):1–8, November 2002.

[231] Holland, K. R. 'Ported Loudspeaker Model'. Technical report, ISVR course handout, 2007.