Predicting the Perceptual Acceptability of Auditory Interference for the Optimisation of Sound Zones

Khan Richard Baykaner

Submitted for the Degree of Doctor of Philosophy

School of Arts
Faculty of Arts & Human Sciences
University of Surrey

August 2014
© Khan Richard Baykaner 2014
Abstract

This work is part of the Perceptually Optimised Sound Zone project (posz.org) which aims to develop sound zoning systems which reproduce audio programmes to multiple listening zones within automotive and domestic environments. This work describes the construction of a model to evaluate sound zoning systems.

A framework for evaluating auditory interference scenarios is described in which either the target or interferer programme is masked, or where both programmes are audible and the listening scenario has some degree of acceptability. Masking and acceptability experiments were conducted to investigate the relationship between the two, and to determine boundaries of audibility. A linear correlation was found between masking and acceptability, and a linear regression model was constructed to predict thresholds of acceptability from masking thresholds. A masking threshold model was adapted and predictions were within 3 dB of the reported mean masking thresholds. Predictions of acceptability, using a linear regression and masking model combination, accounted for three quarters of the variance in acceptability.

Further work focused on speech target programmes based on listener comments that the presence of speech affected acceptability. An experiment was conducted to gather intelligibility and acceptability data. Results showed that a high speech intelligibility marked the lower boundary of acceptability. Existing models for intelligibility prediction were evaluated and a time-windowed speech intelligibility index was shown to predict intelligibility with RMSE = 10.8%.

Subsequently, a model was constructed to predict acceptability within these boundaries. Two experiments were conducted gathering training and validation data, and a training and selection procedure was carried out to methodically identify the most useful features. The selected model predictions had acceptability scores of RMSE = 11.1–17.9% across training and validation data.

Finally, an algorithm was proposed for the prediction of acceptability in auditory interference scenarios. The algorithm consists of first predicting masking thresholds to determine the boundaries of acceptability. Then, for non-speech target programmes, the acceptability is predicted using a linear regression to the masking threshold; for speech target programmes, the intelligibility is calculated to revise the lower acceptability boundary and the speech acceptability model is used to predict acceptability.
Statement of Originality

This thesis and the work to which it refers are the results of my own efforts. Any ideas, data, images or text resulting from the work of others (whether published or unpublished) are fully identified as such within the work and attributed to their originator in the text, bibliography or in footnotes. This thesis has not been submitted in whole or in part for any other academic degree or professional qualification. I agree that the University has the right to submit my work to the plagiarism detection service TurnitinUK for originality checks. Whether or not drafts have been so-assessed, the University reserves the right to require an electronic version of the final document (as submitted) for assessment as above.
# Contents

Abstract iii  
Statement of Originality iv  
Contents v  
List of Figures x  
List of Tables xiv  
List of Equations xvi  
Publications Arising from this Thesis xviii  
Acknowledgements xix  

## Chapter 1: Introduction  
1.1 The aim of this work .............................................. 1  
1.2 Auditory masking in sound zones .............................. 4  
1.3 Speech perception .................................................. 5  
1.4 Research questions ................................................ 5  
1.5 The structure of this thesis ..................................... 7  

## Chapter 2: Auditory Masking  
2.1 The measurement of auditory masking .......................... 8  
2.2 Auditory filters and masking experiments .................... 9  
2.2.1 Auditory filters and the power spectrum model .......... 9  
2.2.2 Off-frequency listening ....................................... 13  
2.2.3 The roex auditory filter ...................................... 16  
2.2.4 The gammatone filter ......................................... 17  
2.2.5 The dual resonance non-linear filter ..................... 18  
2.2.6 Excitation levels, excitation patterns, and the upward spread of masking ........................................... 19  
2.2.7 Extracting partials from complex tones .................. 22  
2.2.8 Two-tone masking ............................................. 24  
2.2.9 Discussion ..................................................... 24  
2.3 Temporal masking .................................................. 28  
2.3.1 Simultaneous masking ........................................ 28  
2.3.2 Forward masking .............................................. 28  
2.3.3 Backward masking & excess masking ....................... 30
# Contents

2.4 Partial masking .................................................. 30
    2.4.1 Evidence for partial masking .......................... 31
    2.4.2 Spectral partial masking ............................ 32
    2.4.3 Temporal partial masking ........................... 33
    2.4.4 Discussion ............................................ 34

2.5 Binaural unmasking .......................................... 34
    2.5.1 Impact of binaural unmasking ....................... 37
    2.5.2 Summary .............................................. 37

2.6 Informational masking ..................................... 38
    2.6.1 Uncertainty masking ................................. 39
    2.6.2 Summary .............................................. 44

2.7 Comodulation masking release ............................. 44

2.8 Summary and conclusions ................................... 47

Chapter 3: Masking and Acceptability Experiment ......... 49
    3.1 Experiment design ....................................... 49
        3.1.1 Methodology ....................................... 49
        3.1.2 Stimuli ............................................ 51

3.2 Results ......................................................... 53
    3.2.1 Masking experiment .................................. 53
    3.2.2 Acceptability experiment ............................ 55
    3.2.3 Subject comments .................................... 67
    3.2.4 Discussion ............................................ 67

3.3 Comparison of masking and acceptability ............... 68
    3.3.1 ANOVA of difference scores ......................... 69
    3.3.2 Correlations .......................................... 71

3.4 Summary and conclusion ..................................... 73

Chapter 4: Masking Prediction ................................. 75
    4.1 Masking threshold prediction models ................... 75
        4.1.1 Fletcher's power spectrum model .................. 75
        4.1.2 Zwicker's loudness model ........................ 76
        4.1.3 Short-time partial loudness model ............... 78
        4.1.4 Computational auditory signal-processing and perception model 80
        4.1.5 Component of relative entropy model ............. 86
        4.1.6 A note on modelling binaural unmasking .......... 88
        4.1.7 Selection ........................................... 88

4.2 Implementing the masking model .......................... 90
    4.2.1 Some complications with computational masking models 90
    4.2.2 Modifications to CASP ............................... 91

4.3 Evaluation of masking threshold predictions ............ 95

4.4 Predicting acceptability by predicting masking ....... 97

4.5 Summary and Conclusions ................................... 99

Chapter 5: Speech Intelligibility Experiment .............. 101
    5.1 Speech intelligibility background ...................... 102
        5.1.1 The nature of speech as a percept ................ 102
        5.1.2 The use of speech intelligibility in auditory interference 102
        5.1.3 Corpus selection ................................... 103
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1.4</td>
<td>Calculation of intelligibility scores</td>
<td>107</td>
</tr>
<tr>
<td>5.1.5</td>
<td>Response method</td>
<td>109</td>
</tr>
<tr>
<td>5.1.6</td>
<td>Application to sound zoning systems</td>
<td>109</td>
</tr>
<tr>
<td>5.2</td>
<td>Pilot experiment</td>
<td>110</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Presentation levels</td>
<td>111</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Stimuli</td>
<td>112</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Methodology</td>
<td>113</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Calculation of intelligibility</td>
<td>116</td>
</tr>
<tr>
<td>5.3</td>
<td>Pilot experiment results</td>
<td>117</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Scoring transcripts and contextual errors</td>
<td>118</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Data transformation and the validity of ANOVAs</td>
<td>118</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Intelligibility</td>
<td>119</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Distraction scores</td>
<td>125</td>
</tr>
<tr>
<td>5.3.5</td>
<td>Acceptability and masking</td>
<td>127</td>
</tr>
<tr>
<td>5.3.6</td>
<td>Summary and consequences for main experiment</td>
<td>130</td>
</tr>
<tr>
<td>5.4</td>
<td>Main experiment</td>
<td>132</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Presentation levels and SNRs</td>
<td>132</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Target corpus selection</td>
<td>133</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Interferer programmes</td>
<td>134</td>
</tr>
<tr>
<td>5.4.4</td>
<td>Methodology</td>
<td>135</td>
</tr>
<tr>
<td>5.4.5</td>
<td>Assumptions and limitations</td>
<td>137</td>
</tr>
<tr>
<td>5.5</td>
<td>Main experiment results</td>
<td>138</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Intelligibility</td>
<td>138</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Distraction scores</td>
<td>152</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Acceptability and masking data</td>
<td>162</td>
</tr>
<tr>
<td>5.5.4</td>
<td>Interferer intelligibility</td>
<td>169</td>
</tr>
<tr>
<td>5.5.5</td>
<td>Overview of results</td>
<td>171</td>
</tr>
<tr>
<td>5.6</td>
<td>Summary and conclusions</td>
<td>174</td>
</tr>
</tbody>
</table>

**Chapter 6: Speech Intelligibility Prediction**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Intelligibility models</td>
<td>175</td>
</tr>
<tr>
<td>6.1.1</td>
<td>The speech intelligibility index</td>
<td>176</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Short-time objective intelligibility</td>
<td>178</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Time frequency multi-look model</td>
<td>179</td>
</tr>
<tr>
<td>6.1.4</td>
<td>Coherence speech intelligibility index</td>
<td>181</td>
</tr>
<tr>
<td>6.2</td>
<td>Model evaluations</td>
<td>182</td>
</tr>
<tr>
<td>6.2.1</td>
<td>SII predictions</td>
<td>182</td>
</tr>
<tr>
<td>6.2.2</td>
<td>STOI</td>
<td>185</td>
</tr>
<tr>
<td>6.2.3</td>
<td>TFMLM</td>
<td>186</td>
</tr>
<tr>
<td>6.2.4</td>
<td>CSII</td>
<td>187</td>
</tr>
<tr>
<td>6.3</td>
<td>Model comparison</td>
<td>188</td>
</tr>
<tr>
<td>6.4</td>
<td>Summary and conclusion</td>
<td>189</td>
</tr>
</tbody>
</table>

**Chapter 7: Acceptability Training and Validation Experiments**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Training experiment</td>
<td>190</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Methodology</td>
<td>191</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Stimuli</td>
<td>194</td>
</tr>
<tr>
<td>7.1.3</td>
<td>Subjects</td>
<td>195</td>
</tr>
<tr>
<td>7.1.4</td>
<td>General results and analysis</td>
<td>196</td>
</tr>
</tbody>
</table>
7.2 Validation experiment ........................................... 201
  7.2.1 Differences in experiment design ........................... 201
  7.2.2 Stimuli .................................................... 203
  7.2.3 General results and analysis ............................... 204
  7.3 Summary and conclusion ..................................... 205

Chapter 8: Building a Model to Predict Acceptability 206
  8.1 Modelling approach ......................................... 206
    8.1.1 Model complexity ...................................... 206
    8.1.2 Data sets for training and testing models ............... 208
    8.1.3 Model metrics ........................................ 208
  8.2 A model using CASP based features ........................ 211
    8.2.1 Features ................................................ 211
    8.2.2 Model building approach ................................ 215
    8.2.3 A benchmark model ..................................... 219
    8.2.4 Constructing the model .................................. 222
    8.2.5 Validation .............................................. 225
    8.2.6 Summary ................................................ 226
  8.3 A search for further features ................................ 227
    8.3.1 Features ................................................ 227
    8.3.2 Constructing the model .................................. 229
    8.3.3 Validation .............................................. 232
    8.3.4 Summary ................................................ 233
  8.4 Including PEASS ............................................. 234
    8.4.1 Constructing the model .................................. 235
    8.4.2 Validation .............................................. 238
    8.4.3 Summary ................................................ 238
  8.5 SNR based hierarchy and model adjustments ............... 240
  8.6 Model selection .............................................. 242
  8.7 Model comparison ........................................... 242
    8.7.1 PEASS comparison ...................................... 244
    8.7.2 PESQ and POLQA comparison ............................. 244
  8.8 Summary and conclusion ..................................... 246

Chapter 9: Predicting Acceptability 248
  9.1 General method of acceptability prediction ................. 248
  9.2 Assumptions and limitations ................................ 251
  9.3 Practical implementation of the acceptability prediction algorithm .... 252
  9.4 Research applications .................................... 253
    9.4.1 Evaluating sound-zoning systems ...................... 254
    9.4.2 Designing and evaluating zone-based public address systems 255
    9.4.3 Evaluating and designing acoustic spaces ............ 256
  9.5 Summary and conclusions .................................. 257

Chapter 10: Conclusion 258
  10.1 Framing the problem ....................................... 260
  10.2 Auditory masking in auditory interference scenarios, and the use of masking prediction for the prediction of acceptability ... 261
  10.3 Speech intelligibility and acceptability in auditory interference scenarios 262
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.4 A model for predicting the acceptability of speech in auditory</td>
<td>263</td>
</tr>
<tr>
<td>interference scenarios</td>
<td></td>
</tr>
<tr>
<td>10.5 A method for predicting acceptability</td>
<td>263</td>
</tr>
<tr>
<td>10.6 Scope, limitations, and further work</td>
<td>263</td>
</tr>
<tr>
<td>10.7 Novel contributions to the field</td>
<td>264</td>
</tr>
<tr>
<td><strong>Acceptability model features</strong></td>
<td>267</td>
</tr>
<tr>
<td><strong>Spectrograms of stimuli</strong></td>
<td>273</td>
</tr>
<tr>
<td><strong>Acronyms</strong></td>
<td>275</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>278</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A room with multiple sound zones and multiple sound sources</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>Overall quality of the POSZ system</td>
<td>2</td>
</tr>
<tr>
<td>2.1</td>
<td>A simple band widening experiment</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Band widening experiment results</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Loudness of noise bands</td>
<td>11</td>
</tr>
<tr>
<td>2.4</td>
<td>Off-frequency listening</td>
<td>13</td>
</tr>
<tr>
<td>2.5</td>
<td>Off-frequency listening and the non-linear addition of masking</td>
<td>14</td>
</tr>
<tr>
<td>2.6</td>
<td>The notched noise method</td>
<td>15</td>
</tr>
<tr>
<td>2.7</td>
<td>An ideal roex auditory filter</td>
<td>16</td>
</tr>
<tr>
<td>2.8</td>
<td>An example gammatone filterbank</td>
<td>18</td>
</tr>
<tr>
<td>2.9</td>
<td>Overview of the dual resonance non-linear filter</td>
<td>19</td>
</tr>
<tr>
<td>2.10</td>
<td>Overlapping tails of auditory filters</td>
<td>20</td>
</tr>
<tr>
<td>2.11</td>
<td>An example excitation pattern</td>
<td>21</td>
</tr>
<tr>
<td>2.12</td>
<td>The upward spread of masking</td>
<td>23</td>
</tr>
<tr>
<td>2.13</td>
<td>Determining the critical bandwidth</td>
<td>25</td>
</tr>
<tr>
<td>2.14</td>
<td>Spectral analysis of complex signals</td>
<td>27</td>
</tr>
<tr>
<td>2.15</td>
<td>Forward masking thresholds</td>
<td>29</td>
</tr>
<tr>
<td>2.16</td>
<td>Partial masking loudness</td>
<td>31</td>
</tr>
<tr>
<td>2.17</td>
<td>Spectral partial masking loudness</td>
<td>32</td>
</tr>
<tr>
<td>2.18</td>
<td>Temporal partial masking</td>
<td>33</td>
</tr>
<tr>
<td>2.19</td>
<td>Binaural unmasking</td>
<td>35</td>
</tr>
<tr>
<td>2.20</td>
<td>Phase characteristics of sound zoning methods</td>
<td>36</td>
</tr>
<tr>
<td>2.21</td>
<td>Effect of phase on binaural unmasking</td>
<td>36</td>
</tr>
<tr>
<td>2.22</td>
<td>BMLD of speech</td>
<td>37</td>
</tr>
<tr>
<td>2.23</td>
<td>Spectral uncertainty masking</td>
<td>40</td>
</tr>
<tr>
<td>2.24</td>
<td>Learning curves for spectral uncertainty masking</td>
<td>42</td>
</tr>
<tr>
<td>2.25</td>
<td>Learning of tonal complexes of different levels of uncertainty</td>
<td>43</td>
</tr>
<tr>
<td>2.26</td>
<td>Comodulation masking release</td>
<td>45</td>
</tr>
<tr>
<td>2.27</td>
<td>A two noise band masking scenario</td>
<td>46</td>
</tr>
<tr>
<td>3.1</td>
<td>Masking experiment test setup</td>
<td>50</td>
</tr>
<tr>
<td>3.2</td>
<td>Mean masking thresholds separated by target and interferer</td>
<td>54</td>
</tr>
<tr>
<td>3.3</td>
<td>Subject 1 acceptability scores compared with all other scores</td>
<td>57</td>
</tr>
<tr>
<td>3.4</td>
<td>Subject 7 acceptability scores compared with all other scores</td>
<td>57</td>
</tr>
<tr>
<td>3.5</td>
<td>Relationship between mean acceptability scores and difference between</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>repeated trials</td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td>A principal component analysis of acceptability scores separated by subject</td>
<td>59</td>
</tr>
<tr>
<td>3.7</td>
<td>Mean acceptability thresholds separated by target, interferer, and filtering</td>
<td>64</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>3.8</td>
<td>Mean acceptability thresholds separated by target and interferer</td>
<td>65</td>
</tr>
<tr>
<td>3.9</td>
<td>Acceptability target to interferer ratio for all listening scenarios split by subject</td>
<td>66</td>
</tr>
<tr>
<td>3.10</td>
<td>A scatter plot of acceptability and masking data</td>
<td>72</td>
</tr>
<tr>
<td>3.11</td>
<td>A scatter plot of averaged masking and acceptability data</td>
<td>73</td>
</tr>
<tr>
<td>4.1</td>
<td>A simple model of simultaneous masking</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Critical band rate</td>
<td>77</td>
</tr>
<tr>
<td>4.3</td>
<td>An overview of the STPL model</td>
<td>79</td>
</tr>
<tr>
<td>4.4</td>
<td>Forward masking data predictions using a temporal integrator</td>
<td>81</td>
</tr>
<tr>
<td>4.5</td>
<td>STPL results</td>
<td>82</td>
</tr>
<tr>
<td>4.6</td>
<td>An overview of the CASP model</td>
<td>83</td>
</tr>
<tr>
<td>4.7</td>
<td>Growth of masking functions</td>
<td>85</td>
</tr>
<tr>
<td>4.8</td>
<td>Masking threshold predictions of the CASP model</td>
<td>86</td>
</tr>
<tr>
<td>4.9</td>
<td>CASP model predictions of forward masking</td>
<td>87</td>
</tr>
<tr>
<td>4.10</td>
<td>Binaural processing with the contralateral inhibition model</td>
<td>89</td>
</tr>
<tr>
<td>4.11</td>
<td>Features of the masking models</td>
<td>89</td>
</tr>
<tr>
<td>4.12</td>
<td>Binary search algorithm</td>
<td>92</td>
</tr>
<tr>
<td>4.13</td>
<td>CASP masking threshold predictions over time</td>
<td>94</td>
</tr>
<tr>
<td>4.14</td>
<td>Correlation between predicted and observed target to interferer ratios</td>
<td>97</td>
</tr>
<tr>
<td>4.15</td>
<td>Mean and predicted acceptability scores</td>
<td>98</td>
</tr>
<tr>
<td>5.1</td>
<td>User interface for speech intelligibility experiment</td>
<td>115</td>
</tr>
<tr>
<td>5.2</td>
<td>User interface for the distraction experiment</td>
<td>117</td>
</tr>
<tr>
<td>5.3</td>
<td>Intelligibility scores across subjects</td>
<td>119</td>
</tr>
<tr>
<td>5.4</td>
<td>Intelligibility scores separated by interferer programme and target corpus</td>
<td>123</td>
</tr>
<tr>
<td>5.5</td>
<td>Intelligibility scores separated by interferer programme</td>
<td>124</td>
</tr>
<tr>
<td>5.6</td>
<td>Mean distraction scores separated by interferer programme and absolute level</td>
<td>128</td>
</tr>
<tr>
<td>5.7</td>
<td>Usage of 'inaudible' and 'acceptable' check boxes</td>
<td>128</td>
</tr>
<tr>
<td>5.8</td>
<td>Mean probability target masked</td>
<td>131</td>
</tr>
<tr>
<td>5.9</td>
<td>An overview of various metrics across signal to noise ratio</td>
<td>133</td>
</tr>
<tr>
<td>5.10</td>
<td>Intelligibility scores for each subject</td>
<td>139</td>
</tr>
<tr>
<td>5.11</td>
<td>Sentence based speech reception thresholds</td>
<td>145</td>
</tr>
<tr>
<td>5.12</td>
<td>Intelligibility scores separated by interferer programme</td>
<td>147</td>
</tr>
<tr>
<td>5.13</td>
<td>Mean intelligibility scores separated by speech content in the interferer</td>
<td>148</td>
</tr>
<tr>
<td>5.14</td>
<td>Logistic fit to intelligibility scores separated by interferer</td>
<td>150</td>
</tr>
<tr>
<td>5.15</td>
<td>Logistic fit to intelligibility scores separated by presence of speech in the interferer</td>
<td>151</td>
</tr>
<tr>
<td>5.16</td>
<td>Mean distraction scores across signal to noise ratio</td>
<td>153</td>
</tr>
<tr>
<td>5.17</td>
<td>Example differences between distraction scores across subjects and interferer</td>
<td>153</td>
</tr>
<tr>
<td>5.18</td>
<td>Mean distraction scores for interferers not featuring speech</td>
<td>156</td>
</tr>
<tr>
<td>5.19</td>
<td>A principal component analysis of distraction scores based on the presence of speech in the interferer</td>
<td>156</td>
</tr>
<tr>
<td>5.20</td>
<td>Mean distraction scores across signal to noise ratio for interferers featuring speech</td>
<td>158</td>
</tr>
<tr>
<td>5.21</td>
<td>Mean distraction scores separated by the presence of speech in the interferer</td>
<td>159</td>
</tr>
<tr>
<td>5.22</td>
<td>Distraction scores separated by the presence of speech in the interferer</td>
<td>160</td>
</tr>
<tr>
<td>5.23</td>
<td>Logistic fits to distraction curves</td>
<td>161</td>
</tr>
</tbody>
</table>
List of Figures

5.24 Acceptability scores separated by the presence of speech in the interferer . . 163
5.25 Acceptability scores separated by interferer and subject groupings . . . . . 164
5.26 Mean acceptability scores separated by interferer and subject groupings . . 165
5.27 Masking scores separated by subject . . . . . . . . . . . . . . . . . . . . . . 166
5.28 Masking scores separated by interferer programme . . . . . . . . . . . . . 168
5.29 The relationship between interferer intelligibility, acceptability and
distraction scores . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 170
5.30 An overview of distraction, acceptability, masking and speech
intelligibility scores across signal to noise ratio . . . . . . . . . . . . . . . . . 172
5.31 An overview of measures across signal to noise ratio considering high and
low tolerance acceptability scores . . . . . . . . . . . . . . . . . . . . . . . . . . 173

6.1 An overview of the STOI model . . . . . . . . . . . . . . . . . . . . . . . . . 178
6.2 An overview of the TFMLM model . . . . . . . . . . . . . . . . . . . . . . . 180
6.3 SII logistic fit . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 183
6.4 SII with time windowing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 184
6.5 logistic functions for mapping SII to intelligibility . . . . . . . . . . . . . . 185
6.6 STOI predictions and intelligibility scores . . . . . . . . . . . . . . . . . . . 186
6.7 Logistic fit to STOI predictions compared with intelligibility scores . . . . . 186
6.8 TFMLM scores and intelligibility scores . . . . . . . . . . . . . . . . . . . . 187
6.9 logistic fit to TFMLM and intelligibility scores . . . . . . . . . . . . . . . . 187
6.10 CSII scores and intelligibility scores . . . . . . . . . . . . . . . . . . . . . . 188
6.11 Logistic fit to CSII and intelligibility scores . . . . . . . . . . . . . . . . . . 188

7.1 The user interface of the acceptability experiment . . . . . . . . . . . . . . . 192
7.2 The distribution of acceptability scores by subject . . . . . . . . . . . . . . . 199
7.3 Acceptability anchor scores separated by musical training . . . . . . . . . 200
7.4 Acceptability scores separated by training . . . . . . . . . . . . . . . . . . . 200
7.5 Acceptability scores across signal to noise ratio . . . . . . . . . . . . . . . 201
7.6 The user interface used by subjects to provide binary acceptability data . . . 202
7.7 Mean acceptability scores plotted against signal to noise ratio . . . . . . . 204

8.1 Cross validation of the benchmark model . . . . . . . . . . . . . . . . . . . . 220
8.2 Benchmark model validation 1 acceptability scores . . . . . . . . . . . . . . 221
8.3 Benchmark model validation 2 acceptability scores . . . . . . . . . . . . . . 222
8.4 Accuracy and generalisability of the CASP based acceptability model . . 223
8.5 Predicted acceptability scores plotted against the features of the CASP
based acceptability model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 225
8.6 Mean acceptability validation scores for the CASP based acceptability model 226
8.7 Accuracy and generalisability of the extended acceptability model . . . . . 230
8.8 Validation 1 acceptability score predictions of the extended acceptability
model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 233
8.9 Mean acceptability validation 2 scores for the extended acceptability model 234
8.10 Accuracy and generalisability of the PEASS based acceptability model . . 236
8.11 Validation 1 acceptability score predictions of the PEASS-based
acceptability model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 239
8.12 Mean acceptability validation 2 scores for the PEASS-based acceptability
model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 239
8.13 PESQ and POLQA predictions for validation 1 . . . . . . . . . . . . . . . . 245
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>A method for the prediction of acceptability</td>
</tr>
<tr>
<td>10.1</td>
<td>Mean and standard deviation magnitude spectra for target programmes (left), and a duplicate showing only frequencies up to 4 kHz (right). Black lines show the mean, with blue and red lines showing upper and lower standard deviations.</td>
</tr>
<tr>
<td>10.2</td>
<td>Mean and standard deviation of the magnitude spectra for interferer programmes (left), and a duplicate showing only frequencies up to 4 kHz (right). Black lines show the mean, with blue and red lines showing upper and lower standard deviations.</td>
</tr>
<tr>
<td>10.3</td>
<td>Difference between the mean magnitude spectra of the target and interferer programmes (left), and a duplicate showing only frequencies up to 4 kHz (right).</td>
</tr>
</tbody>
</table>
## List of Tables

2.1 An overview of the relevance of various masking phenomena .......................... 47
3.1 An analysis of variance of masking experiment data ................................. 55
3.2 Subject mean acceptability differences between repeats .......................... 56
3.3 An analysis of variance of acceptability scores ........................................ 62
3.4 An analysis of variance of the differences between masking and acceptability thresholds ................................................................. 69
3.5 Subsets of differences between masking and acceptability thresholds ........ 70
5.1 Combinations of stimulus levels for pilot speech intelligibility experiment .... 112
5.2 Interferer stimuli for pilot speech intelligibility experiment ....................... 114
5.3 An analysis of variance of intelligibility scores ......................................... 120
5.4 An analysis of variance of keyword intelligibility scores .......................... 121
5.5 Linearly interpolated SRTs for intelligibility separated by interferer programme ................................................................. 124
5.6 An analysis of variance of distraction scores ............................................ 125
5.7 Logistic regression on binary masking data ................................................. 129
5.8 Logistic regression on binary acceptability data ......................................... 130
5.9 Interferer stimuli used in speech intelligibility experiment ........................ 136
5.10 Subject scores for training sentences ...................................................... 139
5.11 Statistics of intelligibility scores separated by signal to noise ratio ............ 140
5.12 Word-based SRTs across interferer programmes ...................................... 143
5.13 Sentence-based SRTs across interferer programmes .................................. 143
5.14 An analysis of variance of intelligibility scores ....................................... 145
5.15 An analysis of variance of keyword intelligibility scores ......................... 146
5.16 Subsets of intelligibility scores .............................................................. 146
5.17 An analysis of variance of distraction scores ........................................... 154
5.18 Subsets of distraction scores ................................................................. 154
5.19 An analysis of variance of distraction scores for non-speech interferers ... 155
5.20 An analysis of variance of distraction scores for interferers featuring speech 157
5.21 Correlations between acceptability, masking, and signal to noise ratio ... 162
6.1 Prediction error for intelligibility models .................................................. 189
7.1 The probability of correctly guessing hidden anchors .............................. 193
7.2 Acceptability subject disagreement ......................................................... 198
8.1 A comparison of acceptability data sets .................................................. 209
8.2 Features, coefficients, and VIF for steps of the CASP based acceptability model ................................................................. 224
### List of Tables

8.3 Features, coefficients, and VIF for steps of the extended acceptability model \(231\)

8.4 Features, coefficients, and VIF for steps of the PEASS-based acceptability model \(237\)

8.5 A side-by-side comparison of the performance of two acceptability models \(243\)

8.6 Accuracy for PESQ and POLQA models \(244\)
## List of Equations

3.1 Pearson’s product-moment correlation coefficient .......................... 71
3.2 Linear regression of masking thresholds to thresholds of acceptability .... 71
3.3 The calculation of root-mean squared error ................................... 72

4.1 A model of loudness based on (Zwicker and Fastl 1990) ..................... 76
4.2 The calculation of the specific loudness based on (Zwicker and Fastl 1990) 78
4.3 The calculation of epsilon insensitive root-mean squared error ............ 95
4.4 the 95% confidence interval based on the normal distribution .......... 95
4.5 The calculation of acceptability thresholds using an adjusted linear regression to masking thresholds ........................................... 99

5.1 The calculation of intelligibility scores ........................................ 108
5.2 The calculation of keyword intelligibility scores .......................... 108
5.3 Prediction of intelligibility using a logistic fit to the SNR ................. 142
5.4 Calculation of speech reception thresholds from a logistic fit to intelligibility 143
5.6 General logistic fit to SNR for intelligibility ............................... 150
5.7 The adapted general logistic fit to distraction ............................... 159
5.8 A general logistic fit to distraction scores ................................... 162

6.1 Formulation of the speech intelligibility Index ............................. 176
6.2 Formulation of the band audibility function used in the calculation of the speech intelligibility index ............................................. 177
6.3 TFMLM equation 1 ............................................................. 179
6.4 TFMLM equation 2 ............................................................. 179
6.5 TFMLM equation 3 ............................................................. 180
6.6 TFMLM equation 4 ............................................................. 180
6.7 CSI equation 1 ................................................................. 181
6.8 CSI equation 2 ................................................................. 182
6.9 CSI equation 3 ................................................................. 182
6.10 Lower limit on the output of the logistic fit to the speech intelligibility index 185

7.1 The probability of correctly guessing hidden high or low anchors ......... 193
7.2 The probability of correctly guessing all hidden anchors .................... 193
7.3 Calculation of the phi coefficient ............................................ 197

8.1 Binary approximation to the normal 95% confidence interval ............ 210
8.2 Linear regression of all other features to test feature ..................... 210
8.3 Calculation of the variance inflation factor ................................ 210
8.4 General formulation of a multi linear regression model ................... 215
8.5 An example of the combinatorial explosion ............................... 216
<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.6</td>
<td>A benchmark model based on a linear regression to SNR</td>
<td>219</td>
</tr>
<tr>
<td>8.7</td>
<td>The linear regression CASP-based acceptability model</td>
<td>225</td>
</tr>
<tr>
<td>8.7</td>
<td>The linear regression of the extended acceptability model</td>
<td>232</td>
</tr>
<tr>
<td>8.8</td>
<td>The linear regression of the PEASS-based acceptability model</td>
<td>238</td>
</tr>
<tr>
<td>8.10</td>
<td>Setting limits on acceptability predictions</td>
<td>241</td>
</tr>
</tbody>
</table>
Publications Arising from this Thesis

Conference Papers


Acknowledgements

A large number of people have aided with the production of this thesis in a variety of ways, and I would like to take this opportunity to offer my gratitude.

Firstly, I would like to thank Bang & Olufsen for the financial support without which this work could not have begun. Additionally, I am grateful for the time and expertise given by all those with whom I have worked on the POSZ project. I am especially grateful to my supervisors Russell Mason, and Christopher Hummersone, as well as Søren Bech of Bang & Olufsen for their guidance and advice over the last three years. Their help was instrumental in keeping me on track when I threatened to get distracted by interesting, but irrelevant, areas of research. Even more importantly, they were extremely generous with their time, whether it was to be spent explaining a mathematical concept or revising the eighteenth draft of a report.

Additional thanks go to my colleagues on the POSZ team, especially to Jon Francombe, Phil Coleman, and Marek Olik for variously helping out with experiments, answering thoughtless Matlab coding questions, discussing mind-boggling mathematics, and most importantly for forming a great team with whom it has been a pleasure to work. Further thanks go to my other colleagues at the Institute of Sound Recording: Will Evans, Daisuke Koya, Andy Pearce, Cleo Pike, Toby Stokes, Kirsten Hermes, Martin Dewhirst, and Tim Brookes, for providing an amiable working environment filled with much deep and engaging conversation (over many Friday morning coffees and pastries).

I also wish to extend my gratitude for the various computational models without which this work could not have been completed. Opticom have been very generous in allowing the usage of their PEXQ suite of quality models for comparative evaluation. Morten Loewe Jepsen was kind enough to provide code for the CASP model, upon which much of this work is based. Additionally, the PEASS toolbox of Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, the PRAAT speech analysis tool of Paul Boersma and David Weenink, the audio systems array processing toolbox of Kevin Donohue, and the SpecCentroid Matlab function of Frederik Nagel and Michael Grobach, were all a great aid throughout the work presented in this thesis.

Finally I would like to thank all those who took part in my listening tests, with special thanks to Anna Grayson and Spencer Hudson for lending their time and voices in preparation of speech stimuli.
This work is part of a larger body of collaborative research aimed at producing a Perceptually Optimised Sound Zone (POSZ) system. The POSZ system is designed to reproduce monophonic audio programmes to multiple listening zones within a single physical space. Although this collaborative research could have applications in many domains, the POSZ project is specifically concerned with audio reproduction within automotive and domestic environments.

The problem arises when multiple listeners reside within a single acoustic space and wish to listen to different audio programmes. There are some cases in which headphones may not be an appropriate solution because they increase physical and social isolation and can sometimes be impractical and uncomfortable when used for extended periods. If multiple audio programmes are reproduced through a conventional audio system (as in fig. 1.1) there would usually be overwhelming levels of crosstalk. This would ordinarily be undesirable for listeners. The POSZ system, therefore, aims to sufficiently minimise crosstalk such that spatial ‘zones’ can be generated within a room for different programmes.

In order to achieve this aim, a multi-disciplinary approach is taken to tackle the various engineering and psychoacoustic aspects of the work. The engineering aspects are focused on the design of the sound zoning system itself, while the psychoacoustic aspects are concerned with predicting and evaluating the effects of the competing sound zones on the quality of the listening experience.

1.1 The aim of this work

This work aims to devise a computational evaluation model for the POSZ system which predicts the performance of the system by modelling the listening experience and reporting useful indicators of quality. Figure 1.2 illustrates the time-varying effect of interference on the listening experience within an arbitrary listening zone. For an ideal POSZ system the evaluation model would produce results indicating a performance line which always remains within category 4.

Unfortunately, however, the sizes of the categories in fig. 1.2 will vary across listeners and for different combinations of target and interferer programme. This variability of
Chapter 1: Introduction

Figure 1.1: A room with multiple sound zones and multiple sound sources

Figure 1.2: An example of the prediction of the overall quality of the POSZ system. The blue line represents the changing level difference between target and interferer over time for an arbitrary combination of programmes.
the effect of interference makes the prediction of the Quality of Listening Experience (QoLE) more difficult, but not impossible. Specifically the metric that is most indicative of listening quality may depend upon the listening task involved. For example if the programme for zone A is music and the programme for zone B is speech, then it is likely that speech intelligibility will be a good indicator of the QoLE in zone B while the same attribute is likely to be of little importance in zone A. It should also be noted that the QoLE will depend upon both the quality of the target programme as well as the effect of the interference, although this work is primarily concerned with interference prediction. Systems developed in this work will not, therefore, predict QoLE directly but only the proportion of QoLE attributed to the effect of the interference (QoLEₐ), which specifically excludes consideration of the quality of the target programme not due to the presence of the interferer. Both QoLEₐ and the proportion of QoLE attributed to the effect of the target (QoLEₜ) are likely to be determined by a wide range of factors including level, programme, and frequency content. The conceptual framework can therefore be expressed with the following statement:

The QoLE constitutes the combined effect of the QoLEₜ and the QoLEₐ.

Since the specific factor of target quality is not considered in this work a POSZ system operating within category 4 will be considered to have maximum QoLEₐ and is therefore considered ideal. Conversely, category 1 represents the worst case performance for a POSZ system. It is not expected that any POSZ system would operate within category 1 for an extended period since even a conventional loudspeaker system replaying two programmes would be likely to exceed this performance most of the time. Even so, this result is possible and should appear on the scale of QoLEₐ. Categories 2 and 3 may more accurately be considered as one continuum of perception, however it is reasonable to posit that there will be some conditions wherein the interference is audible but listeners consider this level of interference to be ‘acceptable’ for consumption at home or in an automobile, while there will be other cases where the level of interference is considered ‘unacceptable’. A line can thus be drawn between these which indicates a ‘threshold of acceptability’. Since it is considered unlikely that a POSZ system will be able to regularly produce listening scenarios falling within category 4, the goal of the POSZ system should therefore be to operate, wherever possible, above the threshold of acceptability. More generally, for populations of listeners one could map the range with a metric called ‘acceptability’ describing the probability that a listener picked at random would find a listening scenario to be acceptable. This research is therefore primarily concerned with making predictions of and mapping out the range of acceptability.
1.2 Auditory masking in sound zones

In order to define which category a listening scenario falls into, it is necessary to predict the audibility of the target and interferer programmes. In order to achieve this it is necessary to have an understanding of auditory masking.

Auditory masking (or simply ‘masking’) is a broad term that refers to a range of psychoacoustic phenomena in which one sound appears to be “obscured, or rendered inaudible by the presence of other sounds” (Moore, 2004). This definition is a useful way of explaining, in simple terms, what is usually meant by masking but a strict definition is not universally agreed upon. It is arguable that the use of the word ‘obscured’ is inappropriate because it allows some phenomena to be described as masking even when it is more meaningful to use another term such as ‘interference degradation’. For example when two speech programmes are simultaneously presented at the same level it is improbable that either will be inaudible while both programmes will be ‘obscured’ to some extent by the other.

Masking experiments usually require a listener to perform one of three processes: detection (noticing the presence of a signal), discrimination (noticing that two signals differ), or recognition (reporting a known signal). Durlach (2006) points out that masking is more closely related to the detection paradigm than to discrimination or recognition. Durlach’s description of masking, however, refers to a target being “degraded” by the presence of a masker, a term which implies discrimination and recognition difficulties. He identifies the primary source of confusion: “in many circumstances discrimination implies that detection was achieved, and in a two-stimulus experiment both terms are interchangeable with recognition”.

Ideally the terms ‘masking’, ‘degradation’ and ‘interference’ would each have meanings which refer to specific mechanisms of auditory perception. In practice, however, the terms are used to refer to phenomena which are the result of mechanisms that are not yet fully understood. For this reason, the various types of masking are named not by their underlying mechanisms but descriptively according to the conditions wherein they occur. For the purposes of this work the term masking will be used to refer to any phenomenon where a signal is rendered inaudible. The various masking phenomena are therefore named, with reference to conventions in the literature, according to the conditions in which they occur.

For an ideal implementation of the POSZ system a target signal would be audible within each sound zone, and all non-target signals (and any other extraneous noises) would be completely inaudible. This could be achieved if the target signal masked the presence of all unwanted sounds. When trying to produce separate sound zones without acoustic barriers, however, the non-target signal is likely to be significant in each sound zone (see fig. 1.1). If the level difference within each zone is great enough, the quieter
signal will be masked and will therefore be imperceptible. This scenario is defined by the two extreme regions in fig. 1.2. An understanding of the factors which determine the audibility of a programme are, therefore, an appropriate objective for this research project.

It is interesting to note that a large proportion of recent research into masking tends to focus on finding mechanisms that enable masking release (unmasking). This approach has been motivated, in part, by consumption in applications such as Automatic Speech Recognition (ASR), where if the mechanisms which cause unmasking in human listeners can be understood they might be applied to improve such systems. In the case of creating sound zones, however, masking of the non-target signal is a highly desirable phenomenon. As such, special cases of unmasking are important because a failure to predict them will produce an exaggerated prediction of the performance of the system, and thus reduce the extent to which it can be optimised.

Accurate prediction of masking, including the effect of unmasking phenomenon, will allow categories 1 and 4 (see fig. 1.2) to be well defined. As such, a method for the prediction of the audibility of programmes under pre-specified scenarios is an important aim for this work.

1.3 Speech perception

If masking prediction accounts for categories 1 and 4 (see fig. 1.2) in the prediction of acceptability, then the prediction of categories 2 and 3 requires determining the extent to which a listener is likely to find the listening scenario acceptable beyond simply the audibility of either programme. This is likely to depend upon the content of the target and interferer programmes as well as the task of the listener. It is reasonable to suggest that for programme combinations involving speech there may be special aspects of speech perception to consider; for example with a speech target programme the listener’s primary task is to understand the meaning of the speech. As such, the speech intelligibility is extremely important and therefore likely to constitute an important part of the acceptability within categories 2 and 3. If the interferer is speech, the intelligibility of the interfering speech may also affect the acceptability of the listening scenario.

1.4 Research questions

With the project aims outlined, a number of research questions present themselves. In order to predict the acceptability of auditory interference scenarios produced by sound zoning systems it is first necessary to define categories 1 and 4 by considering the
audibility of the target and interferer. It is therefore important to know “what are the factors which determine whether an auditory stimulus will be masked by the presence of a second stimulus?”, and “what is the relevance and importance of each factor for sound zones?” Once these research questions have been answered it is necessary to consider how acceptability varies with Signal to Noise Ratio (SNR), and how this relates to audibility, thus posing the research questions: “what is the range of SNRs over which acceptability primarily varies?”, and “is there a relationship between masking and acceptability?”. In order to utilise the findings of these research questions it is important to then consider how predictions can be made about which category of fig. 1.2 any auditory interference scenario is operating under. To do this it is necessary to know “how can auditory masking be predicted?”

It was noted that programmes featuring speech are likely to have special circumstances, in which intelligibility is likely to play an important role. Considering this, then, it is important to consider “what relationships exist between intelligibility, acceptability, and other relevant measures?” If intelligibility is informative about the acceptability of auditory interference scenarios, it would then be useful to know “how can the intelligibility of speech within auditory interference scenarios be predicted?” If intelligibility is not entirely responsible for the acceptability of scenarios featuring speech it would then be necessary to consider the question: “how can the acceptability of auditory interference scenarios featuring a speech target be determined?”, and it may be necessary to consider the related question “what is the general distribution of listener acceptability responses?”.

Finally, the findings of the previous chapters would need to be drawn together to address the primary goal, to answer the question “how can the acceptability of listening scenarios featuring auditory interference be determined?” These questions form the conceptual basis of this project, and are restated as the research questions of this thesis as follows:

1. what are the factors which determine whether an auditory stimulus will be masked by the presence of a second stimulus?
2. what is the relevance and importance of each factor for sound zones?
3. what is the range of SNRs over which acceptability primarily varies?
4. is there a relationship between masking and acceptability?
5. how can auditory masking be predicted?
6. what relationships exist between intelligibility, acceptability, and other relevant measures?
7. **how can the intelligibility of speech within auditory interference scenarios be predicted?**

8. **how can the acceptability of auditory interference scenarios featuring a speech target be predicted?**

9. **what is the general distribution of listener acceptability responses?**

10. **how can the acceptability of listening scenarios featuring auditory interference be predicted?**

### 1.5 The structure of this thesis

Chapter 2 investigates the first two questions by detailing various masking experiments and outlining a range of masking phenomena. In chapter 3 an experiment is described to obtain masking and acceptability data for a variety of ecologically valid programme combinations in order to address the third and fourth questions. Then in chapter 4 masking threshold models are listed and compared before one is selected for implementation and is tested on the data from this experiment; this accounts for the fifth research question. Chapter 5 deals with the sixth question posed by describing experiments investigating the relationship between intelligibility, acceptability, and other related measures for auditory interference scenarios. The seventh question is considered in chapter 6, where speech intelligibility models are described and evaluated using the intelligibility data from the experiment in chapter 5. The eighth and ninth research questions are addressed by chapters 7 and 8; in the former two experiments are described aimed at collecting acceptability data for a wide range of auditory interference scenarios and from a variety of listeners, and the latter of which describes the construction of a model for the prediction of the acceptability of listening scenarios featuring speech as the target programme. Chapter 9 draws the work of all the previous chapters together, and considers how to use this work to build a general method for the prediction of acceptability, thus addressing the tenth and final research question. Finally, chapter 10 restates these research questions and summarises the findings of each chapter, as well as outlining the scope and limitations of the work and proposing relevant future work.
Chapter 2
Auditory Masking

The previous chapter introduced auditory masking as the psychoacoustic phenomenon wherein the presence of one auditory stimulus inhibits the perception of a second. Auditory masking occurs under a range of listening scenarios and these entail a variety of related phenomena which should be understood in order to assess their relevance and importance to the evaluation of sound zones.

The primary goals for this chapter are therefore to answer the following questions: “What are the factors which determine whether an auditory stimulus will be masked by the presence of a second stimulus?”, and “what is the relevance and importance of each factor for sound zones?”

In section 2.1 the measurement of masking is described to provide a point of reference for detailed discussion. Auditory filters and the masking experiments which define them are then introduced in section 2.2, and it is shown that this leads to a basic model of masking. Simultaneous, forwards and backwards masking are subsequently introduced as the fundamental basis of masking phenomena in section 2.3. Partial masking is discussed in section 2.4 to highlight the relationship between masking and loudness. The special cases of binaural unmasking, informational masking and comodulation masking release are considered in sections 2.5 to 2.7 as scenarios not easily explained by the previously discussed models. In each case the likelihood of occurrence and impact of the phenomenon is considered. Finally the various masking phenomena are summarised and prioritised in terms of their applicability to the sound zone problem, potential impact, and modelling complexity in section 2.8.

2.1 The measurement of auditory masking

Before the various masking phenomena are discussed it is worth noting precisely how auditory masking is measured. The simplest way to directly ascertain the conditions of auditory masking is to conduct a listening test and find (for fixed target and interferer programmes) the level of the target programme at which the interferer is just inaudible. This level is then defined as the ‘masking threshold’, because any increase in level would cause the signal to once more become audible and thus no longer be masked. Masking thresholds are usually indicated in decibels (dB) of Sound Pressure Level (SPL).
Given this approach it would seem that a likely predictor of auditory masking would be the Target to Interferer Ratio (TIR) at the listening position. Unfortunately, however, the phenomenon of auditory masking depends heavily on the spectrotemporal content of both the masking programme and the masked programme. Therefore, while the TIR is a useful descriptor of the relative acoustic intensities at the listening position it generally does not predict the phenomena of auditory masking.

Another measure of auditory masking, which is usually used when investigating the specific mechanisms of masking, is known as the masking level. The masking level is the difference in level between a masking threshold and the equivalent threshold of audibility for that signal, or between masking thresholds in two different experimental conditions. Thus masking levels are a measure of the extent of the additional change in level required to render audible a target programme, and are therefore useful for describing the variety in effect for different maskers (masking stimuli) and maskees (masked stimuli).

Now that the circumstance of auditory masking and its measurement have been outlined, an understanding of the basic mechanisms are introduced.

2.2 Auditory filters and masking experiments

A foundational conceptual basis for the phenomena of auditory masking was outlined as a result of early experiments by Fletcher (1940). This work introduced the concept of auditory filters as a way to explain masking experiment results. Auditory filters provide the conceptual framework on which a large portion of masking experiments and models are based, and are therefore deserving of some explanation.

2.2.1 Auditory filters and the power spectrum model

An experiment, originally conducted by Fletcher (1940) and subsequently repeated many times e.g. (Hamilton 1957; Greenwood 1961; Spiegel et al. 1981), found masking threshold curves for individual tones when masked by a band of noise centred on the frequency of the tone. This was accomplished by systematically expanding the bandwidth of the noise signal and observing the change in masking threshold (see fig. 2.1).

As the Narrow Band of Noise (NBN) increases in width there is a resultant decrease in SNR for a fixed signal level. The expected result would therefore be a masking threshold which increases monotonically with noise bandwidth. The results of these experiments, however, indicate that as the NBN is widened the masking threshold increases monotonically only until a specific bandwidth is reached, after which no further increase in masking threshold is observed (see fig. 2.2).
Figure 2.1: A simple band widening experiment in which the signal tone is presented at a level where it is clearly audible. The listener then adjusts the level of the signal tone until it is just masked by the narrow band of noise. The width of the narrow band of noise is then slightly increased and the process is repeated.

Figure 2.2: Results from a band widening experiment in which the masking threshold was determined for a tone signal of 2 kHz with a noise band masker. Adapted from (Schooneveldt and Moore, 1989).
Chapter 2: Auditory Masking

Figure 2.3: The loudness of various bandwidths of noise centred at 1 kHz with various intensities. Adapted from Feldtkeller and Zwicker (1956).

Fletcher (1940) referred to the bandwidth at which the signal threshold no longer increased as the ‘critical bandwidth’. Fletcher then suggested that the auditory periphery behaves as if it contains a bank of bandpass filters, which are now commonly referred to as the ‘auditory filters’ (Moore 2004). Feldtkeller and Zwicker (1956) later demonstrated that the perceived loudness of a band of noise did not increase with increasing bandwidth below a certain threshold (see fig. 2.3). The results of both experiments indicated a common ratio between noise band centre frequency and the bandwidth at which the perception of stimuli changes. This implies that there is a strong link between auditory masking and loudness, and it allows the use of both band widening experiments to determine the critical bandwidth.

Based on his work determining the critical bandwidths and conceptualising the auditory periphery as a bank of bandpass filters, Fletcher (1940) went on to propose the power spectrum model of masking. The power spectrum model suggests that the auditory masking of a signal is determined by the SNR at the output of the auditory filters. This proposition is supported by the band widening experiments because as the bandwidth of the masking noise exceeds the bandwidth of the auditory filter in which the signal tone is presented the additional noise energy (passing through adjacent auditory filters)
Chapter 2: Auditory Masking

is perceived as distinct from the signal and thus disregarded causing no further change in masking threshold. A constant of signal detection, \( K \), therefore can be posited to describe the relationship between the relative levels of signal and noise at which auditory masking occurs. Fletcher (1940) expressed the power spectrum model of masking mathematically in the following way:

\[
K = \frac{P}{N} = \frac{P}{W \times N_o}
\]  

(2.1)

Where \( P \) is the power of the signal (at the masking threshold), and \( N \) is the power of the noise passed through the auditory filter, which is comprised of \( W \), the width of the auditory filter centred on the signal, and \( N_o \), the noise power density (Moore 2004). A simple rearrangement of eq. (2.1) allows for an estimation of the width of an auditory filter assuming that the power of the tone, the noise power density and the signal detection constant are all known. This allows auditory filter widths to be predicted without conducting masking experiments. Another rearrangement allows for the estimation of the noise power required to mask a tone of power \( P \), if the bandwidth of the noise and the signal detection constant of the listener are both known. Under this framework, when the bandwidth of the noise is less than \( W \), all of the noise is passed through the auditory filter so the equation can be simplified to:

\[
K = \frac{P}{N_o}
\]  

(2.2)

The value of \( K \) varies amongst listeners and for different centre frequencies although Scharf (1970) showed that \( K \) is typically around 0.4. Equations (2.1) and (2.2) of the power spectrum model are perhaps the most basic, and fundamental, predictors of auditory masking.

The power spectrum model of masking is, however, based on results from tone in noise experiments and thus any application of the model to other signal types makes a number of assumptions. Firstly this model assumes that when trying to hear a specific signal the auditory filter is centred on that signal and any noise which falls outside of the auditory filter is effectively ignored, and thus plays no part in masking (Moore 2004). Secondly, the model assumes that masking occurs due to the SNR at the output of the auditory filter and is unaffected by other acoustic cues. Thirdly, the model is based on the long-term power spectrum of the signal and masker and thus assumes that any temporal variations are either irrelevant or negligible. As discussed in later sections, each of these assumptions are untenable in more complex listening scenarios.

Where the assumptions of the the power spectrum model of masking are violated, it is useful to compare the predicted outcome with the actual response of human auditory processing. Occasions where the auditory system makes use of spectrottemporal information remote from the signal to aid signal detection (e.g. such as comodulation
masking release discussed in section 2.7), and occasions where the auditory system is unable to ignore spectrotemporal information which hinders signal detection (e.g. such as informational masking discussed in section 2.6), are both examples of the incompleteness of the power spectrum model of auditory masking. The simplicity of the model, however, and its effectiveness under simple listening scenarios, makes it a convenient starting point for describing the mechanism of auditory masking.

2.2.2 Off-frequency listening

The power spectrum model assumes that a listener’s auditory filter is centred on the target signal. It is possible, however, that a listener could achieve an increased SNR by attending to an auditory filter centred at a different frequency in order to minimise the noise level (see fig. 2.4). This phenomenon is known as Off Frequency Listening (OFL).

Figure 2.4 shows a scenario in which the auditory filter is not centred on the signal tone. The output of the auditory filter due to the signal tone is diminished by this behaviour but the output due to the noise content is more significantly diminished. The resultant SNR is therefore greater than it would be if the auditory filter was centred on the signal.

When listening to more complex stimuli, however, OFL might not be beneficial. This is because OFL will yield little or no signal detection advantage for a masker with components both above and below the frequency of the signal (see fig. 2.5). In the top panel of fig. 2.5 sinusoidal maskers are presented separately. The resultant masking threshold of each masker is the same because OFL is utilised to maximise signal detection. In the bottom panel, both maskers are presented simultaneously and OFL no longer offers any significant advantage.
Chapter 2: Auditory Masking

Figure 2.5: OFL and the non-linear addition of masking: Excitation patterns of two sinusoidal maskers at 700 Hz and 1300 Hz (solid line) presented separately (top panel) and together (bottom panel). The dotted line is the resultant masked threshold, and the arrows represent the optimal listening frequency for a tone presented at 1 kHz. Adapted from Van der Heijden and Kohlrausch (1994)
Chapter 2: Auditory Masking

One side effect of OFL is the non-linear addition of masking thresholds. In the top panel of fig. 2.5 OFL can be utilised to diminish the masking threshold. In the bottom panel where both tones are presented simultaneously there is no longer any advantage from OFL and thus the masking threshold increases by significantly more than the 3 dB that would otherwise be expected.

It may be, therefore, that for spectrally rich listening environments the effect of OFL can be generally ignored by a masking model. Even so, OFL cannot be ignored for simple cases such as those found in masking experiments which aim to determine the shape of the auditory filters. In order to control this confounding variable Patterson (1976) devised a masking experiment method called the ‘notched noise method’. This involves using a noise masker centred on the signal but with a notch at the signal frequency (see fig. 2.6). Varying the notch width varies the SNR in the auditory filter, as in band widening experiments, but the presence of equidistant noise bands prevents any advantage from auditory filters positioned off centre from the signal frequency. The notched noise method does assume that the auditory filters are symmetrical, however psychophysical tuning curve experiments have shown this to be a reasonable assumption for low and mid intensity signals (Moore 2004).

Using this method Patterson (1976) found data points which indicated that auditory filters were shaped according to the following equation:

$$10 \log_{10}(B) = 7.91 \log_{10}(f_0) - 2.71$$

(2.3)

Where $B$ is the bandwidth of the auditory filter and $f_0$ is the frequency of the signal, with both quantities measured in Hz.

In much the same way as the band widening experiment, this method can be used to derive the point at which noise no longer contributes to the masking threshold by
2.2.3 The roex auditory filter

Results from notched noise experiments indicate that auditory filters are of the shape illustrated in fig. 2.7 which can be modelled by a rounded exponential curve (roex) defined in (Patterson and Nimmo-Smith 1980) by:

\[ W(g) = (1 + pg)e^{-pg}, \]

where \( p \) tunes the steepness of the filter and \( g \) describes the distance between the frequency, \( f \), and the centre frequency of the filter, \( f_c \). \( g \) is defined as:

\[ g = \frac{|f - f_c|}{f_c}. \]

This auditory filter is not rectangular and thus cannot be meaningfully described using a critical bandwidth. One way to describe an auditory filter of this nature with a single number is by referring to its half power bandwidth (i.e. the 3 dB bandwidth). In psychoacoustics, however, a more common way is to use the Equivalent Rectangular Bandwidth (ERB) which is the bandwidth of a perfect rectangular filter which has a
passband response equal to the peak of the auditory filter and which passes the same total noise power (i.e. the integral will be the same).

Since the width of the auditory filter varies with frequency, Glasberg and Moore (1990) suggested the following equation to calculate the ERB:

\[
ERB = 24.7(0.00437f + 1) \tag{2.6}
\]

Using this relationship an auditory filter for any frequency can be described. Thus the ERB of an auditory filter at 1 kHz would be calculated as:

\[
ERB = 24.7\times5.37 = 132 \text{Hz} \tag{2.7}
\]

According to Glasberg and Moore (1990) the tuning parameter \( p \) should be set such that

\[
p(f_c) = \frac{4f_c}{ERBf_c}. \tag{2.8}
\]

Depending on the level of the stimulus, however, the auditory filters may not be symmetric. Specifically, the lower slope of the auditory filters tend to become flatter as the level of the input stimuli increases (Glasberg and Moore 1990).

### 2.2.4 The gammatone filter

While roex filters are a reasonable way of modelling an auditory filterbank, they are derived from tone in noise experiments and thus reveal only spectral information. The phase response of the roex filter, therefore, is not well defined. As such it is not possible to uniquely specify the impulse response of the filter, which limits its use (Patterson et al. 1988).

A solution to this problem was found by means of a biological experiment using the Reverse Correlation (RC) technique. The RC technique, devised by de Boer and de Jongh (1978), involved using micro-electrodes to measure the response of fibres in the auditory nerve of a cat that was presented with white noise. This technique effectively determined the impulse response of the cat’s basilar membrane. By fitting the statistical gammatone function to this data a gammatone filter was derived to model auditory filters. It has been shown that in general a fourth order gammatone filter can provide a close approximation to the spectrum of a roex filter (Patterson et al. 1988). Thus the advantage of the gammatone filter, over the roex filter, is that the original phase information is retained because the filter is derived directly from the impulse response.

While it is true that the impulse responses used to determine the gammatone filters are those of the basilar membrane of small mammals, Patterson et al. (1988) point out that
the results can be extrapolated to humans by scaling the bandwidths of the filters and that this approximation is preferable to using a roex filter where the phase response is ill defined. An example gammatone filterbank is shown in fig. 2.8.

Gammatone filterbanks are commonly used in models of the auditory periphery, an example of which is (Meddis et al. 2001a).

2.2.5 The dual resonance non-linear filter

A gammatone filterbank can be used to model the auditory filterbank for time-domain applications. The frequency response, however, is almost symmetrical about its centre frequency, while psychophysical tests show that for high stimulus intensities the auditory filters have an asymmetric frequency response with a shallower low frequency roll-off and a steeper high frequency roll-off (Lutfi and Patterson 1984).

One solution to this problem is to use a series of gammatone and low-pass filters in series to account for this asymmetry. The Dual Resonance Non-Linear (DRNL) filter proposed in (Meddis et al. 2001a) and subsequently modified in (Meddis et al. 2001b) is an example of this.

As illustrated in fig. 2.9, the Dual Resonance Non Linear (DRNL) filterbank is a system which uses a number of gammatone filterbanks. The low-pass filters are set such that
Chapter 2: Auditory Masking

Figure 2.9: An overview of the dual resonance non-linear filter, adapted from (Meddis et al. 2001a). The filter models basilar membrane velocity as a function of stapes velocity. In order to account for the asymmetrical frequency response and the nonlinear intensity response a number of gammatone filters and low pass filters are used, as well as a 'broken-stick' nonlinearity.

they produce a reduction of 6 dB at the critical frequency of the previous gammatone filter. This combination of gammatone filter followed by low pass filter produces a psychoacoustically informed asymmetrical response. The 'broken-stick nonlinearity' is a compressive function which mimics the power compression at the output of the basilar membrane occurring for stimuli of intensity 40-70 dB SPL.

In these ways the DRNL filter accounts for both the auditory filter asymmetry and intensity compression phenomena in the auditory periphery. Because of this accuracy the DRNL filter has been adopted for use in physiologically inspired models of the auditory periphery such as the Computational Auditory Signal-processing and Perception (CASP) model described in (Jepsen et al. 2008).

2.2.6 Excitation levels, excitation patterns, and the upward spread of masking

However the auditory filterbank is modelled it will usually contain a number of band-pass filters with overlapping tails. Thus if a sine tone is presented within one of these auditory filters, it is likely to also fall within the tails of a few adjacent filters (see fig. 2.10).

Many auditory models consider that the loudness of a tone can be described by the sum of the excitations produced at the outputs of all of the auditory filters which overlap the tone. This value is known as the excitation level.

Since a single tone is overlapped by auditory filters centred on a range of frequencies, the combined output to higher auditory processes is a combination of excitations of a range of intensities. Each excitation is produced by a different auditory filter and so a graph of intensities and centre frequencies can be drawn (see fig. 2.11).

The excitation pattern, shown in the bottom panel of fig. 2.11, is the combined output from the auditory filterbank for an input signal of a sine tone. Moore et al. (1997)
Figure 2.10: A sine tone presented at 1 kHz overlaps the tails of adjacent auditory filters. The relative intensity passed through each auditory filter is indicated by the gain of each auditory filter at 1 kHz (represented here by the height of the correspondingly coloured circle).
Chapter 2: Auditory Masking

(a) A representation of the (level-equalised) output of the DRNL filterbank (linear path shown only for clarity) of Meddis et al. (2001a). The low frequency slope of each filter is shallower than the high frequency slope, especially for filters with higher frequency centers. The circle representing the energy passed through each filter is shown positioned at the center frequency of the respective filter.

(b) When a line is drawn connecting the circles representing the energy passed by each auditory filter, the resulting shape is the excitation pattern of the 1 kHz tone.

Figure 2.11: An excitation pattern is calculated from a 1 kHz sine tone by marking the gain applied by each auditory filter to the 1 kHz sine tone.
have suggested that loudness (and by extension masking thresholds) may be a direct function of the excitation pattern produced at the output of the auditory filterbank. This idea is a natural extension to the power spectrum model, and seems plausible since the shape of the excitation pattern for a sine tone corresponds very well with the masking threshold curve it produces (Egan and Hake 1950).

Additional evidence can be found by observing the slopes of the excitation pattern. The high frequency slope of the excitation pattern is shallower than the low frequency slope (a direct consequence of the increasing width of auditory filters with frequency). This asymmetry of slopes is a well known characteristic of auditory masking known as the ‘upward spread of masking’.

The upward spread of masking is an interesting side effect of the proportional relationship between frequency and critical bandwidth. Egan and Hake (1950) showed that maskers are significantly more effective for higher frequency maskers than lower frequency maskers. This is an emergent property of the expanding widths of auditory filters with increasing centre frequency. When a signal is overlapped by fewer low frequency filter tails than high frequency tails the resulting masking threshold will tend to spread upwards in frequency, more effectively masking higher frequency sounds, as depicted in fig. 2.12.

It is notable that the asymmetry of auditory filters is dependent upon the level of the input stimulus. At high stimulus intensities the lower slopes of the auditory filters flatten out, which results in an excitation pattern with an exaggerated upward spread. For low to moderate intensities, however, the auditory filters are nearly symmetrical, so the expanding widths of auditory filters with increasing centre frequency is the only cause of asymmetry in the excitation pattern, and the upward spread of masking is less pronounced.

### 2.2.7 Extracting partials from complex tones

Another proposed method for predicting critical bandwidths involves picking out individual sinusoids from a sound comprising multiple sinusoids. The ability of listeners to achieve this task was tested in (Plomp 1964) and (Plomp and Mimpen 1968). In their experiments only the first 5-8 components could be picked out from the complex tone.

When this work is considered in the context of the power spectrum model, however, it could be assumed that sinusoids may only be picked out when they are alone within an auditory filter (i.e. they must be separated from their neighbour by at least one critical bandwidth). Since critical bandwidths increase with increasing frequency, high frequency partials are more difficult to discern.

If this assumption is valid the critical bandwidths could be estimated from the results
Figure 2.12: Results from masking experiments show a nonlinear increase in masking threshold for high frequencies with masker intensity. Adapted from Egan and Hake (1950)

of experiments that determine which partials may be extracted from complex tones. The results from these two studies agree with a curve approximately 1.25 times the bandwidth estimates for ERBs (Moore 2004).

This method is somewhat flawed however because Soderquist (1970) found that musicians were markedly superior at picking out partials from complex tones than non-musicians. This could be explained if musicians have finer auditory filters than non-musicians. Fine and Moore (1993), however, found that ERBs calculated via notched noise masking experiments were the same for musicians and non-musicians. As such it is likely that a higher cognitive process, which musicians have refined by experience, also affects the ability to discern partials from complex tones and this method is therefore unlikely to be the most reliable way of estimating critical bandwidths.

Under the power spectrum model of masking a complex sound that consists of components within one auditory filter should produce a masking threshold determined by the sum of their energies. It is reasonable to assume, therefore, that if the components of the complex sound have a wider frequency range then only information from a single auditory filter will be used, and signal detection should be less successful. Spiegel (1981), Buus et al. (1986), and Langhans and Kohlrausch (1992), however, found evidence to the contrary. Spiegel’s results suggested that the ear is capable of integration over bandwidths much greater than the auditory filter bandwidth, and Buus et al. found that widely spaced tones present in background noise were more easily detectable than any of the individual tones. This is, therefore, further evidence
suggesting that the extraction of partials from complex tones is not the best way to describe the width of the auditory filters, probably because higher level auditory processing is involved in the task.

2.2.8 Two-tone masking

In a scenario effectively the inverse of the notched noise method (two tones surrounding an NBN) Rabinowitz et al. (1980) increased the tone frequency separation until the masking threshold of the noise signal changed. This point could be considered to be the point at the edges of the critical band. Rabinowitz et al. found larger critical bandwidths than other experimental methods. There are complications, however, with this method which arise from combination products caused by interference between the lower tone and the noise. The combination products give cues to the presence of signals, even when the signals would be otherwise inaudible. Additionally the higher frequency tone is likely to be less efficient at masking the noise than the lower frequency tone, which diminishes the symmetry of the masker and further complicates interpretations of the results.

2.2.9 Discussion

Figure 2.13 shows results from harmonic and inharmonic partial extraction experiments alongside the bandwidths found using the noise widening and notched noise experiments. The results have relatively good agreement above 500 Hz. Below 500Hz OFL is particularly effective (because the auditory filters are narrower) which explains, to some extent, the discrepancy between results shown by the band widening and notched noise experiments.

It should be noted that critical bands, auditory filters and ERBs are all determined by the measurement of related phenomena. As such the distinction between these terms can easily become blurred. Fastl and Schorer (1986) elucidate, “[ERB] defines an auditory filter, while the [critical band] describes a change in subjective response, not confined to a certain filter shape”. The distinction is a subtle one, but highlights a difference in emphasis: ERBs and auditory filters relate to physiological mechanisms while critical bands describe perception. Thus for cases which are not explained by the power spectrum model, critical bandwidths may differ from the ERBs of auditory filters.

The notched noise method is generally thought to produce the most reliable estimates of the auditory filters, although modifications are sometimes made to the method to detect their asymmetry (Glasberg and Moore 1990).

A great deal of research into masking describes and illustrates findings by relating results to critical bands instead of absolute frequencies. This is useful in many cases
where there appears to be a special relationship between a specific result and the critical bands but it should not be forgotten that such comparisons inherently make the assumption that masking is determined by the SNR at the output of an auditory filter. Using this assumption, experimental procedures which try to explain the mechanisms of masking have tended to focus on the simplest cases, such as that of a tone masked by a broadband noise. This case has been tested in many different ways during the last century (Wegel and Lane 1924; Fletcher 1940; Greenwood 1961; Hellman 1972; Spiegel 1981), and results have demonstrated the following masking trends:

- Sounds mask other sounds which are of a similar frequency most effectively (Fletcher 1940).
- Sounds mask other sounds of a higher frequency better than they mask sounds of a lower frequency (see section 2.2.6 on the upward spread of masking) (Wegel and Lane 1924).
- Masking thresholds usually relate to the bandwidth of a noise masker in a way which would be expected if the auditory system contains a bank of band pass filters (Hellman 1972).

These results are all explained by the power spectrum model of masking and by the
shapes of overlapping auditory filters. Many other trends, however, have emerged which are not so easily explained. Hellman (1972), for example, identified an asymmetry between the effectiveness of noise and tone as a masker, noting that noise masks tone more effectively than tone masks noise (even when the noise is restricted to a single critical band). Hellman went on to show that while the relative effectiveness of a noise masker is constant for various stimulus levels at a fixed SNR, this was not the case for a tone masker which decreases in effectiveness as the levels of signal and masker increase. This contravenes the assumption made by the power spectrum model that SNR is equal to a constant of signal detection.

This phenomenon may be somewhat explained by the relationship between the width of auditory filters and stimulus intensity levels. As stimulus intensity increases the shape of auditory filters broadens with the low frequency slope becoming more shallow. For an auditory filter centred on the signal tone the broader shape will allow more noise to be passed, thus increasing the excitation caused by the noise, without affecting the perception of the tone.

Hellman (1972) also noted a difference in how tone is perceived as a masker: the region of uncertainty in the vicinity of a masking threshold is much greater when tone is a masker than when noise is a masker.

The power spectrum model of masking is based on the long-time average spectrum of a signal and thus ignores temporal variations. Real signals, however, often include transients which may be more easily detectable. Hirsh et al. (1950) investigated the masking of transients and found a number of trends:

- Tones are generally poor at masking transients, while bands of noise are more effective.
- The masking effectiveness of a band of noise, on a transient, is inversely proportional to its centre frequency.
- The frequency at which masking is most effective is inversely proportional to the level of the transient.

Another masking trend not described by the power spectrum model can be found when signal tones coincide with ‘spectral edges’. Spectral edges are sharp changes in the spectrum of the masker, such as the edge of an NBN. Margolis et al. (1981) found an increase in masking thresholds when spectral edges are in the vicinity of a tone.

Although work such as (Hawkins and Stevens 1950) and (Schafer et al. 1950) has allowed masking threshold curves for various combinations of tone and noise to be well agreed upon, many of the previously discussed masking trends are not predicted or explained by the power spectrum model, and thus a more complete model of masking is likely to give better predictions of masking thresholds.
Masking of unwanted signals within a sound zone is highly desirable, but it is likely that there will be many practical applications where it is not possible. This is because real signals are complex and it is unlikely that the spectrotemporal characteristics of both the signal and masker will be identical or sufficiently similar to allow simultaneous masking to persist throughout the duration of a listening experience. For example, the nature of speech is such that the quiet gaps between sentences, words and some phonemes will render it unable to mask a continuous noise signal.

If there are large differences in time-varying frequency spectra it is possible that simultaneous masking will not fully mask the signal, even if the temporal characteristics of the masker and maskee signals are identical. This is because significant disparities in frequency spectra will result in unmasked components of the signal. Figure 2.14 shows an example in which a combination of speech and violin signals are likely to result in the harmonic peaks of the violin remaining unmasked.

It is also possible that room reverberation could diminish masking effects by changing the temporal properties of the masker and maskee. If the signal sources are in two different locations within a single acoustic space they may not have identical reverb tails (Everest 2000), which may results in the listener perceiving non-simultaneous auditory events.

The power spectrum model is a convenient starting point for modelling masking thresholds because of its relative simplicity, but it is founded on the results of tone in noise masking experiments and thus assumes that the level and frequency content of the signal and masker do not significantly vary with time. These assumptions are rather untenable for ecologically valid listening situations, however, and section 2.3 discusses the temporality of masking phenomenon.

Figure 2.14: Spectral analysis of a sustained violin note (in blue) and male speech (in green).
2.3 Temporal masking

In the previous section it has been assumed that the masker and signal are presented to the listener simultaneously. This may be referred to as simultaneous masking (see section 2.3.1). In some cases masking occurs when the signal is presented after or before the masker, and these cases of non-simultaneous masking are discussed in sections 2.3.2 and 2.3.3 respectively.

2.3.1 Simultaneous masking

Simultaneous masking is the psychoacoustic phenomenon where a maskee is rendered inaudible by the presence of a masker presented at the same time. Simultaneous masking is the most commonly encountered form of masking and nearly all listeners will have experienced this phenomenon while trying to hear speech in a noisy environment. Simultaneous masking, together with forward and backwards masking (described in sections 2.3.2 and 2.3.3 respectively), comprises of the most significant and ubiquitous aspects of masking phenomena (although some special cases of unmasking are also worth attention). All masking described in section 2.2 was assumed to be simultaneous in nature, however this is only one aspect of a broad range range of masking phenomena, collectively referred to as ‘temporal masking’.

2.3.2 Forward masking

Forward masking (sometimes known as post-masking) is the phenomenon in which the presence of a masker renders a signal inaudible even though this signal (the maskee) is presented after the masker presentation has ceased. The time difference between the masker offset and the maskee onset is a critical variable because the masking threshold decays over time. The rate of this decay is proportional to the presentation level of the masker. The result is that the amount of forward masking is always zero after a period of around 100-200ms (the precise value varies amongst listeners). Figure 2.15 shows an example of a signal rendered inaudible by forward masking.

Generally, the amount of forward masking is a function of the logarithm of the delay between the end of the masker and the start of the signal. There is a conditional relationship, however, between the masking threshold and both the masker level and the signal delay. While an increase in masking threshold is observed for increased masker levels, this increase is minimised for greater delays between signal and masker (Moore 2004).

Forward masking phenomena are sometimes described by a growth of masking function. This function describes a non-linear aspect of masking: the change in masking threshold
caused by a change in masker level. For example, a growth of masking function of 1 describes a situation where the masking threshold occurs at a fixed SNR. Like simultaneous masking, noise is a more effective forward masker than tone, and has a growth of masking function with a gradient slightly less than 1 while the sine tone growth of masking function has a gradient closer to 0.5. It should also be noted that forward masking is affected by central processes: when noise masks tone at one ear but noise is also presented contralaterally the growth of masking function is diminished (Neff and Jesteadt 1983).

Another important variable in forward masking is the duration of the masker. For masker durations less than 50 ms the amount of forward masking tends to decrease while, for masker durations greater than 50 ms the amount of forward masking is constant. This is consistent with the hypothesis that loudness takes time to develop, and that masking is a result of masker loudness; masker stimuli of duration less than 50 ms thus produce less forward masking because there has been insufficient time for the loudness of the masker to plateau (Moore 2004).

The forward masking threshold, therefore, is a function of offset-onset delay, masker level, masker duration, and both signal and masker type spectral content.

Further complexity emerges when considering the combination of many forward maskers. Under this circumstance there is usually an increase in the masking threshold although yet in some cases there have been surprising results. Weber (1984) found that adding tone to a noise masker decreased the forward masking on a signal tone. Zwicker and Fastl (1990) suggest that a similar phenomenon can occur for narrow-band maskers.

Figure 2.15: A masker is presented and then terminated before the onset of the signal. If the delay between the two is sufficiently short the signal may not be detected (as in this case). The curved line shows how the masking threshold varies over time after the masker offset.
To account for these masking trends, two hypotheses explaining the physiological mechanism of forward masking are generally presented: persistence and adaptation. The persistence hypothesis of forward masking suggests that neural activity evoked by the masker persists (after masker offset) somewhere in the auditory pathway beyond the auditory nerve (Oxenham 2001). The adaptation model suggests that the masker produces short term fatigue on the auditory nerve itself, such that it is not stimulated by the proceeding signal (Jesteadt et al. 1982).

Oxenham (2001) showed that the difference in predictions by persistence and adaptation models is relatively small, and thus for the practical purposes of the modelling of auditory perception for the prediction of masking the underlying physiological mechanism may be ignored.

2.3.3 Backward masking & excess masking

Backward masking (sometimes known as pre-masking) is the phenomenon where a listener is unable to perceive a sound which is presented just prior to a masker. Pastore et al. (1980) suggest that temporal uncertainty plays an important role (i.e. the listener does not know which (temporal) part of the signal to listen to until after the distracting masker is presented). As a result backward masking usually occurs over a very short period of time (less than 20ms).

Although the severity of backward masking can sometimes be greater than that of forward masking, timing cues can significantly diminish the effect (while they have no effect on forward masking) (Pastore et al. 1980). In addition, the level of backward masking depends considerably on how much listening practice the subjects have, and may be entirely absent for trained listeners (Moore 2004). Backward masking therefore seems to be related to listener uncertainty and the change due to listener training indicates that backwards masking is likely due to a process in the auditory pathway beyond the auditory nerve.

A related phenomenon occurs when a masker is presented both just prior to and just after a signal (i.e. when forward and backward masking phenomena overlap). In some cases the resultant masking threshold is greater than would be predicted by a simple energy summation (Oxenham and Moore 1994), while in other cases it is less (Pastore et al. 1980). The unexplained difference is usually referred to as ‘excess masking’.

2.4 Partial masking

In previous sections masking has been discussed in terms of the dichotomy of audibility, i.e. either a signal is masked (and thus inaudible), or it is not. There is, however, another phenomenon known as ‘partial masking’ wherein a signal is not rendered
completely inaudible but its loudness is decreased due to the presence of an interfering signal.

In this way, masking can more meaningfully be considered to be a specific case of perception where the loudness of a signal is reduced to zero by the presence of another signal. This frame of reference is equivalent to the case of partial masking where the additional presence of a signal increases the overall level of presented stimuli by less than the Just Noticeable Difference (JND). The resultant effect is that the additional signal is perceptually masked, although in absolute terms this may be considered an extreme case of partial masking. The specific case of absolute masking, while conceptually notable, is therefore simply one extreme in a range of interference scenarios.

2.4.1 Evidence for partial masking

Scharf (1964) described an experiment in which listeners adjusted the level of a tone until they considered it to be of equal loudness to another tone presented simultaneously with white noise. The results indicated that the presence of the white noise decreased the loudness of the simultaneously presented tone, relative to a tone of equal level presented in isolation. Gockel et al. (2003) conducted a similar experiment using tone complexes to further extend the work. The results are shown in fig. 2.16, and the differing slopes of the growth of masking functions reveal that the asymmetry of masking between noise and tones is also present for partial masking.
As shown in fig. 2.16, whether the target was white noise or a tone complex, a higher target level was required for the target presented with the interferer than for the target presented in isolation for equal loudness. When the target level is sufficiently high, the partial masking effect of the interferer becomes negligible and the loudnesses become equivalent regardless of the presence of the interferer. It should be noted that for partial masking, as for complete auditory masking, there is an asymmetry between the effectiveness of tone and noise (i.e. noise is the more effective masker).

2.4.2 Spectral partial masking

When a tone is presented to a listener simultaneously with high pass filtered white noise, the perceived loudness of the tone varies as a function of its frequency separation from the high pass cut-off. The relationship is non-linear, and is illustrated in fig. 2.17. This effect is likely to impact the loudness of a target within a sound zone because an interferer with frequency components less than 300 Hz away from the the components of the target will cause a noticeable decrease in loudness. Additionally this effect increases the complexity of loudness prediction because quite a large frequency separation is required to give a loudness value which corresponds to a simple addition of the loudness of each tone. Thus in many cases the overall loudness of a pair of signals cannot be
2.4.3 Temporal partial masking

In much the same manner as backward masking experiments (discussed in section 2.3.3), a partial backward masking experiment is described in Zwicker and Fastl (1990). The results indicate that the level of partial masking varies as a function of the time difference between the offset of the tone and the onset of the noise, as shown in fig. 2.18. This phenomenon could significantly affect the loudness of a programme in a sound zone, because the spill from another programme is likely to have an uncorrelated temporal pattern which causes temporal differences between offset and onset less than 100 ms. As a result the loudness of a target signal in a sound zone may be decreased by the presence of an interfering signal which does not fully mask it.
2.4.4 Discussion

While masking can be considered in terms of the dichotomy of audibility, a more complete understanding concerns the effect of the presence of an interferer on the loudness of a target. As such masking can more meaningfully be considered to be a specific case of partial masking in which the loudness of the masker is zero, or negligibly small.

While these results are fascinating in isolation, they also apply in a very specific way to the sound zone problem. The evaluation model is required to predict the overall quality of a sound zoning system. While an important part of this task is determining whether an interferer will be rendered inaudible by the target, an inaudible interferer is simply an interferer with zero effective loudness. Thus an equally important task lies in determining the effective loudness of the interferer, which is partially determined by the loudness of a target.

2.5 Binaural unmasking

Binaural unmasking is a phenomenon where the masking threshold of a signal is reduced by binaural listening (relative to monaural listening). The difference in masking threshold, the Binaural Masking Level Difference (BMLD), is generally attributed to a spatial separation of signal and masker, since diotic listening offers no advantage when there are no Interaural Phase Differences (IPDs) or Interaural Level Differences (ILDs).

When a listener attempts to identify a signal within a noisy environment the level of success is determined both by the SNR at the ears and by the differences in level and phase of the target and the interferer between the two ears. Cherry (1953) is credited with first demonstrating this phenomena by presenting listeners with dichotic recordings of speech produced by the same speaker. Figure 2.19 shows scenarios in which a signal that is masked by noise becomes audible when the IPD or ILD of the signal is very different from the IPD or ILD of the noise.

Egan (1965) demonstrated binaural unmasking using an experiment in which subjects were presented with a mixture of noise and signal monaurally. The SNR was then adjusted until the signal was just masked by the noise. When a correlated noise signal was applied contralaterally the listener once again perceived the signal even though the SNR at the auditory filter passing the signal had not changed, and the overall SNR had decreased. The masking threshold must, therefore, have been reduced by a central process, i.e. comparing the signals presented at the left and right ears in order to use the IPD and ILD for signal detection.

It should be noted that binaural unmasking may still occur in cases where contradictory or confusing spatial cues are given. This is due to the surprising finding that the
Chapter 2: Auditory Masking

Figure 2.19: Binaural unmasking examples. In a) the signal is presented to each ear in phase accompanied by noise which masks it. In b) the phase of the signal at one ear is reversed. The resulting IPD is different to the IPD of the masker, causing the signal to become audible. In c) the signal and masker are presented monaurally, and the signal is masked. In d) correlated noise is presented contralaterally and the resultant difference between the ILD of the masker and the ILD of the signal provides a cue to detection. Adapted from (Moore 2004).

The greatest binaural masking level difference occurs under antiphase conditions where, “sound images occur diffusely within the head. Thus escape from masking and lateralization/localization seem, to some extent, to be separate capacities” (Moore 2004). Some methods of sound zoning, such as contrast control, can create conflicting phase information at the listening position (Jacobsen et al. 2011) and so might produce scenarios with conflicting IPD and ILD cues. Figure 2.20 shows the phase characteristics produced using two different sound zoning methods side by side. While it is possible for such conflicting cues to unmask an interfering programme, it should also be noted that if the IPD varies over time, however, the resulting cues to binaural unmasking might be negligible.

Hirsh (1948) noted that BMLD for low frequency signals tends to be around 15 dB for broadband maskers when the noise is presented diotically in phase and the signal is presented in antiphase. As the frequency of the signal increased the BMLD decreased, thus it appears that high frequency signals are less prone to binaural unmasking caused by differences between the IPD of the signal and the IPD of the masker.

Zurek and Durlach (1987) further identified that BMLD tends to increase significantly when the masker bandwidth is less than three times the critical bandwidth. This trend can also be found in the data reported by Van de Par and Kohlrausch (1999), who conducted a similar experiment. Subjects were presented with either signal and noise in phase at both ears or with signal in phase but noise in antiphase. Repeating the experiment for a range of signal frequencies and masker bandwidths revealed more of
Chapter 2: Auditory Masking

Figure 2.20: A contrast control method (left) and a wavefield synthesis method (right) of sound zoning. The two methods produce very different phase characteristics. Reproduced with permission from (Olsen and Møller 2011).

Figure 2.21: The graph on the left indicates the masking thresholds found for listening tests with signal and noise both on the frontal axis ($S_0N_0$) using various signal frequencies and a range of masker bandwidths, while the graph on the right indicates the masking thresholds found for the corresponding listening tests with noise presented in antiphase ($S_0N_{\pi}$). Adapted from (Van de Par and Kohlrausch 1999).

the nature of the binaural unmasking effect. When the noise was presented in antiphase, masking thresholds across frequencies did not converge above 1 ERB. Furthermore the masking thresholds did not begin decreasing until around 2–4 ERBs, in some cases increasing slightly at 1 ERB. These results are shown in fig. 2.21. The data seems to confirm that the BMLD is greater for lower signal frequencies, and further indicates that a change in masker bandwidth affects the BMLD in a way which depends upon the signal frequency. Van de Par and Kohlrausch (1999) found similar results when the noise was presented in phase and the signal was presented in antiphase.
2.5.1 Impact of binaural unmasking

As in (Hirsh 1948), Durka (1963) suggested that BMLDs could be as high as 15 dB in antiphase conditions. Breebaart et al. (2001) later suggested that BMLD could even reach as high as 25 dB. These experiments, however, were conducted using simple, controlled stimuli. Bronkhorst (2000) collated a number of studies on binaural unmasking and found that for speech masked by noise BMLD did not exceed 12 dB (see fig. 2.22). Figure 2.22 shows relatively good agreement between a number of binaural unmasking studies which suggest that there is approximately 7-10 dB of BMLD for 60-120 degrees azimuth noise for a frontal speaker.

2.5.2 Summary

Binaural unmasking phenomena are consistent and can have a significant effect on masking thresholds. The effect seems to be strongest for lower frequencies. Whether
binaural unmasking would occur, however, within a sound zoning context depends heavily on the ILD and IPD which, in turn, depend upon the sound zoning method. Some methods do not control phase information and therefore may produce arbitrary IPDs which could be utilised as cues to detect the interferer programme.

2.6 Informational masking

Producing a strict definition of Informational Masking (IM) is a much more complex task than it may seem. Part of the problem is that different authors have used the term to refer to different phenomena (or at least to refer to the results of experiments for which it cannot be guaranteed that the same phenomenon is present). Neff and Green (1987) used IM to describe a phenomenon in which the presence of multitone maskers produced a much greater masking level than is predicted by the power spectrum model of masking. They further demonstrated that this additional masking was largely caused by the listener’s uncertainty of the spectral content of the signal and masker. These results have been replicated using the same terminology in (Durlach et al. 2005) and (Leibold et al. 2010).

Some authors such as Hawley et al. (2004) and Cooke et al. (2008) have used IM to refer to the increase in masking level which occurs when a speech target is masked by speech maskers, a phenomenon which Carhart et al. (1969) described as a particular case of additive masking and referred to as 'perceptual masking'. This use of IM is often based on work such as that of Brungart et al. (2006) which claims that the Energetic Masking (EM) component of speech-on-speech masking is generally very small (where EM comprises all those types of masking described in previous sections). This can be problematic, depending on the aim of the experiment, because the EM component is assumed to be small but is often not calculated and subtracted from the result. Additionally the EM component will not be constant if the number of speech maskers is varied. In general EM, as a confounding factor in speech experiments, should be either controlled or monitored. Where neither of these are carried out experimental results can only be used to illustrate the overall masking of a scenario, and not to discern the underlying processes which cause the masking. Such results may therefore still be of some use in this work which seeks to predict masking thresholds.

In some cases IM is used to refer to masking which occurs when no EM could have occurred (Hawley et al. 2004), where EM is defined as masking caused by the auditory periphery. Such a definition of IM is particularly weak because it is a definition of exclusion and therefore does not actually describe the phenomenon being measured. Work which uses IM in this way does not comment on whether results are indicative of a single process or a group of processes. Furthermore, Durlach (2006) points out that the use of IM to refer to masking which cannot be energetic is equivalent to relating IM
to central masking, and EM to peripheral masking. Making such a comparison shows that using IM and EM in this way is meaningless because “all masking is energetic masking if examined at a sufficiently high level” (Durlach 2006).

Yost (2006) proposes that three areas should be clearly distinguished: “masking in a detection task, interference in a discrimination task, and competition in a recognition/identification task”. Watson (2006), however, counters the suggestion that IM should be split into three terms by pointing out that results for the three are highly correlated and that recognition implies discrimination, which in turn implies detection. Brungart et al. (2006) point out that the division of masking into IM and EM is an oversimplification, arguing that in cases of signal distortion caused by additional masking energy there may be insufficient energy to prevent signal detection, but sufficient distortion to prevent signal recognition. Such a case could reasonably be argued to be the result of either IM or EM.

Since the aim of this work is to predict masking within sound zones, rather than to identify distinct peripheral or cognitive processes involved in masking, a solution to the problematic terminology is not necessary. Some specific cases of IM are introduced in this section, however, and they are named separately in order to avoid confusion. EM will be used to refer to masking effects described in previous sections, however the term IM will be altogether avoided by use of distinct names attributed to individual phenomena.

2.6.1 Uncertainty masking

Uncertainty, in the context in which a target tone is presented, can significantly degrade detection (Watson et al. 1976). Watson termed this phenomenon “informational masking” to distinguish it from masking produced by energy at the signal frequency, but Neff and Green (1987) referred to this masking as that caused by ‘spectral uncertainty’. Since the uncertainty is not always spectral, this work will use the term ‘uncertainty masking’ in order to distinguish it from IM which is used in the context of speech perception.

An increase in masking threshold is observed when the listener is naïve to the frequency spectrum of either the signal or masker. A rapidly changing frequency spectrum maintains this naïveté and thus increases the spectral uncertainty of the listener. Neff and Green (1987) conducted a study in which subjects were presented with a tone signal masked by a multitone complex comprised of randomly selected tones of frequencies below 5kHz. The phase and level of each tone within the multitone complex was randomly selected, but the overall SPL was kept constant. The variable of interest was the number of components in the multitone complex, which was varied from 1 to 100. A reference case of broadband noise (low pass filtered at 5kHz) was also tested.
Chapter 2: Auditory Masking

Figure 2.23: Multitone IM for 1 kHz tone (left) and 4 kHz tone (right). The dashed line indicates the equivalent masking threshold for broadband noise. Adapted from (Neff and Green 1987).

It would be reasonable to postulate that the masking components would produce a masking threshold which varies with their frequency and level. As the number of components increases the multitone complex should be perceived more like broadband noise because many tones of random level and phase will be present within a single critical band. The masking threshold should therefore tend towards the threshold found for broadband noise at the same SPL and the variability of results should diminish.

The results from Neff and Green’s (1987) study for signal tones of 1kHz and 4kHz can be seen in fig. 2.23. As predicted the standard deviation of results appears to be inversely proportional to the number of masking components. The masking level appears to be tending towards that of broadband noise for 100 components however subjects reported that the 100 component tone complex did not sound very similar to noise, and thus it seems a much greater number of components is needed for the masker to be perceptually indistinguishable from noise.

The most striking result is that as much as 55 dB of masking was observed in some cases where relatively few masking components were used. In these cases very little of the masking signal would be found within the critical band of the signal. This evidence indicates that a mechanism of masking is at work which the power spectrum model does not explain, because the SNR at the output of the auditory filter centred on the signal should be very high in these cases. Another interesting result is that the masking levels for more than four components are considerably above the masking level found for broadband noise. This additional masking also cannot be explained by the power spectrum model because the SNR at the output of the auditory filter must be no less than it would be for broadband noise in the vast majority of trials. Neff and Green (1987) indicated that these results were examples of IM. In this case the term indicates
that listeners are sometimes unable to concentrate on a single auditory filter even if it advantageous to do so.

Cases where the masker was changed on every interval were then considered and these were compared with the results from cases where pairs of intervals with the same masker were studied. The results suggested that the subjects experienced significantly greater masking when the masker was randomised for each interval. This implies that subjects make a comparison between the spectral character of the audio presented on repeated trials and then perform a subtraction to discern which of the two cases contained an additional signal. This is conceptually similar to Durlach’s (1963) Equalisation Cancellation (EC) theory of BMLD (which suggests that the frequency spectrum of the signal presented at the left and right ear are aligned and subtracted) although over a spectrotemporal, rather than just spectral, scale. When the masker is randomised for each test case this comparison is either not carried out or gives misleading results and the signal detection therefore suffers.

It could be argued that while the probability of masking tones falling within the critical band of the signal is low, it is still likely that there will be some contribution to the determined average masking level by EM. In order to discern what contribution could have been made by EM Neff and Callaghan (1988) conducted a similar study in which multitone components were never produced within one critical band around the signal (estimated at 160Hz for a 1kHz signal). According to the power spectrum model of masking this new methodology should eliminate all simultaneous masking. The results of the study indicated about 10 dB of masking release when the critical band regions were protected, but 30-40 dB of masking was still present.

In order to confirm the veracity of this work Neff and Callaghan analysed their results to check whether listeners had improved at the signal detection task over a large number of trials. Figure 2.24 shows the learning curves for four of the listeners over 1800 trials. Subjects 2, 3, and 4, did not improve in the signal detection task over a very large number of repetitions, but listener 1 improved considerably over the first 600 trials. Learning to diminish informational masking, therefore, appears to be possible for some listeners but not all. It should also be noted that listener 2 was a musician and did not appear to have considerable unmasking advantage over the non-musician listeners.

These results are remarkable because the power spectrum model of masking predicts no masking at all for cases where measurements indicate between 30 and 40 dB of masking. If the power spectrum model of masking correctly describes the energetic masking produced in the auditory periphery then masking must be a phenomenon with multiple mechanisms, at least one of which must be non-EM but also pre-cognitive processing (because the unwanted spectral components could not be ignored by means of attentiveness or training).

In related studies, Spiegel (1981) found that both signal and masker uncertainty
contribute to uncertainty masking, but that masker uncertainty was responsible for the greater proportion of the masking regardless of the number of components in the masker. Durlach et al. (2005) found that there could be up to 50 dB of difference between the informational masking experienced by different listeners presented with the same stimulus and masker (although 10 dB of difference was much more common). Contrary to Neff and Callaghan, Oxenham et al. (2003) found that musicians were significantly better at detecting a signal in uncertainty tasks than non-musicians. In this experiment the non-musicians experienced a mean of 25 dB of uncertainty masking while the musicians experienced only 10 dB.

It should also be noted that the extent of uncertainty masking varies greatly both among specific listener groups and with individual differences. Wightman et al. (2003) found that children experienced greater uncertainty masking of pure-tone signals than adults under the same acoustic conditions (which might be explained by their greater stimulus naivety), and Oh and Lutfi (1998) found large individual differences in masking functions amongst the adult listeners tested.

Work by Spiegel (1981) indicates that familiar sounds are unlikely to cause uncertainty masking. They found that when listeners were given time to learn various multitone complex maskers the listeners were less prone to the uncertainty of those maskers, even when the maskers were randomly presented. Figure 2.25 shows three conditions in which listeners had varying levels of uncertainty of the masker.

While uncertainty seems to be inversely proportional to the listener’s success rate, it is important to note that the subjects tended to improve in all conditions. Spiegel and Watson (1981) caution that learning takes considerably longer for complex maskers. This is, however, strong evidence that masking caused by spectral uncertainty will be
reduced for learned sounds such as familiar musical instruments and voices.

Oh and Lutfi (1998) found a strong correlation between listeners’ ability to recognise a sound and the decrease in uncertainty masking caused by that sound. They could not, however, rule out the confounding factor that most of the sounds which were easily recognised in their experiment had a harmonic structure; thus it is unclear whether the reduction in uncertainty masking is caused by the recognition of a sound or by its harmonic structure. Kidd et al. (2002) found that uncertainty masking was significantly increased when the signal and masker had spectrotemporal similarities (such as the way that the frequencies of masking components changed over time), which emphasises that the nature of this masking is indeed based on uncertainty. In an earlier paper, Kidd et al. (1994) had found that uncertainty masking could be reduced by three methods: binaural unmasking (because interaural cues allow the sources to be perceived as occupying different spatial locations), a difference in spectral cues (e.g. such as when the frequency spectrum of one signal is varying and the other is not), and a difference in temporal cues (e.g. such as when a signal is recurrent while the masker is static).

On the strength of this evidence, therefore, it seems unlikely that uncertainty masking will be very important for sound zoning scenarios featuring ecologically valid stimuli. This is because such stimuli will be almost universally harmonic in structure, and will usually be recognisable to the listener; there may also be binaural or temporal cues available to the listener to diminish uncertainty.

Other types of uncertainty may also affect masking. Shi and Law (2010) found that speech recognition was masked more by serial and jazz music than by classical music.
2.6.2 Summary

In summary it appears that uncertainty, in many forms, can affect the masking threshold of a listener. It is not entirely clear, however, how the uncertainty of a listener should be predicted, since this would require knowledge of many aspects of the listener’s musical experience, and some cases of spectral uncertainty learning, for instance, still cannot be accounted for.

Although in theory up to 70 dB of masking can occur by listener naivety it can be generally assumed that the overwhelming majority of stimuli will have a harmonic structure or will be at least broadly familiar sounds to the listener (even when listening to new music it is likely that the use of instruments and harmony will not be entirely unfamiliar). Therefore spectral uncertainty masking is unlikely to have much effect in a practical sound zoning system. While it cannot be assumed that signal and masker uncertainty will not be important aspects of quality in sound zones, they can be considered relatively unimportant for the prediction of masking.

2.7 Comodulation masking release

Comodulation Masking Release (CMR) is the drop in masking threshold which occurs when separate frequency components of a masker are amplitude modulated together.

The first study of this phenomenon was conducted by Hall et al. (1984). A standard band widening experiment was conducted featuring a 400 ms 1 kHz tone with a masking noise centred on the signal. As predicted by Fletcher’s (1940) work on critical bands, once a certain bandwidth of noise masker was reached no further influence on masking threshold was observed. When the masking noise was temporally modulated (in this case by multiplying 0-10 kHz noise by a 0-50 Hz band of noise) a decreased masking threshold was observed as the bandwidth of the noise was increased beyond the critical bandwidth (see fig. 2.26).

The explanation for this result is that some kind of modulation detection must be happening both inside and outside of the auditory filter that is centred on the signal. The comparison of the modulation detection occurring in the auditory filters not centred on the target signal with the output of the auditory filter centred on the target signal allows some part of the auditory processing system to unmask the signal.

Hall et al. subsequently conducted another experiment, this time using two bands of noise: one centred on the signal at 1 kHz, and another centred on 900 Hz. The masking threshold of the signal was observed for bandwidths of noise set to 100 Hz, 300 Hz, 500
Hz and 700 Hz under two test conditions: two incoherent bands of noise, and one in which the two bands were generated by multiplying a 300-400 Hz band of noise by a tone complex and applying low-pass filtering such that the temporal envelopes of the two noise bands were coherent. The results are shown in fig. 2.27.

The results show a CMR of around 6 dB for 300, 500 and 700 Hz, but no CMR for 100 Hz. This indicates a step change as the masking bandwidth exceeds the critical bandwidth, but no further masking release beyond that point. Hall et al. (1984) consider that this is probably because the relative coherence between the noise bands was so great that the addition of further flanking bands of noise does not help to predict the noise fluctuations any more successfully.

A third experiment conducted by Hall et al. (1984) investigated whether the distance between the signal frequency and the flanking band centre frequency had a significant effect on the CMR. Although they only tested up to 300 Hz in either direction, they found no evidence that the frequency separation had a significant effect on the extent of the CMR.

Dau et al. (2009) showed that CMR would not be present for wideband flanking bands if a series of gated flanking bands were added after the signal. They interpreted this result as an indication that the gated ‘post-cursor’ bands were perceptually components of the same auditory object as the comodulated flanking bands. As such the entire perceptual auditory object did not have flanking bands which were consistently comodulated, so no additional CMR could be achieved by the listener’s auditory processing system. A
Chapter 2: Auditory Masking

Figure 2.27: A two noise band masking experiment conducted by Hall et al. (1984) in which the circles indicate the case in which the noise bands were uncorrelated, and squares indicate case in which the noise bands were correlated. Adapted from (Hall et al. 1984).

model of CMR must therefore be able to provide some, at least rudimentary, prediction of the perception of auditory objects.

CMR is further evidence that the power spectrum model of masking describes only masking caused by the auditory periphery, and so two proposed mechanisms have been described to explain this collection of results: ‘spectral subtraction’ and ‘dip listening’.

Spectral subtraction suggests that a listener actively makes comparisons of the modulation of signals at the output of different auditory filters. This is an attractive theory because it is partially corroborated by Green (1988) who suggests that listeners are able to compare the outputs of different auditory filters to aid in signal detection.

Dip listening, by contrast, suggests that a listener uses modulation information gathered in the off-frequency filters to pay more attention to the on-frequency auditory filter output at time intervals where the noise is at a minima in the modulation pattern. It is likely that both suggested mechanisms play a role in CMR, and should therefore be considered in a hearing model which predicts this phenomenon.

It seems probable that natural sound sources may have a broad frequency spectrum with coherent amplitude modulations caused by physical resonances. Spatially separated sources might also produce useful CMR cues depending on the reverb characteristics of the room. Even if CMR does not occur in natural sound sources it is likely to occur in some types of synthesised music where amplitude modulation is used as a musical effect. The presence of CMR, therefore, cannot be ruled out for sound zone scenarios.

A simple way to detect CMR would be to compare frequency bands across many time windows to search for repetitive modulation. If similar modulations are detected
in separate frequency bands an adjustment can be made to the predicted masking threshold to account for CMR. A model devised by Unoki and Akagi (1997) does almost exactly this. Their model also performs a power spectrum model analysis and subsequently selects the output with the lowest masking threshold to predict human performance. Their results indicated that for sine tones CMR usually produced no more than 8 dB change in masking level.

### 2.8 Summary and conclusions

This chapter outlined the role that masking experiments have played in describing the operation of the human auditory system. Various masking phenomena were introduced and their relevances to sound zoning scenarios were discussed. Two questions were posed at the start of this chapter: “What are the factors which determine whether an auditory stimulus will be masked by the presence of a second stimulus?”, and “What is the relevance and importance of each factor?”

The literature regarding these two questions were investigated in this chapter, and Table 2.1 outlines some of the key features of the various masking phenomena discussed.

By their nature, the cases of simultaneous, forwards and backwards masking have a severity which is heavily dependent on SNR. The result of these masking scenarios may not be binary, because of partial masking, however one possible outcome is the complete masking of a signal. It is not meaningful, therefore, to attach a masking level to these phenomena which indicates severity because the severity varies so widely, and will sometimes be absolute. Combined, these three phenomena constitute what may more broadly be referred to as auditory masking, with the remaining phenomena changing the effect size. Nonetheless, the effects of backwards masking are sufficiently inconsistent.

#### Table 2.1: An overview of the masking phenomena previously discussed. Where a value is given for ‘Severity’ it indicates the approximate greatest value found in the literature. Phenomena below the double line are specific cases which modify the masking caused by those above the double line.

<table>
<thead>
<tr>
<th>Masking</th>
<th>Severity</th>
<th>Variability</th>
<th>Complexity</th>
<th>Relevance/Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simultaneous</td>
<td>-</td>
<td>Consistent</td>
<td>Simple</td>
<td>High</td>
</tr>
<tr>
<td>Forward</td>
<td>-</td>
<td>Consistent</td>
<td>Simple</td>
<td>High</td>
</tr>
<tr>
<td>Backward</td>
<td>-</td>
<td>Inconsistent</td>
<td>Complex</td>
<td>Low</td>
</tr>
<tr>
<td>Excess</td>
<td>&lt; 10 dB</td>
<td>Inconsistent</td>
<td>Complex</td>
<td>Low</td>
</tr>
<tr>
<td>Binaural</td>
<td>&lt; 25 dB</td>
<td>Consistent</td>
<td>Simple</td>
<td>Mid</td>
</tr>
<tr>
<td>CMR</td>
<td>&lt; 8 dB</td>
<td>Consistent</td>
<td>Simple</td>
<td>Low</td>
</tr>
<tr>
<td>IM$_{SU}$</td>
<td>&lt; 8 dB</td>
<td>Consistent</td>
<td>Complex</td>
<td>Low</td>
</tr>
</tbody>
</table>
that they have are considered to be of low importance to the sound zoning problem (compared with simultaneous and forward masking). Precisely the same objection applies to excess masking, which is thus also considered of low importance.

The effect of binaural unmasking is consistent and can be rather pronounced, however whether or not it is likely to occur within sound zone scenarios depends upon whether the sound zoning method is likely to produce perceptual cues for spatial separation. For this reason, therefore, it is considered to be of secondary importance.

CMR and spectral uncertainty masking are both considered to be of low importance because the impact they will have upon auditory masking is likely to be relatively small in listening environments rich with auditory cues. Additionally, spectral uncertainty masking is considered extremely unlikely to occur within ecologically valid listening scenarios.

With a range of masking phenomena considered, it now becomes necessary to consider the relationship between masking and acceptability. In the next chapter experiments are described investigating this relationship for auditory interference scenarios.
Chapter 3
Masking and Acceptability Experiment

In the previous chapter the various masking phenomena were investigated, and in chapter 1 it was noted that this would be instrumental in defining the boundaries of audibility, and therefore acceptability. This is valuable for mapping the range of possible acceptability scores, and for discerning whether the acceptability might immediately be known in some applications (due to being outside this range). While the range of acceptability is therefore determined by the audibility of the target and interferer programmes, it is likely that acceptability scores mainly vary over a smaller range of SNRs, and the extent of this range is of interest. A research question is therefore posed: "What is the range of SNRs over which acceptability primarily varies?"

Furthermore, it is useful to consider whether there may be a deeper relationship between audibility and acceptability than simply defining boundaries. If so, it may be possible to predict acceptability using an understanding of masking. Thus a research question is posed: "Is there a relationship between masking and acceptability?"

A pair of experiments were conducted to answer these questions; the first for masking data, and the second for acceptability data. Since the two experiments were identical in all but the task posed to the subjects, the method and stimuli are described once in section 3.1. The experiment results are discussed in section 3.2. Finally, section 3.3 considers the use of masking thresholds to predict acceptability thresholds.

3.1 Experiment design

To investigate the relationship between masking and acceptability it was necessary to obtain a set of masking and acceptability thresholds for the same set of listening scenarios. A pair of experiments were conducted to gather masking thresholds and acceptability thresholds for a range of ecologically valid programmes. The experimental methodology and stimulus details are described below.

3.1.1 Methodology

The subjects used an unmarked rotary fader to interact with a computer. The computer replayed one audio programme (the target) via a Genelec 8020A loudspeaker positioned
directly in front at a distance of 1.85 m and a height of 0.78 m, and a different audio programme (the interferer) via a Genelec 1032 loudspeaker positioned directly in front at a distance of 2 m and a height of 1.04 m. Thus both loudspeakers were positioned at approximately head height but with minimal occlusion. A hexagonal array of Genelec 8020As, positioned at a height 78 cm from the floor, was used to reproduce road noise on half of all trials. This loudspeaker arrangement was selected as a simple way to approximate the envelopment experienced when in an automobile. The listening position was near the centre of a room meeting the specifications of ITU-R BS.1116 (1997). Figure 3.1 shows the listening experiment layout.

In the masking experiment the subjects were instructed as follows:

"You will be presented with two audio programmes; you can control the level of one of the programmes. The controllable programme will start at a level where it is audible. Using the rotary folder, please adjust the level of the controllable programme to the point where it is just inaudible."

In the acceptability experiment the instruction differed slightly; they were asked to:

"Imagine you are relaxing (at home or in the car) by listening to music or
sports commentary. With this in mind, adjust the level of the interfering audio programme until you find the listening scenario acceptable”.

The stimuli looped indefinitely until a judgement was made. Once subjects had decreased the level of the interferer programme appropriately they pushed the rotary fader (which incorporated a button) to submit their response.

This methodology, known as the ‘method of adjustment’, is sometimes considered to be less accurate than other psychophysical test methods, such as Alternative Forced Choice (AFC) style procedures. Hesse (1986) tested the effect of a range of psychophysical procedures on masking thresholds for tone masked by noise. Thresholds fell into two groups: AFC style procedures; and non-AFC procedures including the method of adjustment, adaptive control and yes/no procedures. The results showed only small differences between procedures with thresholds obtained using AFC procedures around 2 dB lower than those found using the non-AFC procedures. Similar results were found by Watson and Nichols (1976). In an ecologically valid interference scenario, however, a listener is likely to be concerned only with whether the interfering programme is audible, which is similar to a yes/no paradigm. Since yes/no masking thresholds were very close to those using a method of adjustment, and since the method of adjustment task is the fastest and most intuitive for the subjects (Bech and Zacharov 2006), this procedure was considered appropriate for use.

Ten subjects reporting no hearing impairments, aged between 21 and 38 years, participated in each experiment with eight of the subjects taking part in both the masking and acceptability listening tests. In both experiments the proportions of trained listeners were such that four subjects had training in critical listening and experience conducting and participating in psychoacoustic experiments, four subjects had no such experience but were musicians, and two subjects had no experience in any of these domains.

3.1.2 Stimuli

Three items of target programme material and three items of interferer programme material were selected for use. All stimuli were of duration 10 seconds; this was considered sufficiently short to allow for a reasonable number of trials to be conducted, yet sufficiently long to include realistic programme variability. If the excerpts were of considerably greater duration, the validity of a single masking threshold for the trial would be questionable, whereas if the excerpts were considerably shorter their briefness may diminish their ecological validity and acceptability judgements may be questionable. On each trial the stimuli were looped indefinitely until a decision was made. The targets and interferers were selected to cover a range of programme types and genres. The targets were excerpts of: classical music (Brahms’s Hungarian Dance
No. 18), pop music (The Killers’ On Top), and football commentary (recorded from BBC iplayer). The interferers were excerpts of: classical music (Mahler’s Symphony No. 5 Mov. 4), pop music (The Bravery’s Give in), and male speech (from the BBC Radio 4 show ‘Points of View’).

Any system designed to control interference between signals may have some effect on the magnitude spectrum of the interferer signal. In order to consider this a further six interferers, filtered replicates of the first three, were also used. Three were low pass filtered (LPF) at 200 Hz with a 9 dB/oct roll-off, based upon the results of Akeroyd et al. (2007), and three were high pass filtered (HPF) at 1 kHz with a 16 dB/oct roll-off, based upon the results of Jacobsen et al. (2011).

A single channel recording of road noise was decorrelated according to the method described in Pulkki (2007) and replayed over the 6 channel hexagonal loudspeaker array.

Benjamin and Crockett (2005) identified preferred listening levels for music in the automotive environment at between 70 and 76 dBA for a range of vehicle speeds including stationary (engine off), thus the target programmes were reproduced at a level of 76 dB LAeq measured at the listening position with a time constant of 20 seconds (i.e. programme replayed twice). The road noise was adjusted to 60 dB LAeq which was found to be a good approximation for road noise levels inside automobiles travelling at 30 mph in the above mentioned study. The interferers were set to a starting level which was randomly selected between 70 and 76 dB LAeq in order to minimise the opportunity for listeners to select the masking threshold by recalling the number of rotations of the rotary fader used on a previous trial. Additionally, this range of starting levels ensured that the the interferer programmes were clearly audible before listeners made their judgements.

Levels were verified using a MiniSPL measurement microphone (an omnidirectional microphone with a free-field transducer). In order to prevent accidental hearing damage the user interface was designed such that an increase of no more than +6dB was permitted to the starting level of the interferer.

The experiment design was full factorial with two repetitions per trial, thus there were 108 trials per subject (3 targets x 3 interferers x 3 filtering levels x 2 road noise levels x 2 repetitions). The experiment was carried out with three sessions per subject, with 36 trials per session. Each session contained one target, but the order of sessions was randomised across subjects to minimise any training effect.
3.2 Results

In this section the experiment results are analysed. The dependent variables, masking and acceptability thresholds, were obtained by subjects reducing the level of the interferer with respect to the fixed target level. The thresholds, therefore, are reported in terms of their TIRs, which is analogous to an SNR.

3.2.1 Masking experiment

For the listening scenarios featuring the pop target with the Low Pass Filtered (LPF) classical music (with and without noise), subject 5 reported a masking threshold at 74 dB SNR. Since the other nine subjects reported masking thresholds ranging from 11 to 35 dB SNR, it is likely that these data points are outliers caused by the subject mistakenly identifying a component within the target programme as belonging to the interferer. These two data were removed from further analysis.

Figure 3.2 shows the mean TIR of the masking thresholds for all listening scenarios separated by target and interferer programmes, with error bars representing the 95% confidence intervals. The general trend indicates that the pop target programme was most effective at masking the interferer programmes, with the classical and sports commentary programmes requiring more than an additional 10 dB TIR in most cases.

Shapiro-Wilk tests of sample size $n=20$ (except where outliers were removed) showed that when the data were separated by target, interferer, road noise, and filtering, 11 of the 54 groups were not normally distributed with 95% confidence. Observations of the histograms provided little evidence to support or refute this due to the relatively low sample size per group, so all data were analysed using both parametric and non-parametric tests. No discrepancies between results were found so only results of the parametric tests are reported.

An Analysis of Variance (ANOVA) was conducted with the data split by target, interferer, filtering, and road noise as fixed factors and subject as a random factor. The analysis revealed that all main effects were significant, as well as a number of interactions (see table 3.1). When high level interactions are present an isolated analysis of main effects can be misleading, because any apparent trends amongst main effect levels will vary with the interacting factors. If, however, the interactions have a much smaller effect size than the main effects, it may still be meaningful to consider main effects in isolation. In this case, the five way interaction was significant and had an effect size (partial $\eta^2 = .549$) larger than all other interactions, and larger than all main effects excluding the target programme (partial $\eta^2 = .872$) and the subject (partial $\eta^2 = .641$). The effect sizes of these main effects were not so much larger, however, that they can be considered clearly independent from the interactions.
Figure 3.2: Mean masking thresholds for all listening scenarios split by target and interferer programmes.
Table 3.1: ANOVA of masking thresholds with target programme, interferer programme, interferer filtering, and road noise as fixed features, and with subject as a random factor. Only main effects and interactions with significance $< 0.05$ are shown.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean sq.</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>918860.258</td>
<td>1</td>
<td>918860.258</td>
<td>932.068</td>
<td>$&lt;.001$</td>
<td>.990</td>
</tr>
<tr>
<td>Target</td>
<td>33705.706</td>
<td>2</td>
<td>16852.853</td>
<td>61.139</td>
<td>$&lt;.001$</td>
<td>.872</td>
</tr>
<tr>
<td>Interferer</td>
<td>1105.783</td>
<td>2</td>
<td>552.892</td>
<td>6.249</td>
<td>.004</td>
<td>.188</td>
</tr>
<tr>
<td>Noise</td>
<td>778.253</td>
<td>1</td>
<td>778.253</td>
<td>30.095</td>
<td>$&lt;.001$</td>
<td>.067</td>
</tr>
<tr>
<td>Filter</td>
<td>3924.350</td>
<td>2</td>
<td>1962.175</td>
<td>75.877</td>
<td>$&lt;.001$</td>
<td>.266</td>
</tr>
<tr>
<td>Subject</td>
<td>8873.397</td>
<td>9</td>
<td>985.933</td>
<td>3.576</td>
<td>.010</td>
<td>.641</td>
</tr>
<tr>
<td>Int*Fil</td>
<td>2932.227</td>
<td>4</td>
<td>733.057</td>
<td>28.347</td>
<td>$&lt;.001$</td>
<td>.213</td>
</tr>
<tr>
<td>Noi*Fil</td>
<td>678.453</td>
<td>2</td>
<td>339.227</td>
<td>13.118</td>
<td>$&lt;.001$</td>
<td>.059</td>
</tr>
<tr>
<td>Tar*Fil</td>
<td>2024.378</td>
<td>4</td>
<td>506.095</td>
<td>19.570</td>
<td>$&lt;.001$</td>
<td>.158</td>
</tr>
<tr>
<td>Tar*Int</td>
<td>1805.065</td>
<td>4</td>
<td>451.266</td>
<td>5.100</td>
<td>.001</td>
<td>.274</td>
</tr>
<tr>
<td>Tar*Noi</td>
<td>644.074</td>
<td>2</td>
<td>322.037</td>
<td>12.453</td>
<td>$&lt;.001$</td>
<td>.056</td>
</tr>
<tr>
<td>Tar*Sub</td>
<td>4962.617</td>
<td>18</td>
<td>275.701</td>
<td>3.116</td>
<td>.001</td>
<td>.509</td>
</tr>
<tr>
<td>Tar<em>Int</em>Fil</td>
<td>1222.464</td>
<td>8</td>
<td>152.808</td>
<td>5.909</td>
<td>$&lt;.001$</td>
<td>.102</td>
</tr>
<tr>
<td>Tar<em>Noi</em>Fil</td>
<td>648.349</td>
<td>4</td>
<td>162.087</td>
<td>6.268</td>
<td>$&lt;.001$</td>
<td>.057</td>
</tr>
<tr>
<td>Tar<em>Int</em>Noi</td>
<td>370.735</td>
<td>6</td>
<td>61.789</td>
<td>2.389</td>
<td>.028</td>
<td>.033</td>
</tr>
<tr>
<td>Tar<em>Int</em>Sub</td>
<td>4779.977</td>
<td>54</td>
<td>88.518</td>
<td>3.422</td>
<td>$&lt;.001$</td>
<td>.307</td>
</tr>
<tr>
<td>Tar<em>Int</em>Noi<em>Fil</em>Sub</td>
<td>10789.993</td>
<td>417</td>
<td>25.875</td>
<td>1.572</td>
<td>$&lt;.001$</td>
<td>.549</td>
</tr>
</tbody>
</table>

In conclusion, therefore, the interaction between all factors was significant and explains most of the variation in masking thresholds; the target programme, however, had the largest effect. This conclusion is reasonable, since the spectro-temporal properties of the target programme provided the masking which determines the masking thresholds. The noise programme aids this masking by contributing additional steady state, broadband masking energy, and the interferer programme, along with its filtering, determine what signal is presented to be masked.

3.2.2 Acceptability experiment

Reliability of subjects

The results were first analysed for subject reliability. Since each subject provided two scores for every listening scenario the absolute difference between each pair of scores was calculated and then averaged across trials to give the subject mean difference. The mean difference score is a simple measure of subject consistency. Two subjects (subjects 1 and 7) were identified as having unusually high mean difference scores which warranted further investigation (see table 3.2). The mean scores of subjects 1
and 7 for each listening scenario were also found to disagree with the mean scores of all subjects for each listening scenario. Figures 3.3 and 3.4 show this effect. Although the disagreement with the mean scores is apparent, this is not sufficient grounds for removal of a subject because the dependent variable rated was of a highly subjective nature and some disagreement among subjects was expected.

It was further considered possible that subjects who considered a higher level of interference to be acceptable may be less consistent in their scoring if, as a result of higher interferer levels, they were less able to repeat their previous judgement. In such a case subjects with lower consistency scores should not be removed if they were found to have higher acceptability thresholds. In order to consider this possibility mean scores across repeats were compared with the corresponding absolute differences between repeats. If subjects with a greater tolerance to higher interferer levels usually produce wider discrepancies between their scores a positive correlation would be expected between mean TIRs and absolute difference between repeated scores. Figure 3.5 shows the scatter plot of these data points. No positive correlation is evident, and the greatest differences were found in the middle range (TIRs between 15 and 30 dB).

Finally, a Tucker-1 correlation loading plot was produced using panelcheck ([PanelCheck Analysis Tool](#)) to help identify which, if any, subjects should be removed from the analysis. The plot shows the principal components upon which subjects produced their scores. In this case the first principal component accounts for 70.9% of the total variation of scores, implying that it is the main basis upon which subjects made their judgements (see fig. 3.6). The scores from subject 7 were not strongly correlated along the first principal component with those of the other 9 subjects. On this basis subject 7 was removed from further analysis.

<table>
<thead>
<tr>
<th>Subject</th>
<th>mean difference in TIR / dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.0</td>
</tr>
<tr>
<td>2</td>
<td>2.2</td>
</tr>
<tr>
<td>3</td>
<td>2.2</td>
</tr>
<tr>
<td>4</td>
<td>2.8</td>
</tr>
<tr>
<td>5</td>
<td>1.8</td>
</tr>
<tr>
<td>6</td>
<td>2.6</td>
</tr>
<tr>
<td>7</td>
<td>5.9</td>
</tr>
<tr>
<td>8</td>
<td>2.5</td>
</tr>
<tr>
<td>9</td>
<td>2.6</td>
</tr>
<tr>
<td>10</td>
<td>3.1</td>
</tr>
</tbody>
</table>

Table 3.2: The subject mean (across all trials) difference between trial repeats.
Chapter 3: Masking and Acceptability Experiment

Figure 3.3: Subject 1 scores (red) are plotted against all other subject scores (grey). The line $y=x$ denotes the mean subject response.

Figure 3.4: Subject 7 scores (red) are plotted against all other subject scores (grey). The line $y=x$ denotes the mean subject response.
Figure 3.5: Mean acceptability thresholds plotted against absolute differences between repeated trials.
Figure 3.6: A principal component analysis of the subject responses presented as a Tucker-1 plot. The scores for Subjects 1 and 10 overlap.
Outliers

The identification and removal of outliers must be handled very carefully in this analysis because acceptability is a hedonic judgement, and differing scores are not necessarily evidence of experimental error. Since there are two data points for each subject within each listening scenario, however, it is possible to identify cases wherein a subject was very inconsistent. If, therefore, a subject provided two data points which greatly differ from the other 18 but are similar to each other, it is not reasonable to exclude them from further analysis. In contrast, however, where a subject has provided one data point close to the mean and another data point which appears to be an outlier it may be reasonable to interpret this as experimental error and exclude the outlier from further analysis.

When the data were split by target, interferer, filtering, and road noise, six data points were identified with value at least three times the interquartile range above the upper quartile or below the lower quartile. An additional 24 extreme values were identified where the data point was at least 1.5 times the interquartile range above the upper quartile or below the lower quartile.

The first outlier considered was a data point produced by subject 1 for the pop target with unfiltered classical interferer including road noise. Although this data point was around 20 dB TIR above the upper quartile (31.67 dB TIR) the associated repeat was also an extreme value (20.17 dB TIR). Removal of this outlier would be ignoring the possibility that the judgements of subject 1 genuinely lie well below that of the other subject’s. Since this possibility cannot be ruled out this outlier was not removed.

The second outlier considered was a data point produced by subject 1 for the classical target and High Pass Filtered (HPF) classical interferer including road noise. The data point was 25.00 dB TIR and the associated repeat was 18.67 dB TIR, which was not an extreme value. Although the associated data point was neither an outlier nor an extreme value the absolute difference between the values is 6.33 dB TIR, less than the range of the data. Additionally two extreme values lie between the maximum value and this outlier, which seems to imply that subjects had some difficulties discerning the acceptability threshold for this listening scenario but agreed that the acceptability threshold may be higher than predicted by other subjects. As such this outlier was not removed.

The third outlier considered was a data point produced by subject 1 for the pop target and unfiltered pop interferer including road noise. The data point was 21.5 dB TIR and the associated repeat was 10.5 dB TIR, which was very close to the median score for this listening scenario (10.33 dB TIR) and slightly below the mean (11.04 dB TIR). No other extreme values were present for this listening scenario, and the interquartile range was only 2.67 dB. It seems likely, therefore, that this outlier was indeed a subject input error and it was removed from further analysis.
The fourth outlier considered was a data point produced by subject 1 for the pop target and HPF pop interferer excluding road noise. The data point was 17.50 dB TIR and the associated repeat was 12.17 dB TIR, which was just above the median (11.50 dB TIR) and mean (11.88 dB TIR) for this listening scenario. With an interquartile range of 2.04 dB TIR it may be difficult to justify a discrepancy of 5.33 dB between repeats, however subject 4 also produced an extreme value in this listening scenario of 16.83 dB TIR and it would be difficult to justify the exclusion of one without the other. In any case it cannot be ruled out that the judgements of these subjects were deliberately made and above that of the other subjects. The outlier was therefore not removed from further analysis.

The fifth outlier considered was a data point produced by subject 1 for the classical target and HPF pop interferer including road noise. The data point was 31.67 dB TIR and the associated repeat was 20.83 dB TIR, which was below the median (21.92 dB TIR) and mean (22.38 dB TIR) for this listening condition. The discrepancy of 10.84 dB TIR is large compared to the interquartile range of 2.13 dB TIR. Subject 5 produced the only other extreme value in this listening scenario at 25.67 dB TIR. The difference between this value and the outlier is around three times the interquartile range. It is reasonable to suggest, therefore, that this outlier was a subject input error. The outlier was therefore removed from further analysis.

The final outlier considered was a data point produced by subject 1 for the pop target and HPF pop interferer including road noise. The data point was 24.17 dB TIR and the associated repeat was 8.00 dB TIR, which was below the median (11.67 dB TIR) and the mean (12.03 dB TIR) for this listening scenario. This very large discrepancy of 16.17 dB TIR between repeats, and the lack of any extreme values for this listening scenario, mark this data point as the likely result of a subject input error. This outlier was therefore removed from further analysis.

Three of the six outliers were therefore removed from further analysis, leaving a total sample size of \( n = 969 \) samples. All six outliers were reported by subject one who, it was reported in section 3.2.2, was also found to have an unusually high mean difference between repeats but was not distinctly marked out on the Principal Component Analysis (PCA). This implies that subject one was likely to be performing the correct task, but was occasionally less diligent than the other subjects. It should also be noted that five of the six listening scenarios featuring an identified outlier included road noise. It is possible therefore that the presence of road noise increased the difference between repeated scores for subject one by increasing the subject’s uncertainty about the acceptability threshold.
Table 3.3: An ANOVA of the data separated by target, interferer, filtering, and road noise as fixed factors and by subject as a random factor.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>F</th>
<th>Sig.</th>
<th>Partial η²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>244193.217</td>
<td>1</td>
<td>251.333</td>
<td>.000</td>
<td>.969</td>
</tr>
<tr>
<td>Target</td>
<td>11003.106</td>
<td>2</td>
<td>61.639</td>
<td>.000</td>
<td>.885</td>
</tr>
<tr>
<td>Interferer</td>
<td>2854.942</td>
<td>2</td>
<td>13.192</td>
<td>.000</td>
<td>.622</td>
</tr>
<tr>
<td>Filter</td>
<td>6429.909</td>
<td>2</td>
<td>400.043</td>
<td>.000</td>
<td>.483</td>
</tr>
<tr>
<td>Noise</td>
<td>838.769</td>
<td>1</td>
<td>104.370</td>
<td>.000</td>
<td>.109</td>
</tr>
<tr>
<td>Subject</td>
<td>7773.474</td>
<td>8</td>
<td>6.134</td>
<td>.001</td>
<td>.714</td>
</tr>
<tr>
<td>Interferer * Subject</td>
<td>467.697</td>
<td>4</td>
<td>14.549</td>
<td>.000</td>
<td>.064</td>
</tr>
<tr>
<td>Filter * Noise</td>
<td>234.001</td>
<td>2</td>
<td>14.559</td>
<td>.000</td>
<td>.033</td>
</tr>
<tr>
<td>Target * Filter</td>
<td>1620.362</td>
<td>4</td>
<td>50.406</td>
<td>.000</td>
<td>.191</td>
</tr>
<tr>
<td>Interferer * Subject</td>
<td>1731.591</td>
<td>16</td>
<td>2.769</td>
<td>.007</td>
<td>.581</td>
</tr>
<tr>
<td>Target * Interferer</td>
<td>626.876</td>
<td>4</td>
<td>4.011</td>
<td>.010</td>
<td>.334</td>
</tr>
<tr>
<td>Target * Noise</td>
<td>513.506</td>
<td>2</td>
<td>31.948</td>
<td>.000</td>
<td>.070</td>
</tr>
<tr>
<td>Target * Subject</td>
<td>1428.224</td>
<td>16</td>
<td>2.284</td>
<td>.023</td>
<td>.533</td>
</tr>
<tr>
<td>Target * Interferer * Filter</td>
<td>457.320</td>
<td>8</td>
<td>7.113</td>
<td>.000</td>
<td>.062</td>
</tr>
<tr>
<td>Target * Filter * Noise</td>
<td>357.231</td>
<td>4</td>
<td>11.113</td>
<td>.000</td>
<td>.049</td>
</tr>
<tr>
<td>Target * Interferer * Noise</td>
<td>103.613</td>
<td>6</td>
<td>2.149</td>
<td>.046</td>
<td>.015</td>
</tr>
<tr>
<td>Target * Interferer * Subject</td>
<td>1250.563</td>
<td>32</td>
<td>4.863</td>
<td>.000</td>
<td>.154</td>
</tr>
</tbody>
</table>

ANOVA

Shapiro-Wilk tests of sample size n = 18 showed that when the data were separated by target, interferer, road noise, and filtering, only seven of the 54 groups were normally distributed with 95% confidence. With relatively small sample sizes (17 or 18) in each group, however, this is unsurprising. Observations of the histograms provided little evidence to support or contradict this evidence of non-normality due to the relatively low sample sizes. All data were therefore analysed using both parametric and non-parametric tests. No discrepancies between results were found so only results of the parametric tests are reported here.

An ANOVA was conducted with the data split by target, interferer, filtering, and road noise as fixed factors and subject as a random factor. The ANOVA was then re-run with the non-significant interactions excluded. Table 3.3 shows the result of this second ANOVA.

Effect of interactions

All main effects and many interactions were significant. As shown in table 3.3, many of the two and three way interactions are significant and have an effect size as great...
as some of the main effects. Of the three way interactions the interaction between target, interferer, and subject has the greatest effect size (partial $\eta^2 = .154$). It is not surprising, however, that for various combinations of target and interferer subjects may disagree on the required relative levels for the listening scenario to be considered acceptable. Likewise the two way interactions with the greatest effect size included the subject.

The remaining significant three way interactions have partial $\eta^2 < .065$, the largest of which is between target, interferer, and filter. Figure 3.7 shows this interaction. The effect seems to be that for a given target programme and interferer filtering the male speech interferer programme had the highest TIR, the pop interferer had a lower TIR, and the classical interferer had the lowest TIR. Notable exceptions were present for the classical target where the male speech interferer was considered acceptable with a lower TIR, especially when it was LPF. In general LPF interferers produced lower acceptability TIRs than unfiltered or HPF interferers. The confidence intervals of the unfiltered and HPF cases overlap, while many of the LPF cases do not. For the pop target, however, the mean TIRs occupy a relatively small range (around 9 dB to 15 dB), indicating that changes in interferer programme and filtering are relatively unimportant for this programme. The acceptability TIRs for the classical target and sports commentary target had approximately double this range.

While this interaction is interesting, the differences between the mean acceptability thresholds and those which would be expected without any three-way interaction are quite small (within 2 dB). This is indicative of the relatively small effect size partial $\eta^2 = .062$ compared with those of the corresponding main effects involved.

Of the significant two way interactions, excluding the subject factor, the largest effect was attributed to the interaction between target and interferer. Figure 3.8 shows the result of this two way interaction. For both the classical and sports commentary targets the pop interferer required around 4 dB of level reduction more than the classical interferer, whereas this was around 1 dB for the pop target. For the pop and sports commentary targets approximately 3 dB of further reduction in level was required for the male speech interferer to be considered acceptable, whereas for the classical target the male speech interferer did not require as much level reduction as the pop interferer.
Figure 3.7: Mean acceptability TIRs showing the three way interaction between target, interferer, and filtering. The error bars indicate the 95% confidence intervals.
Figure 3.8: Mean acceptability thresholds for all listening scenarios split by target and interferer programmes.
Chapter 3: Masking and Acceptability Experiment

Figure 3.9: Acceptability TIRs for all listening scenarios split by subject.

Main effects

The main effect sizes were partial $\eta^2 = .885$, .622, .483, .109, and .714 for target, interferer, filtering, noise, and subject respectively. Therefore the effects of the target, the interferer, and the subject were large, while the effects of the filtering and noise were moderate and small.

For targets, a Tukey’s post-hoc test revealed that TIRs for all three target programmes were significantly different with means of 17.46 dB for the classical target, 11.17 dB for the pop target, and 18.97 dB for the sports commentary target. For the interferers, a Tukey’s post-hoc test revealed that TIRs were significantly different for all three groups with means of 13.57 dB for the classical interferer, 16.39 dB for the pop interferer, and 17.67 dB for the male speech interferer. For subjects, the median TIRs ranged from 22.00 dB to 12.17 dB for all listening scenarios (see fig. 3.9).

For filtering, a Tukey’s post-hoc test revealed that TIRs were significantly different for all three groups with means of 17.18 dB for unfiltered interferers, 12.29 dB for LPF interferers, and 18.17 dB for the HPF interferers. For noise, the ANOVA revealed a significant difference between mean TIRs with and without road noise with mean acceptability thresholds of 16.81 dB TIR without road noise, and 14.93 dB TIR with
3.2.3 Subject comments

Subjects commented that the cymbal crash in the pop music interferer and the sibilance in the male speech interferer were particularly important cues for detection. This comment was supported by the masking experiment results which showed that the LPF pop interferer presented with the classical target had a masking threshold about 10 dB higher than the equivalent listening scenario featuring the HPF and unfiltered pop interferers. The masking threshold for the LPF male speech interferer masked by the classical and sports commentary targets was also much higher (greater than 7 dB) than for the respective HPF unfiltered male speech interferer. By contrast the effect of filtering the classical interferer was relatively small, affecting the masking thresholds by less than 2 dB in all cases. This does seem to imply that high frequency cues to detection were particularly important, since cymbals and sibilance tend to have high energy at around 6 kHz.

Subjects also commented that when road noise was present, as the interferers were reduced in level they became obscured by the road noise and were thus more easily acceptable. This was supported by the results which showed a decrease in acceptability thresholds of approximately 2 dB TIR when the road noise was present. This supports the hypothesis that for the production of sound zones within automotive environments, the presence of road noise decreases the contrast necessary in order to achieve the threshold of acceptability. This effect is likely to be even larger when the road noise is louder (e.g. when driving at higher speeds).

3.2.4 Discussion

The results of the acceptability experiment show that most mean acceptability thresholds lie near 15 dB TIR, and that almost all data were included within 10 dB either side of this figure (i.e. between 5 dB and 25 dB TIR), whereas the masking thresholds were higher ranging between 20 and 40 dB TIR. In both cases the largest main effect was the target programme, with lower mean thresholds found for pop music targets than when listening to classical music or sports commentary. The mean thresholds for the pop target were also more consistent across different interferer and filtering levels than the classical or sports commentary targets.

For acceptability the second largest effect (excluding the subject) was the interferer, with subjects finding classical music interference at higher levels acceptable than the levels scored for pop or male speech interference. The two way interaction showed that while the classical interferer had the lowest acceptability thresholds and the pop interferer had slightly higher acceptability thresholds, the male speech interferer had
much higher acceptability thresholds but only for the pop and sports commentary targets. It seems possible that these results might be explained in terms of higher level factors such as dynamic compression, or the interference between linguistic content in programmes.

Filtering of the interferer also had a large effect on the acceptability TIRs. LPF interferers had considerably lower acceptability thresholds than the HPF or unfiltered interferers indicating that, in general, low frequency interference was considered more tolerable. Road noise was also significant, the presence of which diminished the level reduction required to render the listening scenario acceptable. Since the road noise level was fixed it acted like an additional masker, hiding cues to detection of the interferer.

For acceptability the subject factor was significant and had a large effect. Median scores across listening scenarios fell within a range of 10 dB. This degree of subject differences is consistent with Francombe et al. (2012), however this experiment did not show bimodal subject data (i.e. subjects did not appear to be clearly demarcated into ‘low tolerance’ and ‘high tolerance’ groups). It is also possible that this may be due to differing interpretations of the word ‘acceptable’; e.g. the ‘relaxing at home’ scenarios imagined by the subjects might have involved differing time-scales of listening.

### 3.3 Comparison of masking and acceptability

The principal goal of conducting these experiments was to obtain a matching set of masking and acceptability thresholds. This was required in order to investigate any relationship between the two variables which, in the most ideal case, could be used predictively. Section 3.3.1 describes the analysis carried out on a data set computed by taking the difference between acceptability and masking scores and section 3.3.2 outlines the correlation analysis carried out to determine the extent to which masking data can be used to predict acceptability data. Section 3.4 discusses the implications of these analyses.

The investigation into a relationship between masking and acceptability thresholds is not an arbitrary search for correlation. It is based on a theoretical framework of audition which assumes that the two factors must be related. The definition of a masking threshold implies that there should not be acceptability thresholds below the masking threshold (outside of experimental error and subject JNDS). As a minimum, therefore, masking thresholds would be expected to correlate with acceptability thresholds due to this limit. Furthermore, it seems plausible that partial loudness may be related to acceptability, because a loud interferer is unlikely to be considered acceptable. It was argued in section 2.4 that partial loudness and masking thresholds are strongly related, and as such there should be some correlation between masking and acceptability. It is expected, however, that many other factors, such as speech intelligibility, dynamic
range, or relative harmonicity may also influence the acceptability threshold. As such an investigation into the degree of correlation between these variables is required to determine the extent to which masking alone can predict acceptability.

### 3.3.1 ANOVA of difference scores

The masking and acceptability experiments each featured 10 subjects, 8 of which took part in both experiments. An ANOVA was carried out on a set of data computed by first taking the mean score for all repetitions across acceptability and masking data, before subtracting the mean masking thresholds from the mean acceptability thresholds. The resultant data set thus describes the mean difference between acceptability and masking data per subject, target, interferer, filter and noise. Table 3.4 shows the result of the ANOVA.

The ANOVA indicates that all main factors, except road noise, were significant at $p = 0.05$, as well as a few higher order interactions. Since the five way interaction was significant and had the greatest partial $\eta^2$, however, the lower order interactions and main effects can not bemeaningfully considered independently. The finding of this ANOVA, therefore, is that differences between masking and acceptability thresholds depend upon the target programme, interferer programme, interferer filtering, road noise, and subject. Post hoc test (Tukey HSD) revealed that all three target subsets and all three interferer subsets could be considered significantly different, and unfiltered and HPF interferers could be considered significantly different from LPF interferers. These subsets are shown in Table 3.5.

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>F</th>
<th>Sig</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>163455.342</td>
<td>1</td>
<td>109.684</td>
<td>.000</td>
<td>.940</td>
</tr>
<tr>
<td>Target * Interferer *</td>
<td>12592.196</td>
<td>354</td>
<td>1.321</td>
<td>.003</td>
<td>.520</td>
</tr>
<tr>
<td>Filter * Noise * Subject</td>
<td>3674.687</td>
<td>2</td>
<td>8.782</td>
<td>.000</td>
<td>.226</td>
</tr>
<tr>
<td>Target</td>
<td>1993.754</td>
<td>2</td>
<td>4.765</td>
<td>.012</td>
<td>.137</td>
</tr>
<tr>
<td>Interferer</td>
<td>540.179</td>
<td>2</td>
<td>7.593</td>
<td>.001</td>
<td>.041</td>
</tr>
<tr>
<td>Subject</td>
<td>10431.697</td>
<td>7</td>
<td>7.123</td>
<td>.000</td>
<td>.454</td>
</tr>
<tr>
<td>Interferer * Filter</td>
<td>1149.557</td>
<td>4</td>
<td>8.079</td>
<td>.000</td>
<td>.084</td>
</tr>
<tr>
<td>Target * Interferer *</td>
<td>12553.372</td>
<td>60</td>
<td>5.882</td>
<td>.000</td>
<td>.499</td>
</tr>
</tbody>
</table>

Table 3.4: An ANOVA of the differences between masking and acceptability data separated by target, interferer, filtering, and road noise with non-significant interactions excluded and with subject set as a random factor.
<table>
<thead>
<tr>
<th></th>
<th>subset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target</strong></td>
<td>subset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Pop</td>
<td>252</td>
<td>11.5473</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical</td>
<td>252</td>
<td>13.0979</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sports Commentary</td>
<td>252</td>
<td>16.3876</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Interferer</strong></td>
<td>subset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Pop</td>
<td>252</td>
<td>12.0099</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Speech</td>
<td>252</td>
<td>13.2236</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classical</td>
<td>252</td>
<td>15.7996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td><strong>Filter</strong></td>
<td>subset</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>252</td>
<td>12.8724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPF</td>
<td>252</td>
<td>13.2890</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPF</td>
<td>252</td>
<td>14.8717</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sig</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5: Homogeneous subsets of targets, interferers, and filters according to a Tukey HSD post hoc tests.
3.3.2 Correlations

Although the ANOVA failed to identify structures within the data which would allow stronger prediction by grouping according to subsets, a good correlation may still be found between masking and acceptability if variances across the factors considered are relatively small. Since the masking and acceptability experiments had eight subjects in common, the data could be meaningfully analysed separately by subject, target, interferer, filtering and noise. The data were first averaged across each pair of repeats, however, since the selection of masking and acceptability pairs according to repetition would be arbitrary. Following this, Pearson’s Correlation Coefficient (R) was calculated for acceptability and masking thresholds for all data points. R describes the linearity of the correlation between two variables and is calculated with:

\[ R = \frac{\Sigma_{i=1}^{N} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2} \times \sqrt{\Sigma(Y_i - \bar{Y})^2}}, \]  

(3.1)

where \(X\) and \(Y\) represent the prediction and the observation respectively, from Howitt and Cramer (1997).

The square of \(R\), known as the coefficient of determination, describes the quantity of the variance in \(Y\) explained by \(X\) and is also of interest. For acceptability and masking thresholds for all data points the correlation was calculated at \(R = 0.59\), with \(p < 0.001\) and \(n = 432\). The coefficient of determination was therefore 0.348 which indicates that just under 35% of the variance in acceptability scores was accounted for by the masking threshold. Figure 3.10 shows a scatterplot of this data. Few of the data points fall below the line \(y=x\), which is to be expected since the interferer should always be considered acceptable when it is inaudible, thus requiring no further diminution of level.

This analysis considered subjects individually because such a comparison was possible. For practical applications, however, the mean thresholds are likely to be of more interest where acceptability predictions are not required to be tailored to an individual listener. A correlation was therefore re-calculated for acceptability and masking thresholds averaged across repetitions and subjects. \(R\) was calculated as 0.87, with \(p < 0.001\) (one-tailed) and \(n = 54\). With a coefficient of determination at \(R^2 = 0.76\), 76% of the variance in mean acceptability scores was accounted for by the mean masking thresholds. In this model the slope was equal to 0.657 and the constant was 3.471, thus acceptability may be calculated using the equation:

\[ \bar{A}_T = (0.657 \times \bar{M}_T) + 3.471. \]  

(3.2)

where \(\bar{A}_T\) and \(\bar{M}_T\) represent the acceptability and masking thresholds respectively.

Figure 3.11 shows a scatterplot of the data separated by target programme.
Figure 3.10: A scatter plot of acceptability and masking TIRs for data averaged across repeats. The line $y=x$ indicates a maximum positive correlation.

To further evaluate the prediction accuracy of the model the prediction error should be calculated. This is commonly described using Root Mean Squared Error (RMSE). The RMSE describes the average error across all trials (disregarding the direction of the error) and is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n-k} \sum_{i=1}^{i=n} (Y_i - X_i)^2}$$  

where $Y_i$ is the acceptability and $X_i$ is the predicted acceptability for trial $i$. The number of trials is given by $n$ and $k$ represents the number of features on which the model is trained. In this case there are only two features: the masking threshold and the constant.

For this model the accuracy was equal to $\text{RMSE} = 2.63$ dB. For comparison, the root mean squared error between repeats across all subjects and listening scenarios was 4.5 dB, which is to say that the model predicted the mean acceptability thresholds with greater accuracy than subjects were able to repeat their judgements. The consistency of the model can be measured using the Outlier Ratio (OR), which is equal to the ratio of outliers to total data points. In this case a prediction is considered an outlier if it lies
more than 1 standard deviation from the reported mean value. The OR was calculated as OR = 3.7% (with the following two conditions classified as outliers: pop target with LPF classical interferer without road noise, and sports commentary target with LPF classical interferer with road noise).

3.4 Summary and conclusion

Two research questions were posed at the start of this chapter: *What is the range of SNRs over which acceptability primarily varies?* and *Is there a relationship between masking and acceptability?*

The first question was answered, in principle, in chapter 1 by noting that when the interferer is inaudible the listening scenario must be acceptable and when the target is inaudible the listening scenario must be unacceptable. The range of acceptability, therefore, is marked by the audibility of the target and interferer programmes. In practice, however, acceptability mostly varies over a much smaller range of SNRs than
constituted by the edges of audibility.

A pair of listening tests were conducted to answer these questions. The results showed that acceptability thresholds varied from $3 - 40$ dB SNR, and across subjects mean acceptability thresholds varied from $7 - 27$ dB SNR. By contrast, the masking thresholds varied from $5 - 65$ dB SNR, and mean masking thresholds varied from $16 - 43$ dB SNR. As a general rule, therefore, mean masking thresholds for ecologically valid programmes of this sort can be as high as $43$ dB, with mean acceptability thresholds only as high as $27$ dB.

To answer the second research question, the differences between mean masking and acceptability thresholds were calculated across equivalent listening scenarios. The ANOVA of these difference scores revealed that the difference between masking and acceptability scores for various listening scenarios varied according to an interaction between all the factors considered (including subject). The disagreement between subjects implies that some level of individual preference is contained within judgements of acceptability. Further analysis revealed that, while individual subject scores differed, a relatively consistent difference between mean masking thresholds and mean acceptability thresholds exists. Based on this, the use of a linear model to predict acceptability thresholds from mean masking thresholds was suggested.

Predicted and observed acceptability TIRs were fairly well correlated ($R^2 = 0.76$). A linear regression model (see eq. (3.2)) was used to predict acceptability thresholds, and the model had accuracy of RMSE = 2.6 dB and a consistency of OR = 3.7%. By way of comparison the RMSE between repeats across all subjects and listening scenarios was (4.5 dB). This implies that the model predicts the mean acceptability threshold with greater accuracy than subjects were able to repeat their judgements.

If predictions of acceptability thresholds can be made using known masking thresholds then it follows that predictions of acceptability thresholds could also be made using predictions of masking thresholds, although the extent of the compounded error would need to be considered. The next chapter investigates the selection and calibration of a masking threshold model for this purpose.
4 Masking Prediction

In the previous chapter masking and acceptability experiments were carried out, and the resulting masking thresholds were used to predict acceptability thresholds. With this relationship established acceptability thresholds can be predicted for listening scenarios where the masking thresholds are already known. In most practical applications, however, masking thresholds are unlikely to be known in advance and it therefore becomes necessary to predict masking thresholds first. A research question is therefore posed: “how can auditory masking be predicted?”

This chapter addresses this question by considering a range of existing models for the prediction of auditory masking. These models are discussed and compared in section 4.1 before selecting one for implementation. In section 4.2 the implementation and modification to the selected masking threshold model is outlined. In sections 4.3 and 4.4 the prediction accuracy of the model, and subsequent accuracy for predicting acceptability are investigated. Finally in section 4.5 the work is summarised.

4.1 Masking threshold prediction models

A range of models for the prediction of masking phenomena exist; some describe only the occurrence of a specific masking phenomena whilst others aim to model large parts of the human auditory system.

4.1.1 Fletcher’s power spectrum model

The simplest way of predicting masking thresholds would be to consider the frequency spectrum and level of the signal and masker, and calculate the relative proportion of signal and masker passed through each auditory filter. Fletcher’s power spectrum model, introduced in section 2.2, takes this approach. In the power spectrum model the auditory filter with the greatest SNR is identified and subsequently used to determine whether the signal is masked via eq. (2.1) on page 12.

As previously discussed, however, the power spectrum model of masking is based on results from tone in noise experiments and the application of the model to complex, ecologically valid scenarios assumes that the auditory filter will be centred on the peak
signal component, that masking is entirely driven by the SNR at the output of the auditory filter, and that temporal effects are not important (i.e. the model is based on the long-term power spectrum of the signal and masker). Since in complex, ecologically valid listening scenarios these assumptions are likely to be violated frequently, the power spectrum model is unlikely to predict masking thresholds for these programmes very accurately. More recent models of masking threshold prediction, however, have built on the core ideas of the power spectrum model.

4.1.2 Zwicker’s loudness model

The model described in (Zwicker 1977) builds on the ideas of the power spectrum model. Although it is strictly a model for calculating the loudness of temporally variable sounds, as discussed in section 2.4 auditory masking phenomena can be considered a set of scenarios which fall within the bracket of loudness phenomena, including cases where the loudness of a signal are diminished or reduced to less than the JND (and are thus inaudible). This model, therefore, can be used to predict the loudness of a stimuli presented in isolation, and this value can be compared with the prediction of the loudness of a pair of summed stimuli. If the difference between the two predictions is less than the JND, then the second stimulus can be considered masked by the first.

In the model a single masking threshold shaped like a delta is placed on the frequency spectrum for each critical band. Each delta has a fixed low frequency slope, and a high frequency slope which varies depending on masker level (see fig. 4.1). It is suggested that although the auditory system acts as though it contains 640 auditory filters the processing requirement of such a system would usually be impracticable. For this reason Zwicker (1977) recommends the use of 24 auditory filters, where each is centred such that adjacent filters have coinciding upper and lower frequency cut-offs. In this way the entire perceptible auditory frequency range will be covered (see fig. 4.2). Combining this with the power spectrum model, a simple device can be made to detect the audibility of the signal within each critical band.

An updated and extended version of this model is found in Zwicker and Fastl (1990), where it is suggested that the overall loudness of a complex stimulus is given by the integral of the specific loudness for each critical band:

\[
N = \int_{0}^{24\text{Bark}} N' \, dz
\]  

(4.1)

Where \( N \) is the overall loudness in Sones, and \( N' \) is the specific loudness as a function of the critical band rate \( (z) \) (i.e. the loudness per critical band), and is given by:
Chapter 4: Masking Prediction

**Figure 4.1:** A simple model of simultaneous masking.

**Figure 4.2:** An example of the critical band rate scale, which is the selective placement of auditory filters across the range of auditory perception to model audition in a pragmatic way.
Chapter 4: Masking Prediction

\[ N' = 0.08 \left( \frac{E_{TQ}}{E_0} \right)^{0.23} \left[ \left( 0.5 + 0.5 \frac{E}{E_{TQ}} \right)^{0.23} - 1 \right] \] (4.2)

Where \( E \) is the excitation level caused by the stimulus, \( E_{TQ} \) is the excitation threshold in quiet, \( E_0 \) is the reference excitation that corresponds to the intensity \( 10^{-12} \text{ W/m}^2 \). Excitation level is also an intensity and thus has units of \( \text{W/m}^2 \).

Using either of these two more sophisticated models, therefore, the loudness of the stimuli can be calculated. At this stage there is still no consideration of the temporal effects of masking, however, recent models have built further on these ideas.

4.1.3 Short-time partial loudness model

Based on the work of Zwicker and Fastl (1990), Moore et al. (1997) devised a model for the prediction of thresholds, loudness and partial loudness. This model requires an input specifying the spectrum of the sound, and was only designed for steady state sounds. It was later extended and revised into what is referred to in this document as the Sound Term Partial Loudness (STPL) model (Glasberg and Moore 2005), which requires an input waveform specification and is designed for temporally variable sounds.

Figure 4.3 shows an overview of the operation of the STPL model. The monaural input signal and interferer are filtered using a fixed filter which mimics the effect of the outer and middle ear. A Fast Fourier Transform (FFT) of the signal is then conducted six times in parallel, each with a different window length in order to provide high temporal resolution for both high and low frequencies. The windows are then shifted by 1 ms and the FFTs are repeated for the entire signal. The output of each FFT contributes only to a pre-specified frequency range, and other results are discarded. An excitation pattern is then calculated using an auditory filterbank of roex filters spaced at 0.25 ERB intervals.

The Instantaneous Partial Loudness (IPL) is calculated from a formula relating the excitation pattern of the signal to the excitation pattern of the interferer and the excitation pattern of the sum of the two. This formula also performs a number of mathematical translations in order to account for the compressive response of the basilar membrane for mid-intensity signals. This series of equations is described in Moore et al. (1997). Next, the temporal integrator averages a number of IPL values together to give an indication of the STPL. Finally the mean STPL value is calculated, which is considered to be indicative of the long term loudness of the signal.

In line with the persistence hypothesis (discussed in section 2.3.2), the STPL model uses a temporal integrator to smooth the representation of the stimuli over time such that all signals inherently increase the masking threshold of subsequent signals. A previous
Figure 4.3: An overview of the STPL model.
study by Oxenham and Moore (1994) used a temporal integrator to test a forward masking paradigm in a similar way. The results indicated a good agreement between measured and predicted data (see fig. 4.4), with predictions generally well within 3 dB of measured thresholds.

The STPL model was designed with the purpose of predicting the audibility of warning signals in real-world environments. As such it has been tested with a wide variety of realistic, and sometimes informationally rich, time-varying maskers. Figure 4.5 shows a plot of the accuracy of the STPL model in tests conducted by (Glasberg and Moore 2005). In these tests masking threshold predictions differed from measured masking thresholds by, on average, 3 dB. The correlation between predicted and measured masking thresholds was 0.94.

The model requires signal, masker and signal plus masker waveforms as its input. It also requires the specification of the conditions of presentation (e.g. headphone/loudspeaker characteristics), and a reference level on which to base the input translation from waveform to levels. The model does not account for the effects of binaural unmasking or CMR, although these phenomena could potentially be accounted for by including processes from other models. Predictions of uncertainty and elements of excess masking caused by uncertainty are not accounted for, although the prediction of listener uncertainty has significant inherent problems since it assumes a priori knowledge of the listener’s experience.

4.1.4 Computational auditory signal-processing and perception model

Another recent model which broadly models the human auditory system is the CASP model of Jepsen et al. (2008), based on the earlier model of Dau et al. (1997). A flowchart depicting the operation of the CASP model is shown in fig. 4.6.

The basic operation of CASP model is as follows:

1. In the first stage of the model, an audio signal is submitted as the input and is filtered using a fixed transfer function. This function mimics the effect upon a sound wave as it passes through the outer and middle ear. The effect of the outer ear is to filter the sound wave as it passes by the pinna, and through the auditory canal and tympanic membrane (ear drum). The effect of the middle ear is that of a mechanical impedance change (resulting in frequency dependent amplification of the sound wave) as the sound wave is transferred through the ossicles.

2. The second stage of the model comprises a non-linear filter bank known as the DRNL filter (see section 2.2.5) which performs frequency selectivity; thus the signal has a two dimensional matrix representation with axes of time and frequency. This function mimics the effect of the basilar membrane filtering.
Chapter 4: Masking Prediction

After the sound wave passes through the ossicles it enters the oval window to the basilar membrane which, due to its shape being tapered, vibrates with differing intensities along its length depending on resonance due to the frequency of the sound wave.

3. The third stage of the model comprises half-wave rectification followed by low-pass filtering at 1kHz. This mimics the hair-cell transduction, i.e. the transition from vibrations in the basilar membrane to receptor potentials in the inner hair cells.

4. A fourth stage of the model simply squares the output of the hair-cell transduction stage. This stage is included to mimic the square-law relationship between the neural response rate and input level.

5. The fifth stage, adaptation, consists of five feedback loops in series. Within each loop the signal is divided by a denominator determined by the low pass filtered response of the previous samples of the signal; thus the feedback loops have a time-adaptive quality. The net-effect of the five feedback loops is a stationary characteristic similar to logarithmic compression for low frequency signals, and relatively linear for high frequency signals. The time constants of the loops range from 5 to 500 ms. This stage of the model is intended to account for forward
Figure 4.5: A scatter plot comparing measured listener's masking thresholds with those predicted by the STPL model. The various shapes indicate the different types of stimuli used, and the dotted line indicates where points would lie if the model perfectly predicted masking thresholds. From Glasberg and Moore (2005)
Figure 4.6: Overview of the operation of the CASP model. adapted from (Jepsen et al. 2008).
masking effects; the feedback loops account for the change in masking threshold over time. This approach therefore assumes the neural adaptation hypothesis of forward masking.

6. The sixth stage of the model comprises an initial low pass filter with a cutoff frequency of 150 Hz, followed by a modulation filterbank applied to every frequency channel; thus the signal is split into a three dimensional matrix of time, frequency, and modulation frequency. The modulation filterbank has filters at 2.5 Hz, 5 Hz, 10 Hz, and spaced logarithmically thereafter with the highest filter at whichever is the lower value of 1 kHz and one quarter of the centre frequency of current channel. So for example, in a 1 kHz centred frequency channel the maximum modulation filter would be at 250 Hz, whereas in an 8 kHz centred frequency channel the maximum modulation filter would be limited to 1 kHz. This stage of the model accounts for the sensitivity of the auditory system to amplitude modulations and the related masking data, but it does not mimic any particular physiological structure or behaviour.

7. In the final stage of the model Gaussian shaped noise is added to the output of the modulation filters and this signal is known as the ‘internal representation’. Subsequently an optimal detector (based on a template of the signal) is used to determine whether the sound was masked. More specifically, all previous stages of the model are completed three times separately: once to produce an internal representation of the masker alone, once to produce an internal representation of the combined target and masker programmes, and once to produce an internal representation of the masker and suprathreshold target programme (i.e. the target programme is amplified to some level at which it is will definitely be audible). The internal representation of the masking signal alone is subtracted from each of the two other internal representations, and the cross-correlation coefficient of these signals is calculated. The resulting value is a decision variable which is expected to relate monotonically to the probability that the target programme is audible.

The CASP model makes predictions which account for simultaneous and forwards masking phenomena, although it does not specifically account for backwards masking phenomena. In addition to these CASP uses a modulation filterbank technique, based on that of (Unoki and Akagi 1997), to consider the effects of CMR. Like the STPL model, the CASP model requires all target and non-target signals as inputs in order to produce the various internal representations.

Since the CASP model operates by performing cross correlations between intervals and a supra-threshold template, a template level which is known to be above the masking threshold must also be chosen. While this may seem slightly paradoxical, because the
masking threshold is not known before the model is run, in practice any level well above the interferer level can be chosen to all but guarantee audibility. As a default setting Jepsen et al. (2008) suggested a template level of 10 dB SPL above the interferer level.

The CASP model accounts for forward masking by modelling an alternative to the persistence hypothesis modelled in STPL: neural adaptation. An adaptation module, positioned after the DRNL filterbank, consists of feedback loops which contain low pass filters and division operators. In this way the instantaneous signal is continuously summed with the resultant modifications of the temporally prior signal. Figure 4.7 shows the results indicating the accuracy of the CASP predictions of on and off-frequency forward masking.

The prediction accuracy of the CASP model for simultaneous masking paradigms was tested in (Jepsen et al. 2008). For the simplest case of tone masked by noise results were very accurate (see fig. 4.8). Predictions generally fell within 2 dB for all but very brief (less than 15 ms) duration signals.

Additionally, CASP was shown to predict forward masking phenomena with high accuracy; this can be seen by the similarity between measured and predicted thresholds in fig. 4.9.

The CASP model was found to be accurate to within 2 dB for a wide range of test conditions, only failing to meet this requirement for certain extreme conditions (e.g.
signals presented for less than 15 ms, or strong beating cues between signal and masker). Jepsen et al. (2008) point out that certain cases of modulation depth discrimination are not considered by the CASP model, and that an internal noise at the output of the modulation filter could be used to account for this limitation.

While the CASP model was tested for a wide range of masking phenomena, such as forward masking and amplitude modulation paradigms, these conditions used stimuli such as tones and noise, rather than music or speech, and it is not clear how the use of these stimuli will affect the accuracy of the model.

### 4.1.5 Component of relative entropy model

Lutfi’s (1993) model of spectral uncertainty, known as Component of Relative Entropy (CoRE), has been found to be successful at predicting the masking threshold caused by the presentation of complex tonal patterns. The model works on the following principle: component discriminability in an unfamiliar tone pattern is a linearly increasing function of the component’s relative entropy. The model is a good predictor of spectral uncertainty for complex tonal patterns but it is not entirely clear how such a model could be applied to complex musical scenarios involving both familiar and unfamiliar sounds, or how the listener’s unfamiliarity should be estimated in the absence of listener profiling.

That uncertainty masking is so difficult to predict is offset by how rarely it will be an important feature of sound zoning scenarios. While there may occasionally be
Figure 4.9: A comparison between measured and predicted forward masking thresholds as predicted by the CASP model. The open symbols are measured thresholds and the closed symbols are predicted thresholds. Adapted from Jepsen et al. (2008).
scenarios in which listeners are unable to recognise the audio programmes, the large
effects of uncertainty masking found in the experiments discussed in section 2.6.1
were all obtained under highly contrived circumstances (e.g. such as randomised
tone complexes). These effects were found to largely disappear when stimuli had
strong harmonic components, were recognisable to the listeners, or had some spatial or
temporal cues for discriminability; all of these are highly common within the considered
ecologically valid listening scenarios. As a result, uncertainty masking can be considered
a rare occasion which need not be predicted for this work.

4.1.6 A note on modelling binaural unmasking

None of the previously mentioned models explicitly account for binaural unmasking,
although they could all be variously modified to incorporate such effects. Perhaps
the simplest way of predicting binaural unmasking would be to use data gathered
from experiments, such as that in (Bronkhorst 2000), to estimate an adjustment
to a monaurally predicted masking threshold. Such a model would effectively be a
calculation or estimation of the azimuth parameter, followed by a simple look-up table
operation. While the strength of this approach lies in its simplicity and mapping of
known psychoacoustic data, it does not involve analysis of the specific signals being
tested, and thus may be less accurate than models which attempt to mimic the human
auditory processing of binaural stimuli.

Another method would be to use the Contralateral Inhibition (CI) model of Breebaart
et al.’s (2001). The CI preprocessor is based on an early version of CASP, and thus the
two could easily be made to work in tandem. The CI model works by using a matrix of
Excitation-Inhibition (EI) elements with delays and attenuators spaced between each
(see fig. 4.10). The result of this arrangement is that each EI element describes a
specific combination of Interaural Time Difference (ITD) and ILD. The EI element
which outputs the greatest stimulation can be used to indicate the lateralisation of the
signal. This process is carried out for every auditory filter at the output of the CASP
model.

The model was tested over a very large range of experimental conditions and was
generally accurate to within 1 dB, and in rare cases accurate to within 5-10 dB.

Complex systems, such as Breebaart et al.’s (2001) CI model, can be implemented to
predict BMLD, or a much simpler look-up table approach could be utilised instead.

4.1.7 Selection

Table 2.1 gives an overview of the key features of the discussed auditory models. When
considering the available auditory processing models the CASP and STPL models seem
Figure 4.10: Binaural processing using the CI model. In this model the left and right signal paths travel in opposite directions, such that each EI element is indicative of a characteristic ITD and ILD. By finding the EI element with the maximum value, an ITD and ILD is known, and thus a localisation can be inferred. Adapted from Breebaart et al. (2001).

Figure 4.11: The prediction models discussed in the previous sections are listed in the above table, and the masking phenomena they account for are shaded in green. From left to right the phenomena listed are: simultaneous masking, forward masking, backwards masking, excess masking, binaural masking, comodulation masking release, and uncertainty masking.
most appropriate for use because they attempt to model the human auditory system and hence account for a wide variety of masking phenomena (see fig. 4.11). Thus it would be logical to use either of these models for the prediction of masking thresholds, with a view to modifying them as necessary if binaural, or other, phenomena are to be considered. The CASP model was selected for use since the STPL was not available to the author. In any case, the CASP model may be marginally preferable since it can account for CMR and could be adapted to include the binaural CI model if required at a later date. There is also no reason why simpler binaural models like that of Bronkhorst (2000) could not also be implemented if necessary.

4.2 Implementing the masking model

A variety of masking threshold prediction models were discussed in section 4.1. The CASP model was selected for implementation and this section outlines this process.

4.2.1 Some complications with computational masking models

Before summarising the model selections a few general points should be noted about computational masking models.

Firstly it should be considered what exactly is meant by the masking threshold that a model returns. While this might appear simple to answer by referring to the definition of masking thresholds given in section 2.1, the issue is confounded somewhat by individual differences in hearing between listeners. For a randomly selected sample of normal hearing listeners taking part in a masking experiment, the expected results would be a range of masking thresholds. This presents a challenge, however, because a model is required to produce a single value as a masking threshold. Ideally a model should select a masking threshold such that when subjecting a randomly selected normal hearing listener to the same stimuli used in the experiment the audibility of the target signal could be accurately predicted, therefore the masking prediction should be close to the mean. If there is wide disagreement between listeners about the masking threshold for a specific listening environment, this is likely to be indicative that the complexity of the listening scenario is such that other metrics may be of more interest.

Another notable issue regards what is referred to as the ‘internal variance’ of the listener in (Jepsen et al. 2008). Computational masking models which mimic aspects of the human auditory system sometimes detect masked signals more easily than human listeners, even when the peripheral aspects of the auditory system are modelled very accurately. This is usually assumed to be a result of human cognitive errors, sometimes referred to as ‘physiological noise’. Some models (e.g. CASP) incorporate this element of predicting human perception by introducing low level noise into the internal
representation of the signal reference (i.e. an internal variance) or by simply diminishing the detection results. The latter approach may be selected for computational efficiency where its effect is analogous to the former.

4.2.2 Modifications to CASP

The CASP model makes masking threshold predictions by passing a known target and interferer through a series of processes which mimic the response of the human auditory system. The mixed target and interferer are transformed into an Internal Representation (IR) of the signal that is then analysed for correlation with a template IR which is based on the same combination but with the target presented at a high level (and thus known to be audible). The process is repeated to find the correlation between the mixture and the interferer presented in isolation, and the difference between these two correlations is used to calculate a probability of detection via an optimal detector (as described in Green and Swets (1996)). This is performed for the left channel, the right channel, and a summed channel (which produces a simplistic binaural detection known as ‘best ear’). For a detailed description of the CASP model see (Jepsen et al. 2008). Three modifications were necessary in order to adapt the model for the task considered in this report.

Converting probabilities to thresholds

The first modification addressed the output of the CASP model which is a probability of detection for pre-specified signals, rather than a masking threshold. Since the CASP model reports the probability of detection for a specified target and interferer, there is no way to directly infer from this the interferer level for a specific probability. In order to find the masking threshold, the model was run repeatedly with the interferer level adjusted just prior to each run. In this way the interferer level which corresponds to the pre-specified probability could be identified. The first implementation utilised a fixed distance between interferer levels (1 dB) and a fixed minimum and maximum interferer level (20 to 70 dB LAEq). It was found that although smaller signal level increments produced greater resolution, the processing time was significant and an inaccuracy of up to ±0.5 dB was considered acceptable.

Using this method, an interpolation between levels should be considered to approximate the level corresponding to the desired probability on the psychometric function (the curve which describes the relationship between input stimulus and listener response). When the interferer level is incremented sequentially there is sufficient data to perform linear interpolation between levels. Alternatively a curve matching algorithm, which may be more accurate, could be utilised but since the difference in accuracy would be proportional to the level increment it is likely to be substantially less than half the level increment (< 0.5 dB).
Chapter 4: Masking Prediction

**Figure 4.12**: A binary search algorithm implemented to decrease the number of times it is necessary to run CASP before finding the masking threshold. $P_D$ is the probability of detection returned by the CASP model, $P_T$ is the 'test probability', $P_A$ is the 'accuracy probability', $L_T$ is the test level, $L_I$ is the increment level.

While the psychometric functions produced by this series of repetitive calculations may be useful for some purposes, the processing time required was still great. In order to reduce the processing time a simple binary search algorithm was implemented as shown in figure 4.12. The CASP model was run with an interferer set to an initial ‘test level’ ($L_T$) of 20 dB LAeq. The Probability of Detection ($P_D$) returned by the CASP model was then tested against the Test Probability ($P_T$) (set here to 0.5) to see if it is sufficiently close to the Accuracy Probability ($P_A$) which is the minimum inaccuracy allowed to complete the algorithm. If the $P_D$ was not close enough to the $P_T$ the interferer level was increased or decreased by the Increment Level ($L_I$) depending on whether the probability of detection was too low or too high. The $L_I$ was then halved and the process was repeated until the $P_D$ was sufficiently close to the $P_T$. This binary search algorithm ensures that a predicted masking threshold can always be obtained within a fixed number of iterations of the masking model. With the $P_A$ set to 0.01, the algorithm was always completed within eight iterations of the CASP model with a maximum inaccuracy equivalent to 0.16 dB. Thus the binary search algorithm improved the accuracy of the selected masking threshold prediction (over linear interpolation) while severely decreasing the processing time (more than 6 times faster). It also allowed for a limited trade-off between accuracy and processing time by adjusting the value of $P_A$. The disadvantage of the binary search algorithm is that the CASP model was not run at every level between the minimum and maximum specified levels, so any psychometric function interpreted from the data will be extremely coarse. This is only slightly problematic in that the prediction of the gradient of the psychometric function may be used to estimate the spread of data of the masking thresholds, however this measure was considered of secondary value in comparison to the mean masking thresholds themselves.
Calibrating the model

The second modification addresses the calibration of the model. In Jepsen et al. (2008) it is suggested that the model should be calibrated (by modifying the value of the internal variance) using a tone intensity discrimination task. This calibration was found to produce overly sensitive predictions (i.e., signals were predicted to be much too easily detected than the results from the experiment described in chapter 3 indicated), so the cross correlation value was divided by a constant which was empirically derived to give a more accurate set of results. The value was determined such that average observed mean masking thresholds were most closely aligned with the average detection probability of 50% for the samples under test (although in practice a different detection probability could be selected), thus the model performs optimally for the data set available, but a different calibration could be used for different data sets.

Temporal windowing

The third modification addresses the temporally variable nature of the programmes. Because the target and interferer programmes are spectrotemporally complex, it cannot be assumed that a single, meaningful masking threshold can be isolated for the entire duration of a programme. In order to address this, the programmes are split into many short temporal windows before processing. Once the levels which correspond to a probability of detection of 50% have been calculated for each temporal window, the temporal window with the highest TIR is selected as the masking threshold. It is apparent that selecting an appropriate duration of the temporal windows is an important consideration which affects the resultant masking threshold prediction. To make an appropriate selection some understanding of the task performed by the listener in an interference scenario is required.

Understanding precisely what task the listener is performing, and the limitations of the listener's capacity to perform that task, is one of the greatest challenges to the prediction of auditory masking within interference scenarios. In the experiment outlined in chapter 3 listeners were asked to adjust the level of the interferer until it was just inaudible, but it is not known precisely how the listeners achieved this. It is assumed that listeners completed their task by attending to the entirety of the 10 second programme, isolating the section wherein the interferer was most easily detectable, and adjusting the level until the interferer was no longer audible in that section. Some listeners reported that this description closely resembled their behaviour. Even if this assumption is correct, however, the duration over which such judgements were made is unknown.

Yabe et al. (1998) suggested that the temporal integration time of the auditory system is about 160–170 ms, although this value is likely to be frequency dependent (Wassenhove et al. 2007). Even if this response time is known precisely the audibility of the interferer
could be affected if, for example, the section in which the interferer was most easily audible was immediately preceded by a section in which the target was especially loud. Such a condition could cause the listener to doubt the previous detection even though an optimal detector would report the interferer as audible. (Baykaner et al. 2013) describes a study conducted to address this question for the CASP model, and the optimal time windowing solution was found to be using time frames of 400 ms stepping forwards in 100 ms increments.

There are many other ways of interpreting the set of masking predictions. In some systems an averaging process is used in favour of the selection of a single temporal window (Glasberg and Moore 2005). In a similar manner it would be possible to average a subset of the lowest predictions, take running averages, or select a temporal window only if it does not precede a very low TIR. Although in this work these alternative interpretation of masking predictions were not utilised, it is worth noting that a study into such methods could potentially reveal even more accurate approaches than the method adopted here.
4.3 Evaluation of masking threshold predictions

Both R and RMSE were introduced in section 3.3.2 as measures for prediction accuracy. A useful extension to the RMSE is the Epsilon-insensitive Root Mean Squared Error (RMSE*). It describes the average additional error after accounting for subjective variance (in this case that means inter-subject disagreement). The RMSE* is calculated with:

\[
RMSE^* = \sqrt{\frac{1}{n-k} \sum_{i=1}^{n} (\max(0, |Y_i - X_i| - CI_{95}))^2}
\]  

(4.3)

where \(Y_i\) and \(X_i\) are the observation and prediction for trial \(i\), and \(CI_{95i}\) is the 95% confidence interval of the subjective scores for trial \(i\). The confidence intervals, based on the normal distribution, were calculated using:

\[
CI_{95i} = 1.96 \times \frac{\sigma}{\sqrt{n}}
\]

(4.4)

where \(\sigma\) is the standard deviation of the masking thresholds for each trial.

In this way the RMSE* is calculated as the sum of the squared error for all trials where the predicted acceptability falls outside the 95% confidence interval, with the error for all other trials being set to 0. The RMSE* will therefore always be less than or equal to the RMSE, however the difference between the RMSE and the RMSE* gives an indication of the extent to which the inaccuracy of the model is driven by subjective disagreement.

As well as considering the accuracy of the model predictions, it was also important to consider the robustness of the model to new data, so the 54 programme combinations were split into training and cross validation data sets. In each case the data was split into 38 training items (around 70% of cases) with 16 (around 30%) cross-validation items. Ideally every possible combination of training and cross-validation programmes would be analysed however, since there were \(2.1 \times 10^{13}\) possible combinations, this would be very computationally expensive. Instead 5000 random permutations were analysed with the assumption that this would provide a sufficiently representative sample of these combinations.

The average accuracy, across the 5000 permutations, of the predictions for the training and the cross validation data are both presented. The difference between the two can be considered a measure of the model’s ability to extrapolate to new data.

A Kolmogorov–Smirnov test showed that the 5000 cross validation RMSEs were not
normally distributed so median, rather than mean, scores are reported. The median cross validation RMSE was 3.58 dB, while the median RMSEs for the training data was 3.37 dB. The RMSE’s were 1.78 and 1.87 dB for the training and cross validation data respectively, indicating that the average additional error (beyond the edge of the human listener’s 95% confidence interval) was less than 2 dB.

The difference between the RMSE of the training and cross validation data was therefore 0.21 dB, indicating that the model extrapolates to new cases with little additional error. It may be argued, however, that the CV and T data are not truly independent, since many programmes are filtered replicates, and thus this test of extrapolating to new data is optimistic. As a point of comparison, therefore, the RMSEs for the single cross validation (of the 5000 tested) with the greatest disparity between training and cross validation data, was 2.86 dB and 5.12 dB for training and cross validation respectively, giving a much greater difference of 2.26 dB. Since choosing the single worst case is also an unfair representation, the true robustness of the model could most fairly be represented by stating that extrapolation to new data is likely to result in a loss of 0.21–2.26 dB RMSE.

Figure 4.14 shows a scatter plot of the observed TIRs against the median predicted TIRs. A strong positive correlation was found between the predicted and the observed TIRs of $R = 0.87 \ (p<0.001)$, indicating a relatively linear relationship.

Further analysis revealed that prediction errors were fairly evenly distributed across different levels of the factors: target programme, interferer programme, road noise and filtering, with a few exceptions. Two cases were predicted noticeably worse than the others: those where the target was pop music and the interferer was HPF male speech with and without road noise. The predictions were that the speech would need to be reduced by 8.0 and 7.3 dB, respectively, more than was observed as necessary in order to be masked.

It was also found that the median absolute difference between predictions and observations for classical target programmes and for sports commentary target programmes were 1.99 dB and 2.38 dB respectively, whereas the median absolute difference between predicted and observed TIRs for the pop target programmes was 3.99 dB. The majority of these cases (13 of the 18) were predicted as requiring greater reduction in interferer level than was necessary, with 10 of those cases having an error exceeding 3.5 dB. Notable exceptions were for pop targets and classical interferers for all levels of filtering without road noise, where the model underestimated the reduction in the interferer level which would be required by 5.86, 4.28, and 3.72 dB for unfiltered, LPF, and HPF respectively.

It seems, therefore, that the model tended to overestimate the reduction in interferer level which was necessary when the target was pop music (most severely when the interferer was speech), except where the interferer was classical music, where the model
underestimated the necessary reduction.

4.4 Predicting acceptability by predicting masking

With a model already established to predict acceptability thresholds using known masking thresholds (see eq. (3.2) on page 71), the same approach can be used to make predictions about the acceptability of auditory interference scenarios where the masking thresholds are unknown but where an appropriate masking threshold prediction model is available. Using the set of masking threshold predictions obtained from the modified CASP model the linear regression model described in eq. (3.2) was used to produce a set of acceptability threshold predictions. Figure 4.15 shows the relationship between the mean acceptability thresholds and the predicted acceptability predictions.

The accuracy of the predictions was equal to 4.2 dB RMSE, and 2.7 dB RMSE*, with $R = 0.88$. As fig. 4.15 shows, and these metrics imply, the predicted acceptability scores correlated with the the acceptability scores about as well as the known masking thresholds correlated with the acceptability scores, however these predictions had a constant offset such that they tended to be at TIRs slightly higher than the TIRs for the acceptability scores. This implies that in order to predict acceptability scores the
Figure 4.15: A scatter plot showing mean acceptability and predicted acceptability TIRs. The
diagonal line represents the line $y = x$, i.e. an ideal prediction. Classical target programme
scenarios are indicated with circles, pop target programme scenarios are indicated with triangles,
and sports commentary target programme scenarios are indicated with diamonds.
following adjusted linear regression model should be used:

\[
\text{Acceptability threshold} = (0.685 \times \text{Predicted masking threshold}) + 0.756. \quad (4.5)
\]

The accuracy of the predictions based on this linear regression model were \(= 2.4 \, \text{dB RMSE}\), and \(= 1.2 \, \text{dB RMSE}\). Notably, the slope of the linear regression differs from the linear regression to the subjective masking data by 0.01, whilst the constant offset differed by 3 dB, implying that the predictions are relatively stable.

Further improvements to the accuracy of the model might be achieved by including further categories in specific cases where more details are available. For this dataset, however, while it may be possible to model subsets of the data more closely, the result would likely be an over fit to the data.

4.5 Summary and Conclusions

This chapter posed the research question: “how can auditory masking be predicted?” A variety of masking prediction models were described and compared and the STPL and CASP models were considered to be appropriate starting points upon which to base an evaluation model due to their comprehensive approach to modelling large parts of the human auditory system, and in doing so predicting a wide range of masking phenomena. Due to lack of availability of the STPL model, the CASP model was selected for use. The implementation of the CASP model was outlined and associated modifications included the empirically based calibration, the implementation of temporal windowing, and the use of a binary search algorithm to more quickly determine masking thresholds from probabilities of detection.

The modified CASP model was used to make masking threshold predictions for the masking thresholds obtained in the experiment described in chapter 3. The median cross-validation accuracy was 3.6 dB RMSE and 1.9 dB RMSE*. The median training accuracy was 3.4 dB RMSE and 1.8 dB RMSE*, so the model appears to have excellent robustness to new data.

After this, the application of the modified CASP model to acceptability scores was investigated. In section 3.3 it was shown that there is a relatively consistent difference between mean masking thresholds and acceptability thresholds, and that as a result acceptability thresholds can be predicted using a linear regression to known masking thresholds. This regression model was used to make acceptability threshold predictions based on the predictions of the modified CASP model. Although prediction error increased to RMSE = 4.2 dB, the correlation remained very high \((R^2 = 0.78)\). On observing a scatter plot of the data, it became apparent that the prediction accuracy was
suffering an increased RMSE due to a linear offset. A new linear regression model was constructed based on the masking threshold predictions of the modified CASP model, rather than the subjective masking data directly, and the accuracy of the predictions was 2.4 dB RMSE, and 1.2 dB RMSE*.

This shows a way that auditory masking can be predicted, therefore, but subjects noted differences in acceptability between scenarios featuring speech in both the target and interferer programme, and those which featured speech in only one of these. In addition, it is possible that the approach taken here to the prediction of auditory masking may not perform well for spectrotemporally sparse programmes, such as speech, if the result is a selection of the lowest threshold (which occurs during silent gaps). For these reasons the next chapter investigates the perception of speech, focusing on acceptability for cases where the target programme is speech-based.
Chapters 3 and 4 showed that a relationship exists between auditory masking and acceptability, and described a method for predicting acceptability based on a masking prediction model. The model is not necessarily appropriate for use when a listening scenario features speech as the target programme, however, because the gaps in the speech are likely to produce masking threshold predictions close to the absolute threshold of audibility in silence, even though listeners will often find a low-level audible interferer acceptable. Furthermore, subjects reported that when both target and interferer were speech the listening scenario required a greater TIR before it could be considered acceptable.

Due to these considerations it was decided that further work should focus on scenarios where the target programme is comprised primarily of speech. In the overwhelming majority of situations, the goal of speech is to convey an intelligible message; and thus while other aspects, such as the timbre of the speaker, may be important to the listener, the priority will tend to be ensuring a high speech intelligibility. Based on this, it is reasonable to suggest that the intelligibility of the speech may be an important aspect of acceptability in auditory interference scenarios featuring speech. In addition to this, it is worth considering the relationships between intelligibility, acceptability, and any other relevant measures describing the interference scenario. Thus a research question is posed: “what relationships exist between intelligibility, acceptability, and other relevant measures?”

This chapter outlines a series of experiments conducted to investigate these relationships. Section 5.1 consists of a discussion regarding the role of intelligibility in sound zoning scenarios and provides an introduction to resources for speech intelligibility experiments. Section 5.2 outlines a pilot experiment to determine the appropriate selection of a target corpus, discern the importance of presentation level for intelligibility, and select an appropriate range of SNRs; the results of this are presented in section 5.3. In section 5.4 the subsequent main experiment designed to gather a range of speech intelligibility, acceptability and other relevant measures is described. The results are presented and discussed in section 5.5. Finally, conclusions are drawn in section 5.6.
5.1 Speech intelligibility background

This section gives a brief overview of the potential use of speech intelligibility for the evaluation of auditory interference scenarios, before introducing some resources and methods for conducting speech intelligibility experiments.

5.1.1 The nature of speech as a percept

The nature of speech as a percept is fundamentally different to that of other acoustic cues. It is primarily not the acoustic differences, however, which distinguish speech from other auditory percepts, but the decoding of the acoustic signals by the brain. For this to be so, the brain must first categorise incoming acoustic signals as either linguistic or non-linguistic in nature. The evidence that the brain does this comes from various sources discussed below.

The first experimental evidence for this mode of operation is the categorical perception of phonemes. In (Liberman et al. 1967) it was demonstrated that by manipulating the second formant of a phoneme in gradual shifts, listeners either failed to notice a change or determined that a different phoneme was present. Similar changes to non-linguistic cues usually produce gradual shifts in perception, whereas when the stimulus is categorised as linguistic the perception becomes closer to that of discrete states. A second set of evidence for a distinction between speech and non-speech comes from experiments in which competing stimuli are presented simultaneously to both ears. If the stimuli are both speech, the speech presented to the right ear is generally better identified than that presented to the left, while the reverse is true when the stimuli are melodies (Broadbent and Gregory 1964). Neurological evidence additionally supports this with indications that the left hemisphere plays a primary role in speech perception (Broadbent and Ladefoged 1959). Further evidence is found in (Remez et al. 1981) where signals consisting of three sine tones, which varied over time in a manner consistent with speech, were presented to listeners. Listeners who were told nothing about the stimuli perceived them as music or beeps, whereas listeners who were instructed to transcribe a "strangely synthesised English sentence" were able to do so.

The results of these experiments indicate that speech is perceived as something fundamentally different to other forms of acoustic stimuli. The categorical distinction between percepts which are linguistic and those which are not is sometimes referred to as the ‘speech mode of perception’ (Moore 2004).

5.1.2 The use of speech intelligibility in auditory interference

The most direct use of the speech intelligibility predictions pertains to establishing a lower limit of acceptability. If the intelligibility of the target signal is low (when the
target signal is speech) this will be a strong indication that the listener will not be able to carry out even the most basic task of understanding the information which the target signal attempts to convey. In this case, a listening scenario would be considered unacceptably poor. Conversely, a very high speech intelligibility (close to 100%) does not necessarily mean that the acceptability is high, since the interferer may still be audible, and other percepts like naturalness may also be important. Therefore the speech intelligibility, as applied to the target signal, has a simple but critical role of applying a lower limit to the performance of the sound zoning system. If, therefore, an adaptive sound zoning system attempts to improve the separation between target and interferer but in doing so finds that the speech intelligibility of the target drops appreciably, it may be wise to sacrifice the increased separation and revert to a state in which the target signal has higher intelligibility, or it may be that even greater separation is required. In both cases a more informed decision can be made about the appropriate response by measuring the intelligibility of the target.

Another possible use of speech intelligibility predictions may be to consider the effect of the interferer as a distractor. In (Martin 1988) a series of experiments were conducted to investigate the effect of various auditory distractors on a reading comprehension task. It was demonstrated that for a reading comprehension task an instrumental musical auditory distractor had little effect on performance. When the distractor contained verbal material, whether spoken or sung, the reading comprehension performance decreased. In a subsequent experiment a sequence of meaningful words were shown to have greater interfering effect than a meaningless speech background, which had similar effect to a background of white noise. These scenarios were not auditory interference scenarios, however further evidence corroborates the findings for speech-on-speech interference. Further evidence is found in (Simpson and Cooke 2005), which identifies that speech masking is almost entirely informational when there are fewer than 6 interferers, and in (Calandruccio et al. 2010), wherein it is reported that informational masking was found to be irrelevant when interferer intelligibility falls below 80%. These are all consistent with the theory of the speech mode of perception, wherein an auditory stimulus will be processed in a different manner while it is not recognised as being linguistic in nature. An estimation of the interferer intelligibility, therefore, would contribute to the evaluation of the performance of a sound zoning system by producing a feature describing how distracting the interfering speech is likely to be.

5.1.3 Corpus selection

In order to select a corpus for a speech intelligibility experiment the relevant usage and environment must be considered. In this case target and interferer programmes are likely to be primarily of the following types:
Chapter 5: Speech Intelligibility Experiment

- music, e.g. C.D. audio
- radio
- television/film
- telecommunications speech,

and likely scenarios are those which involve combinations of these programmes. As such, listening scenarios under consideration are likely to contain connected speech which may be conversational (e.g. telecommunications and radio talk shows) or clear (e.g. news reports), and will have a wide range of sentence structures with a set size equivalent to the full vocabulary of the language. Talkers will also have a wide variety of accents and vocal characteristics.

A variety of speech intelligibility experiments have been conducted with various points of inquiry. Some corpora have been designed with specific test procedures in mind in order to gain understanding of specific fields of interest, and so should be considered as pairs. The coordinate response measure (CRM), modified rhyme test (MRT), and speech perception in noise (SPIN), are examples of such tests requiring a specific corpus.

**Coordinate response measure**

CRM, developed by Bolla et al. (2000), is a widely used corpus featuring set phrases, each including a colour-number keyword following a callsign. There are 8 callsigns, 4 colours and 8 numbers, spoken by 8 talkers (2048 sentences in total). CRM has been largely used to test speech-on-speech interference tasks with the callsign allowing the listener to identify the target sentence within mixtures, and speech intelligibility is usually determined by the probability that the listener correctly identifies the colour-number keyword. Brungart (2001) used CRM to investigate the effect on speech intelligibility when the interferer was: the same talker, a different talker of the same gender, a talker of the opposite gender, a multitude of talkers, and a range of noise types. Where the interferer was a single talker, non-monotonic functions described the relationship between speech intelligibility and SNR. However, when the interferer was speech shaped noise (SSN), multiple talkers, or envelope modulated noise, the relationship between speech intelligibility and SNR was monotonic (Brungart et al. 2001). Eddins and Liu (2012) further determined the psychometric functions for the CRM corpus for two-talker, four-talker, and cafeteria noise interferers, and found them to be monotonic.

**Modified rhyme test**

The MRT uses 50 lists, each containing 6 similar sounding monosyllabic, consonant-vowel-consonant (CVC) words. Allowable differences between words in the same
sentence are changes in first or final consonant, e.g. cat, bat, and cap could appear in the same list where ‘cat’ is the target word. The listener must identify the target word (usually within a carrier sentence) by selecting from the provided list. This procedure, tested in House et al. (1963), revealed no learning effect for listening exposed to the test programmes for 30 consecutive days. This procedure is useful for audiologists, who can use the specific nature of the confusions discerned for diagnostic purposes, and more generally for the evaluation of communication channels.

**SPIN**

Speech perception in noise (SPIN) was developed by Kalikow et al. (1977) to investigate the extent to which listeners infer unintelligible words from sentence context. This procedure involves lists of paired sentences: a high probability sentence (where the context gives a cue to the word) and a low probability sentence (where the context gives little or no cue to the word) in each pair. The differences between the scores is used to give an indication of the relative importance of contextual cues.

**Interim discussion**

In the sound zone scenarios under consideration it is unlikely that listeners would receive cues to the target programme which are as direct as a predetermined callsign, however some level of expectation about the target programme (such as timbre of the speaker’s voice, or contextual relevance) is likely to act as a persistent cue. CRM-style procedures are therefore not ideal for use in this experiment, but a familiarisation stage in which the listener is able to learn the pertinent cues to the target programmes would likely be valuable. MRT-style procedures, although resistant to learning effects and useful for fine identification of the specific confusions involved in speech-speech listening scenarios, provide the listener with a multiple-choice style solution which is dissimilar to events within ecologically valid listening scenarios. Likewise SPIN procedures provide a level of detail about the effect of context within sound zones which is not necessary for this work; while it is interesting to note that context plays an important role in intelligibility, the listening scenarios under consideration will often provide such context, so its removal is not desirable in this experiment. Other corpora should therefore be used for this speech intelligibility experiment.

**Harvard sentences**

The ‘IEEE Harvard sentences’ (hereafter simply referred to as ‘Harvard sentences’) are worthy of note as a long list of phonetically balanced phrases (i.e. the relative proportions of the phonemes in the entire list are similar to that within general usage of the language). These phrases are the test material recommended for testing speech quality in IEEE Speech (1969). Ordinarily, the Harvard sentences are recorded at commercial telecommunication quality (up to 8 kHz). Rodman (2006) noted that
consonants are far more important than vowels in English (and many other languages), and that many consonants are differentiated using high frequency energy which is lost in band limited speech such as that of telecommunication quality recordings. Sentences recorded in this way are therefore of interest for usage of telecommunications devices in sound zones, but are an unsuitable representation of broadcast and C.D. quality speech. Several lists of recorded Harvard sentences are freely available from the Open Speech Repository (http://www.voiptroubleshooter.com/open_speech/), recorded at 8 kHz with 8 talkers (3 British English men and 5 American English men and women). Another set of recordings by a single, British English male speaker is available from McCorry (2011) recorded at 44.1 kHz.

Dantale 2

Another potential source of sentences is the Dantale 2 corpus, described in (Wagener et al. 2002). Dantale 2 is a Danish corpus with rigid phrasings of the form ‘name’, ‘verb’, ‘numeral’, ‘adjective’, ‘object’ (e.g. ‘Anders had seven new flowers’). This rigid phrasing is somewhat similar to that of CRM, but the randomised semantics and lack of a keyword allow the sentences to be less contextually predictable. Sentences were recorded by a single female talker at C.D. quality (44.1 kHz and 16 bits).

GRID

Similar to Dantale 2, but recorded in English, the Grid corpus has rigid phrasings, with the fixed structure ‘command’, ‘colour’, ‘preposition’, ‘letter’, ‘digit’, ‘adverb’ (e.g. “place green by A 3 please”). Grid has a significant benefit over the previously mentioned corpora, however, in that it contains over 1000 sentences per talker, and features 34 talkers (18 men and 16 women) with a variety of accents. Gender and accent differences have already been shown to be significant variables within speech intelligibility scenarios (Brungart et al. 2001; Calandruccio et al. 2010; Barker and Cooke 2007) and it is important to minimise the effect of these by using a range of both, randomised across test conditions. Additionally, the large vocabulary of sentences ensures that, while the grammatical structure is fixed, the specific phrases should not be predictable. The Grid corpus is freely available from http://spandh.dcs.shef.ac.uk/gridcorpus/.

Selection

Firstly it should be noted that other speech corpora are available from the field of speech recognition, of which TIMIT and NIST are examples. These corpora, however, are similar to the Harvard sentences, which is preferentially selected for this work since it has seen prior use in speech intelligibility work, e.g. Hawley et al. (2004); Bent et al. (2009); Lavandier and Culling (2010).

As such, the listening scenarios under consideration are likely to contain connected speech as the material of interest, so monosyllabic corpora such as that used in MRT are
unlikely to be ideal test material. While callsign-keyword phrases such as those of CRM use connected speech, the fixed sentence structure offers greater contextual cues than would be expected in the considered listening scenarios, which might produce slightly improved speech intelligibility scores than would be expected in the listening scenarios. Experiment results might therefore indicate (Speech Reception Thresholds (SRTs)), at which intelligibility is equal to 50% which are slightly lower (i.e. the target is easier to understand) than when using phrases of arbitrary linguistic structure where no specific cues to the keywords are given. The Grid corpus is similarly semantically predictable to CRM, although it is far more extensive, including a wide range of talker accents. Dantale 2 somewhat diminishes these contextual cues by using grammatically correct but semantically arbitrary sentences. The Harvard sentences are yet more ecologically valid, being both grammatically and semantically appropriate. It is not clear whether there will be appreciable differences between use of the Grid corpus and the Harvard sentences in this speech intelligibility experiment because the interferer programmes (ecologically valid music, speech, and mixed programmes) are rarely used in the literature. It has also been established that 'clear' speech is more intelligible than 'conversational' speech (see Amano-Kusumoto and Hosom (2011) for a review), yet all of these corpora are comprised of phrases which lend more readily to the former description. Finally the corpora from the field of speech recognition are similar to that of the Harvard sentences, and could be used in such a speech intelligibility experiment. Since they have not been utilised in speech intelligibility experiments in the literature, however, the Harvard sentences are selected preferentially.

Ideally the speech intelligibility test would use high quality recordings of multiple male and female speakers with various accents repeating both clear and conversational speech. Such a corpus is not readily available, although something close to this could be constructed by making recordings of the Harvard sentences if necessary. Other reasonable options may be to use the Grid or Harvard (Open Speech Repository) corpora, although usage of these may diminish ecological validity of the results. The two following questions are thus posed and investigated in a pilot experiment:

1. Is there a difference between speech intelligibility of high quality and low quality recordings of the Harvard sentences within auditory interference scenarios?

2. Is there a difference between speech intelligibility of high quality Harvard sentences and GRID sentences within auditory interference scenarios.

5.1.4 Calculation of intelligibility scores

In order to analyse intelligibility scores based on subject responses from listening experiments, the intelligibility scores must be calculated based on interpretations of the subject's responses. Intelligibility scores can be very simply calculated as:
Chapter 5: Speech Intelligibility Experiment

\[ I = \frac{C}{N}, \]  

(5.1)

where \( I \) represents the intelligibility score, \( C \) represents the number of correct words (i.e. the number of words in the subject response which match those in the target phrase), and \( N \) represents the total number of words in the target phrase.

When corpora with limited test sets are used, such as the CRM, some consideration must be taken of the probability that listeners will guess correctly, despite not hearing the work. This work, however, is focused on more ecologically valid scenarios, where the practical test set is equivalent to the vocabulary of the speakers. This vocabulary will ordinarily be so large that random guessing would be an extremely poor strategy, and would have a negligibly small effect on the results if utilised.

There are heuristics, however, which enable listeners to make contextually informed guesses, and accounting for these cases is more difficult. One way to attempt to minimise the effect of such contextual guesses is to calculate intelligibility scores by only considering those words which are key to the informational content of the sentence. This can be done by pre-specifying the keywords in the target phrases and then calculating keyword intelligibility as:

\[ I_K = \frac{C_K}{N_K}, \]  

(5.2)

where \( I_K \) represents the keyword intelligibility score, \( C_K \) represents the number of correct keywords, and \( N_K \) represents the total number of keywords in the target phrase.

The GRID corpus features 3 keywords per sentence (colour, letter, and number), whereas the Harvard corpus features 5 key words per sentence with the 5 key words being those which carry the informational content (e.g. the girl at the booth sold fifty bonds, where the keywords are emphasised).

There are some benefits and disadvantages to using the keyword intelligibility score. Firstly, the length of the subject responses do not necessarily match the length of the answer response, so it is possible for very long subject responses (which do not match the answer well) to include many of the correct low-information words (such as prepositions), and therefore have an artificially high intelligibility score. In this situation keyword intelligibility scores will be relatively unaffected because keywords are generally not comprised of common prepositions (which tend to carry less information) and so the length of the subject’s response becomes mostly irrelevant. In general, however, for relatively short phrases, such as those used in this experiment, any improvements to intelligibility scores achieved in this way tend to be small and infrequent where subjects perform their task correctly. Secondly, keyword marking
compresses the intelligibility scores across conditions (because fewer words are marked) and so may hide significant differences which the intelligibility score reveals. Finally, the keyword intelligibility score effectively gives a closer indication of speech understanding than speech intelligibility, since only those words which are considered to hold important information are marked. As such, both intelligibility and keyword intelligibility are of interest because the former measures a feature describing words ‘intelligible’ to the listener while the latter measures a feature describing phrases for which the key information was intelligible, and therefore likely to be ‘understood’ by the listener.

5.1.5 Response method

The calculation of intelligibility scores depends, in part, on the response method of the listening test. Tests may be designed such that listeners indicate the word presented to them, either by speaking or typing the words they hear, or by some other means such as pressing a button whenever an agreed word occurs. This latter response method changes the task the listener is performing, however, from one wherein a listener openly interprets all speech presented, to one where the listener attempts to match a prespecified word to the words being spoken, and is arguably less ecologically valid. Whether listeners repeat or type the speech they hear makes little fundamental difference to the results, although there are practical advantages to each method. In favour of speaking, it may be faster for listeners to repeat, rather than type, the words heard. Additionally, listeners may produce typographic errors in typed responses. In favour of typing, however, listeners can carry out listening tests without an experimenter present, and communication errors are still possible when a listeners repeats heard speech. Additionally, results are required to be typed at some point prior to analysis, so allowing listeners to type their responses is more efficient.

In this work, the response method selected was that listeners would type their responses for matters of practical efficiency.

5.1.6 Application to sound zoning systems

Although some of the previously mentioned procedures and corpora have been used to gather speech intelligibility data in listening scenarios similar to those under consideration here, the author is not aware of any studies which directly investigate speech intelligibility for a musical interferer. This may be because it is assumed that such scenarios, where the music contains no vocals, will involve no target-interferer confusion and will therefore be entirely within the domain of masking paradigms i.e. the intelligibility of the target is degraded only where it is partially masked by the interferer, and it has been shown that energetic masking can be effectively utilised as a feature in the prediction of speech intelligibility (Barker and Cooke 2007).
interfering music contains vocals it may be assumed that under most circumstances the vocals will be sufficiently different from the target speech, in rhythm and timbre, that little confusion is likely to occur and the paradigm remains one under the purview of masking. If this assumption is correct, listening scenarios featuring a speech target and interferer would be expected to be more objectionable than those featuring a speech target and music interferer (a point reported by subjects in the previous experiment). More specifically, the threshold of acceptability would be expected to be close to the masking threshold if the target and interferer are both speech (because the possibility of confusions must be negated); whereas the threshold of acceptability would be expected to be significantly higher than the masking threshold if the target and interferer are speech and music (in either arrangement).

5.2 Pilot experiment

The main experiment was concerned with gathering speech intelligibility and acceptability scores for a range of signal to noise ratios (SNRs) and programme combinations. In addition, it was considered worthwhile to gather masking data at the same time, since a relationship between acceptability and masking had already been established for cases not focused on speech-based target programmes. In addition to these measures, recent work has shown that ‘distraction’ may be a perceptually relevant descriptor in auditory interference scenarios (Francombe 2012); the relationship between distraction and acceptability, however, is unclear, so an experiment incorporating both distraction and acceptability could clarify this.

Before the main experiment investigating these relationships could be conducted, a number of questions regarding stimuli and methodology remained. Firstly, in section 5.1.3 three potential corpora were identified as possibly being appropriate for the experiment. It was noted that it is first necessary, however, to determine whether significant differences in intelligibility would be found before undertaking the task of recording a set of Harvard sentences. Secondly, it was unclear whether absolute presentation level would have a significant effect on intelligibility, or the other measures of interest, within auditory interference scenarios. If no effects, or only very small effects, were found, the main experiment could use a fixed presentation level, which would allow for more data to be gather investigating the other factors (i.e. programme combinations, SNRs). Finally, the range of SNRs required to cover the full range of intelligibility and acceptability scores was unclear. Thus the three primary goals of the pilot experiment were:

1. Will GRID, or the low quality Harvard sentences, (both of which are readily available) produce different intelligibility scores than high quality recordings of
the Harvard sentences? If these are different then recordings should be made of the Harvard sentences to minimise the effect of target quality degradations.

2. Does absolute presentation level affect the intelligibility or other measures? If it does not, or if the effects are very small, then a single presentation level can be used in the main experiment.

3. What range of SNRs is suitable to cover the full range intelligibility scores, and other measures, for ecologically valid auditory interference scenarios of the type under test?

The pilot experiment also represented an opportunity to test the methods for obtaining intelligibility scores and additional measures such as masking, acceptability, and distraction, to highlight any potential problems with collection strategies.

5.2.1 Presentation levels

It was shown in Kjems et al. (2009) that, using the Dantale 2 corpus for the target, the SRT occurred at SNRs of $-7.3$ dB for SSN, $-8.8$ dB for cafeteria noise, $-20.3$ dB for car interior noise, and $-12.2$ dB for bottling noise. In all cases the intelligibility was at 100% for 10 dB SNR, and in all cases the intelligibility was at 0% for $-30$ dB SNR. This indicates that an operating range of 40 dB SNR would be suitable for the experiment. Presentation levels varied between 60 and 68 dBA SPL in Kjems et al. (2009), which approximately lines up with the preferred listening levels in automotive environments discussed in Benjamin and Crockett (2005).

Since Benjamin and Crockett (2005) indicated preferred listening levels ranging from 60 to 76 dB SPL in automotive environments, and since Pearsons et al. (1977) found that conversational speech levels generally range from 55-67 dB Leq (measured at one metre for ambient noise levels ranging 48-70 dB), the range of presentation levels of interest are approximately 55 to 75 dB. If target levels were fixed at points within the preferred listening levels range, and interferer levels shifted to give various SNRs, the range $-30$ to $+10$ dB SNR could only have been achieved using very loud stimulus presentations (e.g. a target at 60 dB and interferer at 90 dB SPL for $-30$ dB SNR). If interferer levels were fixed, however, and the target shifted in level, the target would have been rather quiet in some cases, which may not have been ecologically valid (e.g. interferer at 60 dB and target at 30 dB SPL). Thus the total stimulus presentation levels were held fixed, at 55, 65 and 75 dB SPL, with relative levels of target and interferer varying as in Kjems et al. (2009). The level combinations which were used are laid out in table 5.1.
### Chapter 5: Speech Intelligibility Experiment

#### 5.2.2 Stimuli

For the pilot experiment the main priority was that interferers should be representative of those in the listening environments considered. For a pilot experiment, however, a small selection was required to simply confirm that there are differences between interferer programme types and to help determine the maximum necessary range of levels for the main experiment plan. It was therefore decided to use 1 (pop) music interferer, 1 speech (radio interview) interferer, and 1 mixed (film sound) interferer (see table 5.2). Additionally a ‘reference’ interferer of speech-shaped noise (SSN), constructed from the average frequency spectrum of the high quality Harvard sentences, was included. The inclusion of the SSN interferer allows for some comparison with the literature, and it was shown in Wong et al. (2012) that SSN can be fairly representative of a range of background noises so it may be possible to use SSN as a general case interferer for broader predictive purposes.

For speech intelligibility testing each target stimulus may not be presented more than once because this allows subjects to guess the sentence based on their memory of a response to a prior trial. Due to this constraint, intelligibility scores for repeated trials are confounded with target phrase; thus repeated measurements are affected both by the consistency of the subjects and by the inherent intelligibility of the target phrase.

It was also necessary, for the experiment design, to decide whether target phrases

<table>
<thead>
<tr>
<th>IV level</th>
<th>Presentation Level</th>
<th>Target to Interferer Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55 dB</td>
<td>−30 dB</td>
</tr>
<tr>
<td>2</td>
<td>55 dB</td>
<td>−20 dB</td>
</tr>
<tr>
<td>3</td>
<td>55 dB</td>
<td>−10 dB</td>
</tr>
<tr>
<td>4</td>
<td>55 dB</td>
<td>0 dB</td>
</tr>
<tr>
<td>5</td>
<td>55 dB</td>
<td>10 dB</td>
</tr>
<tr>
<td>6</td>
<td>65 dB</td>
<td>−30 dB</td>
</tr>
<tr>
<td>7</td>
<td>65 dB</td>
<td>−20 dB</td>
</tr>
<tr>
<td>8</td>
<td>65 dB</td>
<td>−10 dB</td>
</tr>
<tr>
<td>9</td>
<td>65 dB</td>
<td>0 dB</td>
</tr>
<tr>
<td>10</td>
<td>65 dB</td>
<td>10 dB</td>
</tr>
<tr>
<td>11</td>
<td>75 dB</td>
<td>−30 dB</td>
</tr>
<tr>
<td>12</td>
<td>75 dB</td>
<td>−20 dB</td>
</tr>
<tr>
<td>13</td>
<td>75 dB</td>
<td>−10 dB</td>
</tr>
<tr>
<td>14</td>
<td>75 dB</td>
<td>0 dB</td>
</tr>
<tr>
<td>15</td>
<td>75 dB</td>
<td>10 dB</td>
</tr>
</tbody>
</table>

*Table 5.1: Target and interferer levels*
should be fixed across subjects (for the same conditions). If the target phrase for each condition (and repeat) is fixed across subjects, this allows for a better comparison of intelligibility scores across subjects. The effect of this, therefore, is that subject differences can be analysed without confounds and that measurements by different subjects can be treated as if they are repeat measurements. If the target phrase for each condition differed for each subject this comparative analysis would be weaker, due to the additional confound of target phrase, however the effect of the target phrase would be more thoroughly confounded (by using a larger sample of target stimuli). Since there is an advantage (and corresponding disadvantage) to both options, the decision was made based solely on ameliorating the effect which was expected to be the larger obstacle during analysis. Since for each SNR and interferer programme the target phrase was selected from a phonetically balanced corpus (for the Harvard corpus) or from a very limited corpus (for the GRID corpus) it was considered that the effect of the target phrase would likely be significant but small. In contrast the inability to compare intelligibility measurements across subjects was considered to be a serious limitation since it would deny the possibility to verify inter-subject consistency, and therefore also limit the ability to exclude or separate data from any subjects found to have poor accuracy or found to be performing a different task. Since greater importance was attributed to the need for across-subject consistency, it was decided that target stimuli should be fixed across subjects for each condition.

5.2.3 Methodology

Six native English speakers individually completed the pilot experiment. The subjects were seated near the centre of a listening room meeting the specifications of ITU-R BS.1116 (1997) with one Genelec 1032 loudspeaker positioned directly in front at a distance of 2 m. The Genelec 1032 was positioned at a height of 1 m (approximately head height for a seated listener), and was used to replay the target-interferer programme combinations.

The experiment was divided into two stages. Before commencing the first stage a familiarisation stage was completed by the subject with the experimenter present and describing how to operate the user interface. In this familiarisation stage, as in the first stage of the experiment, the subjects were asked to transcribe as many words in the target sentence as they were able to identify for each trial. Masking threshold data (for validation of previous work), and acceptability data (for correlation with speech intelligibility scores), were also gathered during this stage of the experiment by means of two tick boxes marked ‘inaudible’ and ‘acceptable’ respectively (see fig. 5.1).

Although it may initially seem to be challenging to transcribe the target speech while also deciding whether the target was masked and whether the listening scenario was acceptable, the task is in fact much simplified by the simple fact that when the target
<table>
<thead>
<tr>
<th>Programme</th>
<th>Description</th>
<th>Selection Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pop music interferer</td>
<td>Pixie Lott’s Mama Do</td>
<td>Pixie Lott’s Mama do reached number 1 on the UK Singles Chart in June 2009, and is therefore a reasonable selection as a piece of representative popular music.</td>
</tr>
<tr>
<td>Radio interview speech interferer</td>
<td>Radio 1 interview with Jeremy Vine recorded from live digital radio online</td>
<td>The Jeremy Vine show is one of the more popular shows on BBC radio 2, which is the most-listened to radio station in the UK with over 15 million listeners for the quarter ending December 2012 (figures obtained from (RAJAR 2013)).</td>
</tr>
<tr>
<td>Film sound mixed interferer</td>
<td>Audio extracted from dining hall scene in Titanic (featuring chatter, strings, and Foley)</td>
<td>Titanic is the third highest grossing film ever in the UK (the first two are relatively recent films, and may therefore have been viewed less by the general populace) and is therefore a reasonable choice from which to take an extract.</td>
</tr>
<tr>
<td>Speech Shaped Noise (SSN)</td>
<td>Noise based on the average frequency spectra of the Harvard high quality sentence recordings, produced using Acustyk (Bartus 2013) and PRAAT (Boersma and Weenik 2013)</td>
<td>Allows for comparison with the literature to check that intelligibility scores are consistent. It is also useful to investigate the possibility of using SSN as a general case interferer.</td>
</tr>
</tbody>
</table>

*Table 5.2: The interferer programmes used in the experiment.*
Please transcribe the speech you hear, ignoring any interfering programmes.

Figure 5.1: The user interface used by subjects to provide transcripts of the target speech as well as indicate those cases where the listening scenario was acceptable or where the target was masked.
speech is inaudible there is nothing for the subject to transcribe, and subjects only marked the listening scenario as being acceptable when they found it very easy to transcribe the target speech. Subjects were required to complete at least 15 of the 30 trials during the familiarisation stage, however most understood the task and were comfortable with the work flow within completing only the first 10 trials. In fact, some (experienced) subjects reported finding the task much easier than other listening tests they had previously taken part in because the amount of qualitative judgement involved in completing the task was minimal.

For 4 interferer programmes and 15 presentation level combinations, with 3 target corpora there were 180 trials to be completed in stage 1 of the experiment. Most subjects completed this stage in approximately 35 minutes, thus they completed each trial in approximately 12 seconds on average. By way of comparison, in Barker and Cooke (2007) subjects were able to complete trials at an average rate of one trial per 3-4 seconds, however they had only to identify the colour, letter, and digit spoken for GRIDs phrases.

Distraction was rated during the second stage of the experiment using the interface shown in fig. 5.2, with no reference or anchor. Twelve pages with ten stimuli on each were presented (2 repetitions for each listening scenario). This part of the experiment utilised the same programme and level combinations as the first stage. Subjects were instructed to rate all stimuli in which the target was totally inaudible as 100% distracting (overpowered), but were otherwise free to rate stimuli however they felt appropriate according to the provided scale.

Stage 2 of the experiment generally took 10-15 minutes to complete, thus subjects spent approximately 1 minute on each page of stimuli. The stimulus combination presentation order was randomised across subjects, and the presentation order was automatically stored in a text file for later analysis.

5.2.4 Calculation of intelligibility

An answer sheet was constructed featuring the correct target phrase for each trial. An automated comparison was first conducted: those subject responses which were identical to the answer phrase (ignoring capitalisation) were automatically marked with a score of 1, while those subject responses which were entirely blank were marked with 0. The remaining subject responses were manually scored according to two distinct metrics: intelligibility (word score), and keyword intelligibility (keyword score). The intelligibility scores were calculated as in eq. (5.1), and the keyword intelligibility scores were calculated as in eq. (5.2). In this way two measures were obtained, one representing intelligibility directly and the other giving a better indication of understanding.

Marking rules were established for interpreting transcripts featuring homophones or
typographic errors. Homophones were marked as correct, e.g. ‘wipe the Greece off his dirty face’ was scored at 100% for the answer phrase ‘wipe the grease off his dirty face’. This strategy was applied because for homophones the transcribed word perfectly represents the correct phonemes reproduced (even if the meaning differs), and so no marks can reasonably be deducted on grounds of intelligibility. Misspellings, however, sometimes required more careful interpretation. The response, “lat blue at p 6 now”, for an answer phrase, “lay blue at p 6 now”, appears to be a typographic user error because ‘t’ and ‘y’ were adjacent keys on the subject’s user interface and because ‘lat’ is not an English word. By contrast it is far less clear how the response, “these days a chicken’s egg is a rare dish”, for a target phrase, “these days a chicken leg is a rare dish”, should be scored. While it may be that the response word ‘egg’ is an example of the subject hearing correctly but deciding poorly based on prior cultural/contextual cues, this type of interpretation is somewhat speculative and the word ‘egg’ was marked as incorrect in such similar analyses. The word ‘chicken’s’, however, contains within it the entire correct answer word ‘chicken’, featuring only an additional coarticulation with the following word, and such cases were therefore marked as correct.

5.3 Pilot experiment results

This section outlines the results of the pilot experiment, describing the findings for the dependent variables (intelligibility, acceptability, masking, and distraction), which answer the questions posed as motivation for the pilot experiment.
5.3.1 Scoring transcripts and contextual errors

Some response errors were worthy of note. The correct phrase ‘The girl at the booth sold fifty bonds’, was responded to four times with ‘the girl at the booth sold fifty buns’ and once with ‘the guard at the booth sold fifty bombs’. It seems likely that these responses were a result of a partial degradation of the word ‘bonds’ and the relative word frequencies involved, i.e. it is suggested that ‘buns’ is more likely to follow phrases of the form ‘the girl at the booth sold fifty’ than ‘bonds’ in the general usage encountered by the subjects in this experiment, and that ‘bombs’ was also a response affected by the contextual effect of the earlier misheard word ‘guard’. Pollack et al. (1959) demonstrated that the frequency of occurrence of words has a substantial effect upon their intelligibility, especially when the target phrase can feature words from a large set. This type of contextually informed error, however, rarely occurred when the target phrase was from the GRID corpus since it was usually clear to the subject that the phrase would necessarily feature a colour, letter, and number.

As well as word frequency, missing context errors were also found in the subject responses. In cases where a limited portion of the keywords were understood by the subject, there was sometimes insufficient context for the subject to correctly interpret the surrounding partially degraded phonemes such that the subject response featured words which sound similar to the correct response but which would clearly not make sense if the subject had access to other important contextual words. For example one response to the target phrase ‘The colt reared and threw the tall rider’ was ‘the cold grip went through the tall tiger’. The response phrase features many incorrect words but is semantically correct. It seems possible that if the earlier words ‘colt’ and ‘reared’ had been correctly heard, the subject may have been able to combine this contextual information with the correctly heard phoneme ‘er’ and the correctly identified number of syllables in the word (2) to identify the correct word ‘rider’. While it is not clear for any particular specific phrase whether such information would have changed the subject’s response, it is clear that such possibilities do not exist for target corpora featuring small sets of predictable types (e.g. as in the colour, letter, number structure of GRID).

5.3.2 Data transformation and the validity of ANOVAs

Data which vary between two fixed end points have variances which are intrinsically capped by those end points and are therefore compressed. The assumption of the homogeneity of variance implicit in the use of ANOVAs is therefore violated in such cases (as well as the assumption of normality). It is sometimes appropriate to perform ANOVAs on data which has been transformed in order to normalize this error variance, as in Calandruccio et al. (2010) where intelligibility scores were transformed to
5.3.3 Intelligibility

With the transcripts marked according to two dependent variables, intelligibility and keyword intelligibility, the resulting data were investigated for subject differences before ANOVAs and post hoc tests were carried out to investigate which factors had significant effects.

**Subject differences**

For mean keyword intelligibility the ranges of scores, across subjects, were 0%, 3.7%, 19.3%, 22.6%, and 5.7% for SNRs −30, −20, −10, 0, and +10 dB respectively. While some subjects reported that this task was particularly easy to comprehend, subject 1 reported initial difficulties with performing the task and, as fig. 5.3 shows, the mean intelligibility and keyword intelligibility scores for this subject were consistently lower.
Table 5.3: An ANOVA of intelligibility scores with all non-significant (at 95% confidence) factors and interactions removed ($R^2 = 0.916$, adjusted $R^2 = 0.899$).

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>179</td>
<td>54.88</td>
<td>&lt;0.001</td>
<td>0.916</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>9331.04</td>
<td>&lt;0.001</td>
<td>0.912</td>
</tr>
<tr>
<td>SNR</td>
<td>4</td>
<td>44.57</td>
<td>&lt;0.001</td>
<td>0.908</td>
</tr>
<tr>
<td>Absolute Level</td>
<td>2</td>
<td>3.79</td>
<td>0.023</td>
<td>0.008</td>
</tr>
<tr>
<td>Interferer</td>
<td>3</td>
<td>6.83</td>
<td>&lt;0.001</td>
<td>0.022</td>
</tr>
<tr>
<td>Target</td>
<td>2</td>
<td>62.57</td>
<td>&lt;0.001</td>
<td>0.122</td>
</tr>
<tr>
<td>Absolute Level by Interferer</td>
<td>6</td>
<td>3.65</td>
<td>0.001</td>
<td>0.024</td>
</tr>
<tr>
<td>SNR by Interferer</td>
<td>12</td>
<td>13.02</td>
<td>&lt;0.001</td>
<td>0.148</td>
</tr>
<tr>
<td>Interferer by Target</td>
<td>6</td>
<td>3.15</td>
<td>0.005</td>
<td>0.021</td>
</tr>
<tr>
<td>SNR by Target</td>
<td>8</td>
<td>29.45</td>
<td>&lt;0.001</td>
<td>0.207</td>
</tr>
<tr>
<td>SNR by Absolute Level by Interferer</td>
<td>32</td>
<td>2.29</td>
<td>&lt;0.001</td>
<td>0.075</td>
</tr>
<tr>
<td>Absolute Level by Interferer by Target</td>
<td>16</td>
<td>2.45</td>
<td>0.001</td>
<td>0.042</td>
</tr>
<tr>
<td>SNR by Target by Interferer</td>
<td>24</td>
<td>3.12</td>
<td>&lt;0.001</td>
<td>0.077</td>
</tr>
<tr>
<td>SNR by Absolute Level by Interferer by Target</td>
<td>64</td>
<td>2.24</td>
<td>&lt;0.001</td>
<td>0.137</td>
</tr>
</tbody>
</table>

than the other subjects. Despite this the trend of scores for this subject were generally in line with those trends of the other subjects, and this data was not removed from analysis.

**ANOVA**

Tables 5.3 and 5.4 show the ANOVAs for the intelligibility and keyword intelligibility scores respectively. The analyses show that SNR, absolute level, interferer programme, and target corpus are all significant, as well as many of their interactions. The effect sizes, however, are different for each independent variable, with SNR having the largest effect (partial $\eta^2 = 0.91$ and 0.90), target corpus having a much smaller effect (partial $\eta^2 = 0.12$ and 0.13), and both absolute level (partial $\eta^2 = 0.01$ and 0.01) and interferer programme (partial $\eta^2 = 0.02$ and 0.01) having effect sizes an order of magnitude smaller.

Tukey HSD and Bonferroni post hoc tests were carried out to further investigate the nature of the significant effects. Each SNR could be considered a significantly different group (at $\alpha < 0.001$), for intelligibility and keyword intelligibility scores, with the exception of $-30$ and $-20$ dB SNR (which had $\alpha = 0.912$ and $\alpha = 0.886$ respectively for the Tukey HSD test, and which had $\alpha = 1.000$ in both cases for the more conservative Bonferroni test). This indicates that for $-20$ dB SNR intelligibility and keyword intelligibility were so close to zero that they could not be distinguished from the scores at $-30$ dB SNR.
Table 5.4: An ANOVA of keyword intelligibility scores with all non-significant (at 95% confidence) factors and interactions removed ($R^2 = 0.910$, adjusted $R^2 = 0.892$).

The Tukey HSD and Bonferroni post hoc tests for target corpus revealed significant differences between all three target corpora for the intelligibility score ($\alpha < 0.001$ for all comparisons), but only revealed significant differences between the low quality Harvard sentence recordings and the other two corpora for keyword intelligibility with $\alpha = 0.330$ and 0.467 for Tukey HSD and Bonferroni tests between the keyword scores for GRID and high quality Harvard sentences respectively. Figure 5.4 shows the intelligibility scores for each target corpus separated by various interferer programmes. The effect size of the target corpus was slightly smaller than the effect size of the interaction between SNR and target corpus (partial $\eta^2 = 0.126$, and 0.186 respectively). This is unsurprising, however, because the differences between the mean scores separated by target corpus are effectively compressed by the inclusion of scores at $-30$ dB SNR which are all 0% regardless of the interferer programme.

For intelligibility score the Tukey HSD and Bonferroni post hoc tests revealed no significant difference between 55 and 75 dB SPL presentations ($\alpha = 0.998$ and 1.000 respectively). The Bonferroni post hoc test showed no significant difference between intelligibility scores for 55 and 65 dB SPL ($\alpha = 0.057$) but found a significant difference between scores for 65 and 75 dB SPL ($\alpha = 0.048$), whereas the less conservative Tukey HSD test showed significant differences between both pairs ($\alpha = 0.049$ for 55 and 65 dB SPL, and $\alpha = 0.042$ for 65 and 75 dB SPL). For keyword intelligibility scores the Bonferroni post hoc test showed a significant difference between 55 and 65 dB SPL ($\alpha = 0.002$), but no other significant differences ($\alpha = 0.269$ for 55 and 75 dB SPL,
and $\alpha = 0.250$ for 65 and 75 dB SPL). The Tukey HSD post hoc test also only showed
significant differences between 55 and 65 dB SPL ($\alpha = 0.002$), but no other significant
differences ($\alpha = 0.206$ for 55 and 75 dB SPL, and $\alpha = 0.193$ for 65 and 75 dB SPL).
The significant differences found between absolute levels for intelligibility scores were
relatively marginal, while those found for keyword intelligibility were more definitive,
however for both cases the effect size was extremely small. The result of which is that
the mean keyword intelligibility scores, averaged across all cases and subjects, were
39.6%, 43.4%, and 41.5% for 55, 65, and 75 dB SPL respectively, and the intelligibility
scores, averaged across all cases and subjects, were 40.8%, 43.3%, and 40.7%. Thus
the effect of absolute level (over a range of 20 dB) is likely limited to a few percentage
points.

Figure 5.5 shows the interaction between intelligibility scores, interferer programme
and SNR. The Tukey HSD post hoc tests revealed that for intelligibility score the
pop interferer programme was significantly different from the other three interferers
($\alpha = 0.002, 0.027,$ and $< 0.001$ for comparisons with radio, SSN, and film interferers
respectively) and the other three interferer programmes were not significantly different
from one another. For keyword intelligibility scores, however, only the pop and
film interferer programmes were significantly different ($\alpha = 0.013$) with all other
comparisons being non-significant. The Bonferroni post hoc tests were in agreement
with these. For both intelligibility and keyword intelligibility the lowest mean scores
were found for the film and radio interferers, both of which were highly linguistic in
nature, however the effect size of the interferer programme differences was very small.
While the effect size of the interferer programme was very small (partial $\eta^2 = 0.011$),
the effect size of the interaction between SNR and interferer was an order of magnitude
larger (partial $\eta^2 = 0.113$). This is unsurprising, however, because the differences
between the mean scores separated by interferer programme are effectively compressed
by the inclusion of scores at very low and very high SNRs which are all 0% and 100%
respectively regardless of the interferer programme.

**Comparison with the literature**

Table 5.5 shows SRTs for each interferer programme and intelligibility metric calculated
by making simple linear interpolations between the $-10$ dB and 0 dB mean scores. Such
linear interpolations are unlikely to be very accurate, but give an impression of the likely
scores.

In Barker and Cooke (2007) SRTs ranging from $-8$ to $-14$ dB SNR were reported
for GRID keywords with a SSN interferer and for male speakers. The overall keyword
intelligibility SRTs were $-9$ and $-11$ dB SNR respectively. In Kjems et al. (2009)
the SRT, for a SSN interferer, was estimated at $-7.3$ dB SNR, however this was for
a sentence SRT, (i.e. the SNR at which 50% of sentences were reported at 100%
Figure 5.4: Intelligibility and keyword intelligibility scores separated by interferer programme and target corpus.
Chapter 5: Speech Intelligibility Experiment

Figure 5.5: Intelligibility and keyword intelligibility scores separated by interferer programme.

<table>
<thead>
<tr>
<th>Interferer Programme</th>
<th>Intelligibility Metric</th>
<th>SRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>pop</td>
<td>keyword intelligibility</td>
<td>−7.4 dB</td>
</tr>
<tr>
<td>pop</td>
<td>intelligibility</td>
<td>−8.1 dB</td>
</tr>
<tr>
<td>radio</td>
<td>keyword intelligibility</td>
<td>−5.6 dB</td>
</tr>
<tr>
<td>radio</td>
<td>intelligibility</td>
<td>−5.7 dB</td>
</tr>
<tr>
<td>SSN</td>
<td>keyword intelligibility</td>
<td>−5.9 dB</td>
</tr>
<tr>
<td>SSN</td>
<td>intelligibility</td>
<td>−5.7 dB</td>
</tr>
<tr>
<td>film</td>
<td>keyword intelligibility</td>
<td>−5.1 dB</td>
</tr>
<tr>
<td>film</td>
<td>intelligibility</td>
<td>−5.1 dB</td>
</tr>
</tbody>
</table>

Table 5.5: Linearly interpolated SRTs for both measures of intelligibility for each interferer programme.
Table 5.6: An ANOVA of distraction scores with all non-significant (at 95% confidence) factors and interactions removed ($R^2 = 0.877$).

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>29</td>
<td>168.96</td>
<td>&lt;0.001</td>
<td>0.877</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>31155.39</td>
<td>&lt;0.001</td>
<td>0.978</td>
</tr>
<tr>
<td>SNR</td>
<td>4</td>
<td>1181.59</td>
<td>&lt;0.001</td>
<td>0.873</td>
</tr>
<tr>
<td>Absolute Level</td>
<td>2</td>
<td>6.91</td>
<td>0.001</td>
<td>0.020</td>
</tr>
<tr>
<td>Interferer</td>
<td>3</td>
<td>3.54</td>
<td>0.014</td>
<td>0.015</td>
</tr>
<tr>
<td>SNR by Interferer</td>
<td>12</td>
<td>11.06</td>
<td>&lt;0.001</td>
<td>0.161</td>
</tr>
<tr>
<td>SNR by Absolute Level</td>
<td>12</td>
<td>2.05</td>
<td>0.038</td>
<td>0.023</td>
</tr>
</tbody>
</table>

accuracy). In contrast Lavandier and Culling (2010) reported an SRT of $−6$ dB for SSN with the Harvard sentences. These results confirm that the $−6$ dB SRT found for SSN in this test is reasonable.

It should be noted that there may have been a slight negative impact upon SRTs due to usage of three separate corpora. When only one target corpus is used, the subject is better able to predict the form of the phrases to come, thus potentially gaining an advantage in knowing what to listen for; in this pilot experiment some of that advantage was removed because the test phrases in the three corpora used were of different structural forms. This removal of advantage may be partially responsible for the small differences between these scores and those in the related literature.

Summary

In summary all four independent variables were significant with SNR having by far the largest effect and target corpus having a moderately large effect also while absolute level and interferer programme had very small effects which were smaller than that of some of the higher level interactions (on the order of a few percent). The linearly interpolated SRTs were approximately in line with those found in the literature, but smaller SNR steps are required in the main experiment to produce more precise threshold estimates.

5.3.4 Distraction scores

The ANOVA (see table 5.6) showed that SNR, presentation level, and interferer programme were all significant at 95%, however the effect size for SNR was much greater (partial $\eta^2 = 0.873$) than for level or interferer (partial $\eta^2 = 0.020$ and 0.015 respectively). The effect size for the interaction between SNR and interferer programme was relatively large (partial $\eta^2 = 0.161$), and is discussed further in section 5.3.4.

Tukey and Bonferroni post hoc tests showed that distraction scores for all SNRs were significantly different ($\alpha < 0.001$), with the exception of $−30$ and $−20$ dB SNR ($\alpha =$
For presentation level, Tukey and Bonferroni post hoc tests showed that 55 and 65 dB SPL were not significantly different ($\alpha = 0.573$ and $0.943$ for Tukey HSD and Bonferroni post hoc tests respectively), but both were significantly different from 75 dB SPL ($\alpha = 0.001$ and 0.026 for Tukey HSD comparisons between 55 and 75 dB SPL, and between 65 and 75 dB SPL respectively, and $\alpha = 0.001$ and 0.029 for Bonferroni comparisons between 55 and 75 dB SPL, and between 65 and 75 dB SPL respectively), for which distraction scores were slightly higher. Although no significant differences were found between distraction scores for 55 and 65 dB SPL, a general trend of higher distraction scores for higher presentation levels was found (mean scores 71.2, 72.2, and 74.9 for 55, 65 and 75 dB SPL respectively). The effect size of the interaction between SNR and absolute level, however, was greater (partial $\eta^2 = 0.023$) than the effect size of the absolute level alone (partial $\eta^2 = 0.020$), and so it is difficult to conclude any such relationship from this data.

For interferer programme, the SSN and film interferers were not significantly different from any other programmes ($\alpha = 0.111$ and 0.086 for Tukey HSD comparisons between pop and SSN, and between pop and film interferers respectively, and $\alpha = 0.843$ and 0.891 for Tukey HSD comparisons between radio and SSN, and between radio and film interferers respectively), however the Pop and Radio interferers were significantly different from one another ($\alpha = 0.012$ for Tukey HSD comparison). The finding that SSN was not significantly different from any of the alternative interferer programmes supports the argument for its inclusion in the main experiment as a general case interferer for distraction scores (previously also suggested for intelligibility).

**Effect of interaction between SNR and interferer programme on distraction**

A relationship was found between mean distraction score and interferer programme. The bottom right quadrant of fig. 5.6 shows these for the range of SNRs tested.

The change in distraction scores, due to SNR, for the pop, SSN, and film interferers were very similar whereas the radio interferer produced a shallower curve with distraction scores lower than the other interferer programmes for unfavourable SNRs, but with distraction scores higher than the other interferers at more favourable SNRs. For the negative SNRs the radio interferer was likely considered less distracting because the temporal sparsity of the interfering speech allowed for at least some degree of understanding of the target programme through ‘dip listening’, whereas the other three interferers had relatively consistent interferer energy across time. For the positive SNRs, however, the result is in line with previous indications that when target and interferer programmes are of the same ‘type’ (either linguistic or musical/non-linguistic) the listening scenario is more distracting than when the two programmes are of opposed types. The possibility of such a result was mentioned in section 5.1.3, and previous...
research in sound zoning also implies this relationship.

It is notable that the film interferer, although featuring multiple interfering speakers, followed closely the trend of the pop and SSN interferers. One might expect the film and radio interferers to agree since they both feature interfering linguistic content, however it has been noted that when multiple interfering speakers are simultaneously present the scenario is much more closely related to that of energetic masking rather than informational masking (speech confusion masking) (Brungart et al. 2001).

It is also interesting to note that the distraction scores for the SSN match those of the pop and film interferers reasonably well. This is slightly surprising because, unlike the pop and film interferers, SSN is a temporally stationary interferer which provides only spectral masking of the target. This result implies that it might be possible to use distraction scores for interfering SSN to predict distraction scores for arbitrary interferer programmes (when those programmes interfere energetically instead of informationally). This possibility will be re-evaluated after the main experiment, however, where more data will be available for detailed comparisons.

As fig. 5.6 shows, when the results are split by presentation level the interpretation changes slightly. While the trends for 65 and 75 dB SPL are very similar to one another, at 55 dB SPL the distraction score for the radio interferer at $-20$ dB SNR is slightly higher than expected. This difference might be explained by considering that dip listening is likely to be less effective when the target level approaches the threshold of hearing. Additionally at 55 dB SPL and $-10$ dB SNR the pop interferer distraction score is slightly lower than the trend for 65 and 75 dB SPL.

5.3.5 Acceptability and masking

The acceptability and masking data gathered were binary and thus require a slightly different method of analysis. First, however, general trends across SNR are considered. Figure 5.7 shows the relationship between SNR and masking, and between SNR and acceptability. For the former this trend is a negative correlation between SNR and the likelihood that the target is masked, and for the latter this trend is a positive correlation between SNR and the likelihood that the listening scenario was acceptable. For this experiment, however, both buttons were used relatively infrequently throughout the experiment because the range of SNRs tested provided little scope for their use, i.e. with a maximum SNR of 10 dB the listeners rarely found the listening scenario to be acceptable, and when the SNR was at its minimum ($-30$ dB) the listeners still clicked the ‘inaudible’ button on only around half of these trials (for each SNR each option could have been selected a maximum of 216 times).

The crossover at a low probability of approximately 5% implies the intuitively reasonable notion that when the target is masked the listening scenario cannot be
Chapter 5: Speech Intelligibility Experiment

Figure 5.6: Mean distraction scores separated by interferer programme and absolute level

Figure 5.7: Proportions for usage of the ‘inaudible’ and ‘acceptable’ check boxes for various SNRs.
Table 5.7: The results of a logistic regression on the binary masking data for those factors whose inclusion in the model makes a significant difference to predictions based on guessing and where the coefficient of the factors are significantly different from 0. Cox & Snell R square = 0.47, Nagelkerke R square = 0.71.

<table>
<thead>
<tr>
<th></th>
<th>B(SE)</th>
<th>Sig.</th>
<th>Lower</th>
<th>Odds Ratio</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Included</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.29(1.67)</td>
<td>0.00</td>
<td></td>
<td>541.34</td>
<td></td>
</tr>
<tr>
<td>SNR</td>
<td>-2.28(0.71)</td>
<td>0.01</td>
<td>0.03</td>
<td>0.10</td>
<td>0.41</td>
</tr>
</tbody>
</table>

acceptable (because these results include all subjects, target corpora, and interferer programmes, thus there were no cases in which the listening scenario was simultaneously masked and acceptable), and in fact the listening scenario is only likely to be acceptable at a much more favourable SNR.

**Logistic regressions**

The categorical variables (interferer programme and target corpus) cannot be considered using a bivariate correlation analysis, and since the masking and acceptability data are binary the assumption of linearity required to conduct an ANOVA is not satisfied either. As such a logistic regression was carried out to investigate whether these factors were significant and to determine the sizes of their effects.

Table 5.7 shows the result of a logistic regression carried out on the masking data and the four independent variables and all possible interactions. Those factors for which inclusion in the logistic model significantly improved predictions compared with guessing and for which the coefficient of the factor was significantly different from zero were included. While a number of factors and interactions fulfilled the first criterion, only SNR fulfilled the second, and this is likely to be due to the limited sample sizes and SNR range involved in data collection. The results indicate that the SNR significantly affected the masking score, and that for an increase in 10 dB SNR the probability that a subject will not be able to hear the target is decreased by a factor of 10.

Table 5.8 shows the results of a logistic regression carried out on the acceptability data and the four independent variables and all possible interactions. As with the masking data, SNR was the only factor for which the inclusion in the logistic model significantly improved predictions compared with guessing and for which the coefficient of the factor was significantly different from zero. A number of factors fulfilled the former, but not the latter, condition, again, and this is likely due to the limited sample sizes involved. The results indicate that the SNR significantly affected the acceptability score, and that for an increase in 10 dB SNR, the change in probability that a subject will find the listening scenario acceptability is a factor of 10.
Table 5.8: The results of a logistic regression on the binary acceptability data for those factors whose inclusion in the model makes a significant difference to predictions based on guessing and where the coefficient of the factors are significantly different from 0. Cox & Snell R square = 0.15, Nagelkerke R square = 0.42.

Inferring thresholds

The data collected can be used to infer the appropriate thresholds of masking and acceptability by fitting a curve to the data. Doing so in fig. 5.8, and taking a masked probability of 50% to be the masking threshold, would produce masking thresholds of approximately \(-18\), \(-22\), and \(-14\) dB SNR for the pop, SSN, and film interferers respectively, and a masking threshold for the radio interferer which is known only to be less than \(-30\) dB. These are reasonable estimates, however the large steps of 10 dB per data point is a serious limitation upon the accuracy of the estimate, as is the lower limit of \(-30\) dB.

When separated by target corpus, however, the results change slightly. For the GRID corpus the masking thresholds were, \(-16\), \(-23\), \(-10\), and \(<-30\) dB SNR for the pop, SSN, film and radio interferers respectively. For the Harvard low quality target corpus the masking thresholds were non-monotonic, although the variance is great and it is difficult to interpret the masking thresholds with more precision than to say that they must all be lower than \(-10\) dB SNR. For The Harvard high quality target corpus the masking thresholds were \(-16\), \(-23\), \(-15\), and \(<-30\) dB SNR. The only difference between the masking thresholds for GRID and Harvard high quality appears to have been for the film interferer, for which a lower masking threshold was found with the Harvard high quality sentences (i.e. the GRID sentences were more easily confused amidst the babble). This result is intuitively sensible since the GRID sentences were taken from a wide range of speakers, so the subjects effectively did not know which vocal characteristic to listen for amidst the babble.

5.3.6 Summary and consequences for main experiment

The pilot experiment was conducted primarily to answer the following questions:
1. Does absolute presentation level affect speech intelligibility?

The results indicated that absolute presentation level did affect speech intelligibility scores, however the effect was extremely small (in fact smaller than some interactions).
Similar results were found for the additional measures, therefore, a single presentation level (65 dB SPL) will be used in the main experiment.

2. Does target corpus affect speech intelligibility?

The results indicated that target corpus does affect speech intelligibility, and the effect size was moderate. The low quality Harvard sentences produced much lower intelligibility scores than the high quality Harvard sentences, and the GRID phrases also produced slightly lower scores (for intelligibility but not for keyword intelligibility). Additionally, contextual errors (see section 5.3.1) were found for responses using the Harvard target phrases, yet such responses would not occur in an experiment which only features target phrases contained in small, predetermined sets (such as GRID). Such contextual errors are likely within ecologically valid listening environments. For the main experiment, therefore, high quality recordings will be made of the Harvard Sentence lists since this is considered to be more externally valid.

3. What range of SNRs is suitable to cover the full range of speech intelligibility scores for ecologically valid auditory interference scenarios?

The results indicated significant differences between scores at all SNRs tested except between $-30$ and $-20$ dB. This implies that the full range of intelligibility scores can be produced in the main experiment using only the range $-20$ to $+10$ dB. The distraction scores, however, ranged from 100 down to a lowest score of 5, but with a mean score of 27.9 for $+10$ dB SNR, implying that SNRs above 10 dB should be included to investigate the full range of distraction scores. Furthermore the lack of a significant difference between distraction scores at $-30$ and $-20$ dB implies that nearly the full range of distraction scores can be captured between $-20$ and $+20$ dB SNR.

Further than these proposed questions, the method of obtaining additional data for
acceptability and masking thresholds was apparently effective, however while the collection of binary data is faster (and simpler) for the subjects than scalar data, larger sample sizes are required to detect statistically significant differences in the interferer programme. In the pilot experiment there were 6 subjects and no repeats, so for a sample size of $n = 6$ per condition, many conditions had overlapping confidence intervals even though it seems likely from the mean values calculated (and from previous masking experiments) that there would be significant differences between all four interferer programmes. The main experiment must therefore involve a larger number of subjects as well as repeats if this method is to be used.

Finally, the set of results now allows for a simplistic way of describing the SNR space in terms of multiple metrics, as shown in fig. 5.9. The independent variables (interferer programme, target programme, and absolute level) have been averaged across in order to produce such a simple graphical representation, and the representation also does not take account of any other factors (such as relative spectral content, or speaker accent) which may have effects. The representation, however, is informative in several ways. Firstly, it shows the close but non-linear relationship between acceptability and masking of the interferer programme. The red line represents the logistic fit to the masking data, and the orange line is the reflection of the red line about the point SNR $= 0$ dB. This reflection approximates the logistic fit expected for similar cases where the interferer programmes are masked by the target. The changing width of the gap between the orange and blue lines indicates the non-linear relationship between the two. Secondly, it shows that while there may be a strong relationship between target intelligibility and acceptability it is likely to be of the form where a listening scenario is considered totally unacceptable without very high intelligibility, i.e. a high acceptability marks the lower bound of acceptability. Thirdly, it is apparent that distraction and acceptability scores are not merely the inverse of one another, at least for the situations currently tested, since 50% distraction seems to occur close to 0 dB SNR, whilst 50% acceptability occurs at around 20 dB SNR.

5.4 Main experiment

The main experiment was conducted similarly to the pilot experiment but with some alterations inspired by the results of the pilot experiment.

5.4.1 Presentation levels and SNRs

The pilot experiment demonstrated that it is not necessary to further investigate the effect of absolute level upon speech intelligibility, since the magnitude of the effect was small. The range of SNRs tested, however, did not cover the entire range of distraction
scores, and covered only a small portion of the range of acceptability and masking. Additionally the SNRs tested in the pilot were spaced by 10 dB; such large changes in level made it difficult to interpret thresholds very precisely.

For the main experiment, therefore, all stimuli were presented at 65 dB SPL and the SNRs under test were \([-35,35]\) dB in 5 dB increments, resulting in a total of 15 levels. By testing at intervals of 5 dB the SRTs (as well as other thresholds of interest higher up the curve, e.g. 80% and 90%) could be interpolated with much greater precision than in the pilot. The extreme SNRs at \(-35, -30, -25, +25, +30, \) and \(+35\) dB were chosen to allow for a better indication of masking and acceptability thresholds which, between \(-20\) and \(+20\) dB SNR were sparingly used in the pilot.

5.4.2 Target corpus selection

The results from the pilot experiment showed significant differences between the target corpora under test with a moderate effect size. The Harvard low quality sentences had very low intelligibility scores relative to the GRID phrases, which in turn scored as slightly less intelligible than the Harvard high quality sentences. As such, it was decided that the low quality Harvard sentences should not be used for the main experiment since the results would be expected to be skewed by the low target quality.

The GRID corpus has the advantage of including speakers of both genders and multiple accents. This is useful because it has been widely noted that female speech is generally more intelligible than male speech, and that the speaker (and listener) accent can affect speech intelligibility (Brungart et al. 2001; Bradlow et al. 1996; Calandruccio
et al. 2010; Barker and Cooke 2007), and the use of a corpus with both genders and many different accents implies that the effect of vocal characteristics would be well confounded in an experiment and could thus be excluded from analysis. The results of the pilot experiment, however, revealed that it can be detrimental to use multiple target speakers because for some interferer programmes the subject may produce low intelligibility scores through a failure to identify the correct target speaker (especially for multiple speech interferers). If vocal characteristics are not confounded by use of multiple speakers, a different experimental approach must be adopted. To avoid skewing results by the use of speaker accents unfamiliar to the listeners, an alternative method would be to select a single target speaker with speech expected to be highly intelligible to the listeners. The resulting intelligibility scores would therefore be close to ideal, since practical scenarios will feature speakers with various degrees of intelligible speech, yet no target speaker identification errors would be expected.

For this experiment, high quality recordings of the Harvard sentences were made for use in the main experiment. The recordings were made in a listening room meeting the specifications of ITU-R BS.1116 (1997) with the speaker seated at 0.75 m away from a Sony C-48 microphone positioned at a height of 1.1 m set to omni-directional mode. The speaker was a female native English speaker with a south-eastern accent (it was expected that most subjects involved in the main experiment would consider this accent highly intelligible). Harvard sentence lists 1-36 inclusive were recorded. In order to ensure that intelligibility was very high under low background noise conditions, as well as to allow the subject the advantage of first becoming acquainted with the target speaker’s vocal characteristics, subjects were asked to score test phrases which did not feature an interferer programme during the familiarisation stage. This score would allow for a baseline comparison against which to compare the experiment results. The subjects were also presented with the interferer programmes during this stage in isolation to allow them to become familiar with the character of the interferers. With this arrangement the best possible speech intelligibility scores were ensured, and results could be interpreted noting that more ecologically valid scenarios may sometimes feature speaker characteristics which reduce the intelligibility of the speech for an arbitrary listener.

5.4.3 Interferer programmes

A wider range of programmes were tested in the main experiment than in the pilot. The selection was designed to cover a range of situations including interfering music, speech, and mixed interferers. The selected interferer programmes are shown in table 5.9. SSN was retained from the pilot experiment since the results indicated that SSN might be useful as a general case interferer. Three excerpts from musical programmes were selected: a classical instrumental piece, a pop song, and a film score featuring prominent
vocals. Two scenes from films were selected: a conversation scene featuring a male and female speaker with background chatter and sound effects, and an action scene featuring an explosion, music, helicopter rotor sounds, and gunshots. These two acoustic scenes are substantially different from one another in character, and both quite common in film. Three excerpts from radio were selected: a radio interview featuring two male speakers and background music, a traffic report featuring female conversational speech and background music, and a news report featuring clear female speech in isolation. These three were distinct both in terms of programme combinations as well as in the tone of speech involved. The 9 interferers selected therefore cover a wide range of possible, and representative, interferer types including music, speech, sound effects, and various mixtures.

5.4.4 Methodology

The method comprised a familiarisation stage and two experiment stages. The familiarisation stage consisted of listening to 10 test phrases, which did not appear in the main experiment, spoken by the target speaker in isolation at 65 dB SPL. This stage was scored, and the scores provided a baseline against which to ensure that problematic speaker-listener combinations were avoided. The subjects then listened to the interferer programmes in isolation, before completing 10 practice trials identical to the procedure used in stage 1 of the main experiment.

In stage 1 of the main experiment the work flow was identical to stage 1 of the pilot experiment. There were 270 trials (15 levels \( \times \) 9 interferers \( \times \) 2 repetitions). Subjects were expected to proceed at a similar rate as in the pilot experiment so 270 trials were expected to require approximately 1 hour to complete. Short mandatory breaks were enforced after completion of every 90 trials (approximately 20 minute sessions).

The second stage was a multiple stimulus rating test precisely as in the pilot experiment, with 270 trials spread over 27 pages. A short mandatory break was enforced after each 9 pages. This stage was expected to be completed at a similar rate as the pilot experiment, and thus was expected to required around 45 minutes to complete. Including the familiarisation stage of the experiment, the whole experiment was concluded in approximately 2 hours. Listeners generally completed stage 2 on a separate day to stage 1 of the test.

After the experiment was completed the transcripts were marked using the same method as the pilot experiment, producing both intelligibility and keyword intelligibility scores. Results from subjects scoring less than 95% average keyword intelligibility, or less than 90% average intelligibility, for the familiarisation stage target phrases in isolation were intended to be excluded from the analysis.
## Programme Description and Selection Justification

<table>
<thead>
<tr>
<th>Programme</th>
<th>Description and Selection Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>Noise spectrally shaped to have the same magnitude at each frequency as the average of the target phrases. Selected in order to test whether intelligibility scores for SSN can be used as a general case where the interferer is unknown.</td>
</tr>
<tr>
<td>‘Clarinet Concerto’ by Wolfgang Amadeus Mozart</td>
<td>An example of classical orchestral music with no linguistic content. 130 beats per minute, 4/4 timing, in A major. This piece is #2 on the Classic FM ultimate hall of fame, and the highest rated piece in a major key signature (ClassicFM 2013). Classic FM is the most listened to classical music radio station in the United Kingdom reaching 10% of the population during the last quarter of 2012 (RAJAR 2013).</td>
</tr>
<tr>
<td>‘Pompeii’ by Bastille</td>
<td>An example of pop music featuring prominent vocals. 126 beats per minute, 4/4 timing, in A minor. This song was featured on the BBC Radio 1 A list for 13th March 2013 (Radio 1 2013). BBC Radio 1 is one of the most listened to pop radio stations in the United Kingdom reaching 21% of the population during the last quarter of 2012.</td>
</tr>
<tr>
<td>‘Skyfall’ by Adele</td>
<td>An example of an orchestral film score featuring vocals. 76 beats per minute, 4/4 timing, in C minor. The theme song to the highest ever grossing film in the United Kingdom box office (25th Frame 2013).</td>
</tr>
<tr>
<td>‘Conversation scene’ from Skyfall</td>
<td>A conversation featuring male and female dialogue with bar/casino Foley in the background.</td>
</tr>
<tr>
<td>‘Action scene’ from Skyfall</td>
<td>Action scene featuring music, helicopter rotor noise, male speech and an explosion. Replete with transients and broad spectral content, this is expected to be an efficient energetic masker.</td>
</tr>
<tr>
<td>radio interview on Chris Evans BBC Radio 2 breakfast show</td>
<td>A conversation between two people on the most popular radio show of the most popular radio station in the United Kingdom (RAJAR 2013).</td>
</tr>
<tr>
<td>Traffic report</td>
<td>A traffic report by Lynn Bowles on Chris Evans BBC Radio 2 breakfast show; features female speech with background music.</td>
</tr>
<tr>
<td>News report by Moira Stuart on Chris Evans BBC Radio 2 breakfast show</td>
<td>A news report featuring female speech in isolation; the content is clear speech rather than conversational speech.</td>
</tr>
</tbody>
</table>

*Table 5.9: The interferer programmes used in the experiment.*
5.4.5 Assumptions and limitations

It was assumed that during the familiarisation stage the subjects would become sufficiently familiar with the interferer programmes that subjects would avoid making target selection errors during the main experiment, especially in those cases where both the target and interferer feature linguistic content. In ecologically valid listening scenarios, however, it is possible that such target selection errors might occur. It is possible, therefore, that the main experiment intelligibility scores would represent greater intelligibility than would be found within an equivalent ecologically valid listening scenario. It is also likely, however, that within ecologically valid listening scenarios target selection errors made by listeners will tend to diminish over time as the listener identifies the target vocal characteristic, so this source of error is likely to be small.

Another source of error is caused by the subject’s familiarity with the fixed interferer phrases. This familiarity may allow subjects to listen ‘in the gaps’ with greater success than they might be able to in many ecologically valid listening scenarios since listeners would not usually have access to such strong a priori information about the temporal structure of interferer audio programme.

It should be noted, therefore, that due to these two effects intelligibility within listening scenarios under consideration may be slightly less than the main experiment scores indicate in cases where the interferer is primarily linguistic. The difference, however, is expected to be relatively minute, especially for interferer programmes featuring a single speaker with SNRs greater than 0 dB (because listening in the gaps of the quieter programme is unlikely to have a large effect), and for interferer programmes featuring multiple speakers with SNRs greater than $-10$ dB (because target identification is simple when the target speech is noticeably louder than the competing speakers).

As has been previously mentioned both speaker gender, speaker accent, and listener’s familiarity with these characteristics affect speech intelligibility. The effect of these can vary substantially, however in general female speech is more intelligible than male speech by up to 20% for a range of SNRs (Brungart 2001), and everyday experience makes it plain that the effect of accent can sometimes be strong enough to prevent intelligible communication even in low background noise conditions. While these effects are known to be significant, since they can be large it would not be effective to include a very large range of speakers and simply average across these results; the extra variance would be likely to hide the significance of other factors. Instead, the approach taken in this experiment was to simply use a single speaker that was considered likely to have high intelligibility for the subjects involved (a female speaker with a south-eastern British accent), and to verify that the intelligibility of this speaker was high by including a brief ‘no interferer’ condition in the familiarisation stage. This provided
a baseline against which to compare the intelligibility results produced by the main experiment. The result of this process was that speech intelligibility scores obtained in the experiment represent an approximate best case listening scenario for gender and accent. Speech intelligibility predictions, based on the results of this experiment, could likely be improved with prior knowledge of the gender of the speaker, and such predictions would necessarily assume that the speaker and listener do not have prohibitively different accents (unless further modifications are made to predictions to account for this).

5.5 Main experiment results

This section outlines the results of the main experiment, describing the findings for the dependent variables (intelligibility, acceptability, masking, and distraction), which answer the questions posed as motivation.

5.5.1 Intelligibility

The transcripts were marked using the same procedure as that in the pilot experiment, which resulted in two dependent variables: intelligibility and keyword intelligibility. These data were investigated for subject differences before ANOVAs and post hoc tests were carried out to find the significant factors.

Subject differences and outliers

Subject 1 reported a medical condition involving cognition and memory which might affect their results. The baseline scores (see table 5.10), where subjects transcribed the target speaker without an interferer programme, did not reveal any large differences between subjects scores, however, and all subjects exceeded the 95% correct threshold designed to identify any subjects who may have serious difficulties with the accent of the target speaker. Figure 5.10 shows the intelligibility scores for each subject averaged across interferer programme and repeats. Although subject 1 did not show a decrease in performance compared with other subjects in the baseline test, the intelligibility scores for subject 1 in the main test were the lowest. This can be seen both in the offset of intelligibility curves and in the wider confidence intervals at the highest SNRs in fig. 5.10.

From an inspection of the data it seemed very likely that subject 2 had accidentally skipped two trials during the test. The subject’s responses to speech at +10 dB SNR with the interview interferer, and speech at −5 dB SNR with the Casino interferer were both blank, yet these were cases for which no other subjects (or subject 2’s repeat trials) had intelligibility scores of 0%, with most of these cases at, or close to, 100%
Figure 5.10: Mean intelligibility (right) and keyword intelligibility (left) scores for each subject and SNR, averaged across 9 interferer programmes and 2 repeats. Error bars represent 95% confidence intervals, although care should be taken for relying on these since the data is averaged across interferer programmes.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Keyword Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100.0</td>
<td>98.7</td>
</tr>
<tr>
<td>2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>3</td>
<td>96.0</td>
<td>96.0</td>
</tr>
<tr>
<td>4</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>5</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>6</td>
<td>98.0</td>
<td>98.7</td>
</tr>
<tr>
<td>7</td>
<td>100.0</td>
<td>100.7</td>
</tr>
</tbody>
</table>

Table 5.10: Subjects scores for 10 training sentences which did not feature an interferer programme.
Chapter 5: Speech Intelligibility Experiment

Mean Scores

<table>
<thead>
<tr>
<th>SNR</th>
<th>Keyword Score</th>
<th>Score</th>
<th>Median Scores</th>
<th>Keyword Score</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>−35</td>
<td>2.4</td>
<td>2.1</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>−30</td>
<td>1.6</td>
<td>1.6</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>−25</td>
<td>11.4</td>
<td>12.6</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>−20</td>
<td>38.9</td>
<td>40.0</td>
<td>40.0</td>
<td>38.0</td>
<td></td>
</tr>
<tr>
<td>−15</td>
<td>74.8</td>
<td>77.1</td>
<td>80.0</td>
<td>87.0</td>
<td></td>
</tr>
<tr>
<td>−10</td>
<td>92.7</td>
<td>93.9</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>−5</td>
<td>98.1</td>
<td>98.4</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>98.6</td>
<td>98.8</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>99.4</td>
<td>98.5</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>98.4</td>
<td>98.7</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>99.7</td>
<td>99.7</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>99.4</td>
<td>98.9</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>99.4</td>
<td>99.4</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>98.9</td>
<td>98.4</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>98.9</td>
<td>99.1</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.11: Descriptive statistics of the intelligibility scores separated by SNR.

Intelligibility. On this basis, these two data points were removed from further analysis.

**Overview of the intelligibility data**

Table 5.11 shows some general descriptive statistics of the intelligibility scores averaged across subjects, repeats, and interferer programmes. It is worth noting that all median scores below −20 dB SNR were 0% and all median scores above −15 dB SNR were 100%, indicating a relatively steep psychometric function.

**Comparison with pilot results**

It should be noted that the SRTs of the speech intelligibility data gathered in this experiment seem to be between −15 and −20 dB SNR, whereas the SRTs for the listening scenarios in the pilot experiment were between −5 and −10 dB SNR. This is quite a large difference in intelligibility and warrants explanation. The methodology of the pilot experiment presented listeners with a number of disadvantages relative to the main experiment.

Firstly, the pilot experiment featured target phrases selected from up to 37 speakers (across three corpora) whereas the the main experiment featured a single target speaker, thus the subjects were able to learn to recognise the target speaker’s vocal characteristics in the main experiment whereas this was not possible in the pilot. It is difficult to quantify the size of this effect, but as a point of comparison in (Nygaard
et al. 1994) a speech in noise intelligibility test was conducted using a group of listeners familiar with the target speakers and a control group of listeners (unfamiliar with the target speakers). An improvement of around 10% intelligibility was found for positive SNRs and a smaller improvement of around 5% was found at -5 dB SNR.

Secondly, the target speakers in the pilot experiment had a range of different accents with which listeners were likely to have had varying degrees of prior experience, whereas in the main experiment a target speaker with a local regional accent was selected in order to maximise the likelihood of accent familiarity with the selected listeners.

Thirdly, many of the target speakers in the pilot experiment were male (although a few were female) while the target speaker in the main experiment was female. This was a deliberate selection made to maximise baseline intelligibility, as it has been widely noted that female speech tends to score higher on speech intelligibility test, e.g. (Brungart et al. 2001; Barker and Cooke 2007). In (Barker and Cooke 2007) the difference in SRT due to gender was around 2 dB SNR.

Fourthly, three target corpora were used in the pilot experiment (since one of the objectives was to identify differences between these corpora) whereas only one target corpus was used for the main experiment; the differences in prosody (see Amano-Kusumoto and Hosom (2011) for a review of such effects) and informational content, are likely to have diminished listener performance in the pilot experiment relative to the main experiment.

The design principle underlying the main experiment was to maximise baseline intelligibility in order to minimise the effect of uncontrolled factors (such as vocal characteristics), rather than to confound (or investigate) these other variables as in the pilot. The difference between these two approaches carries another, inherent factor with it: the variability of the content (and therefore the listener’s expectations). Considering this difference in approach it would be surprising if the SRTs of the pilot and main experiment were very similar. The results of the main experiment, however, must be interpreted with this consideration in mind, i.e. these intelligibility scores are expected to be close to optimal, such that in ecologically valid listening scenarios there will always be potential for external factors (such as a listener’s lack of familiarity with the target speaker’s vocal characteristics) to severely degrade intelligibility, yet these factors cannot reasonably be incorporated into a model of speech intelligibility designed to evaluate a perceptually optimised sound zoning system.

At this point two objections might be raised to the usage of the intelligibility scores gathered for the purposes as stated. The first is that very high intelligibility scores fall very close to 100% at SNRs below 0 dB, yet sound zoning systems would be expected to perform much better than this. The second is that the intelligibility scores are very high for low SNRs in general, and thus may be so optimal as to be unrepresentative of ecologically valid listening scenarios.
In response to the first objection it should be stated that this is broadly correct, and that for a wide range of scenarios it can simply be assumed that while performance exceeds 0 dB SNR, and assuming no quality degradations imposed by the sound zoning system, intelligibility should remain very high. This implies that the intelligibility of the target programme, therefore, would only be useful as a predictor of acceptability or distraction for highly unacceptable and distracting scenarios. It does not, however, rule out the possibility of using the intelligibility of the interferer programme as a predictor of distraction or acceptability. The gathered data may still be used to evaluate and train models for the prediction of intelligibility, with which the intelligibility of interferer programmes in speech-on-speech listening scenarios can be estimated, and any potential relationship between this and distraction or acceptability can be evaluated.

The second objection concerns the usefulness of intelligibility data which has been optimised to minimise the effect of confounds (such as talker-listener dialect interactions) which are expected in ecologically valid listening scenarios. It is true that the intelligibility data, being optimised in this way, should be interpreted in consideration of this. Therefore if, for a particular listening condition, a very low intelligibility is reported at -20 dB SNR and a very high intelligibility is reported at -10 dB SNR, then it can be stated with great confidence that the intelligibility will be low for similar cases at -20 dB SNR, while for similar cases at -10 dB SNR it could only be stated that intelligibility will be high in the absence of a variety of confounds. The wide variation in ecologically valid speech signals ensures that some degree of variability is necessary for any estimates of general intelligibility estimates.

**SRTs and logistic fits**

Tables 5.12 and 5.13 show word-based SRTs and sentence-based SRTs for each interferer programme (for data averaged across repeats but not across subjects) calculated by linear interpolation and by fitting a logistic function to the data as in (Festen and Plomp 1990; Kjems et al. 2009). The equation for the logistic function is:

\[ I = \frac{1}{1 + e^{\frac{\alpha - \text{SNR}}{\beta}}} \]  

(5.3)

such that the slope of the logistic function is determined by \( \beta \) and the offset by \( \alpha \).

The logistic regression was performed using the Matlab curve fitting tool. Using a rearrangement of this formula (and setting \( I \) to 50%) the SRT can be calculated as:

\[ \text{SRT} = \alpha - \beta \left( \ln \left( \frac{1}{0.5} - 1 \right) \right) \]  

(5.4)

\[ = \alpha - \beta (\ln(1)) \]
Chapter 5: Speech Intelligibility Experiment

<table>
<thead>
<tr>
<th>Interferer</th>
<th>SRT (lin)</th>
<th>SRT (log)</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>adj. $R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>−15.6</td>
<td>−15.54</td>
<td>−15.54 ± 0.28</td>
<td>1.601 ± 0.296</td>
<td>0.9879</td>
<td>0.0477</td>
</tr>
<tr>
<td>Casino</td>
<td>−19.6</td>
<td>−19.21</td>
<td>−19.21 ± 0.58</td>
<td>2.919 ± 0.511</td>
<td>0.9561</td>
<td>0.0820</td>
</tr>
<tr>
<td>Explosion</td>
<td>−17.8</td>
<td>−17.91</td>
<td>−17.91 ± 0.1</td>
<td>1.757 ± 0.247</td>
<td>0.9800</td>
<td>0.0595</td>
</tr>
<tr>
<td>Interview</td>
<td>−22.7</td>
<td>−21.98</td>
<td>−21.98 ± 0.84</td>
<td>3.759 ± 0.749</td>
<td>0.9129</td>
<td>0.1064</td>
</tr>
<tr>
<td>News</td>
<td>−25.9</td>
<td>−25.19</td>
<td>−25.19 ± 1.16</td>
<td>4.184 ± 1.054</td>
<td>0.7974</td>
<td>0.1379</td>
</tr>
<tr>
<td>Pompeii</td>
<td>−14.8</td>
<td>−14.88</td>
<td>−14.88 ± 0.21</td>
<td>1.509 ± 0.339</td>
<td>0.988</td>
<td>0.0479</td>
</tr>
<tr>
<td>Traffic</td>
<td>−21.4</td>
<td>−20.84</td>
<td>−20.84 ± 0.4</td>
<td>1.126 ± 0.468</td>
<td>0.9771</td>
<td>0.0598</td>
</tr>
<tr>
<td>Mozart</td>
<td>−19.7</td>
<td>−19.68</td>
<td>−19.68 ± 0.39</td>
<td>1.768 ± 0.469</td>
<td>0.9645</td>
<td>0.0767</td>
</tr>
</tbody>
</table>

Table 5.12: SRTs for each interferer programme based on word error and calculated by linear interpolation and logistic fitting.

<table>
<thead>
<tr>
<th>Interferer</th>
<th>SRT (lin)</th>
<th>SRT (log)</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>adj. $R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>−10.3</td>
<td>−11.15</td>
<td>−11.15 ± 0.83</td>
<td>2.441 ± 0.733</td>
<td>0.9209</td>
<td>0.1289</td>
</tr>
<tr>
<td>Casino</td>
<td>−14.0</td>
<td>−13.54</td>
<td>−13.54 ± 1.49</td>
<td>3.418 ± 1.313</td>
<td>0.7953</td>
<td>0.1966</td>
</tr>
<tr>
<td>Explosion</td>
<td>−15.0</td>
<td>−15.06</td>
<td>−15.06 ± 0.49</td>
<td>1.532 ± 0.763</td>
<td>0.9404</td>
<td>0.1093</td>
</tr>
<tr>
<td>Interview</td>
<td>−9.2</td>
<td>−10.9</td>
<td>−10.9 ± 2.52</td>
<td>5.367 ± 2.219</td>
<td>0.6484</td>
<td>0.2645</td>
</tr>
<tr>
<td>News</td>
<td>−15.6</td>
<td>−10.82</td>
<td>−10.82 ± 3.68</td>
<td>9.67 ± 3.373</td>
<td>0.5595</td>
<td>0.2863</td>
</tr>
<tr>
<td>Pompeii</td>
<td>−9.5</td>
<td>−10.91</td>
<td>−10.91 ± 0.46</td>
<td>fixed at 1</td>
<td>0.9249</td>
<td>0.1271</td>
</tr>
<tr>
<td>Traffic</td>
<td>−15.6</td>
<td>−13.33</td>
<td>−13.33 ± 1.68</td>
<td>4.314 ± 1.480</td>
<td>0.7980</td>
<td>0.1969</td>
</tr>
<tr>
<td>Skyfall</td>
<td>−16.7</td>
<td>−16.52</td>
<td>−16.52 ± 1.39</td>
<td>2.557 ± 1.228</td>
<td>0.7598</td>
<td>0.2149</td>
</tr>
<tr>
<td>Mozart</td>
<td>−16.5</td>
<td>−15.97</td>
<td>−15.97 ± 1.28</td>
<td>1.096 ± 1.297</td>
<td>0.8478</td>
<td>0.1756</td>
</tr>
</tbody>
</table>

Table 5.13: SRTs for each interferer programme based on sentence error and calculated by linear interpolation and logistic fitting.

$$= \alpha$$

For the calculation of the logistic function fitting the Pompeii interferer $\beta$ was fixed at 1. Although it is possible to fit a logistic function to this data with marginally better accuracy using a value $\beta < 0.01$ such fits are made at the cost of assuming a very wide variance in $\alpha$ which renders them effectively useless for our purposes. Moreover, as $\beta$ falls below 0.5 the logistic curve begins to differ very little from a step function.

As previously discussed in section 5.3.3, SRTs for SSN have been reported ranging from $-6$ to $-14$ dB SNR. For reasons discussed in the previous section, the intelligibility scores gained in this experiment were expected to be close to optimal (i.e. higher intelligibility scores are unlikely for similar repeated conditions). As a result the SRTs found for SSN, −15.54 and −11.15 dB SNR for word and sentence-based respectively,
were close to but slightly lower than SRTs reported in similar experiments in the literature.

In (Festen and Plomp 1990), SRTs are reported for interferers of steady-state noise, modulated noise, and interfering voice. The resulting (sentence-based) SRTs were −5, −9, and −12 dB SNR respectively. To the latter of these a comparison may be drawn with the News interferer programme which had a (sentence-based) SRT of −10.82 dB SNR. This value greatly differs from the linearly interpolated SRT of −15.6 dB SNR. This is explained by the high variability of scores across SNRs, and implies that the logistic fit gives a more appropriate estimate for the SRT. Festen and Plomp (1990) report that the gradient of the intelligibility curve for the speech interferer was 12.0% per dB. The gradient of the curve at the SRT is given by the differential of the logistic function:

\[
\frac{d}{dx} \left( \frac{1}{1 + e^{\frac{-\text{SRT}}{\beta}}} \right) = \frac{\alpha - \text{SRT}}{\beta} \left( 1 + \frac{\alpha - \text{SRT}}{\beta} \right)^{2} = \frac{1}{\beta(1 + 1)^{2}} = \frac{1}{4\beta} \tag{5.5}
\]

Thus the gradient for the logistic function fitted to the news interferer is \( \frac{1}{4\beta} = \frac{1}{4 \times 9.67} = 0.0259 \), i.e. 2.6% per dB. This is substantially shallower than the gradient reported by Festen and Plomp (1990). Although the SRTs are similar, the difference in gradient might be explained by the only moderate fit of the logistic function to the sentence-based data (adjusted \( R^2 = 0.56 \)). The data does, however, agree with the results of Festen and Plomp (1990) in that temporally steady interferers produce intelligibility curves with steeper gradients than temporally modulated interferers. The gradients for the interferers featuring speech were between 2.6 and 8.0% per dB, whereas the gradients for the interferers not featuring speech were 9.6 and 22.7% per dB.

ANOVA

Tables 5.14 and 5.15 show the ANOVAs for the intelligibility and keyword intelligibility scores respectively. Shapiro Wilk tests showed that the residuals of these ANOVAs were not normally distributed. An inspection of the histograms showed a leptokurtic normal
Figure 5.11: Sentence based SRTs for normal hearing listeners adapted from Festen and Plomp (1990). The squares show SRTs for a speech interferer, the circles show SRTs for a modulated noise interferer, and the diamonds show SRTs for steady-state noise.

Table 5.14: An ANOVA of intelligibility scores with the non-significant (at 95% confidence) interaction removed

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>Sig</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>5420.848</td>
<td>&lt;0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>SNR</td>
<td>14</td>
<td>637.692</td>
<td>&lt;0.001</td>
<td>0.991</td>
</tr>
<tr>
<td>Interferer</td>
<td>8</td>
<td>17.135</td>
<td>&lt;0.001</td>
<td>0.741</td>
</tr>
<tr>
<td>Subject</td>
<td>6</td>
<td>5.469</td>
<td>&lt;0.001</td>
<td>0.296</td>
</tr>
<tr>
<td>SNR by Interferer</td>
<td>112</td>
<td>13.801</td>
<td>&lt;0.001</td>
<td>0.489</td>
</tr>
<tr>
<td>Interferer by Subject</td>
<td>48</td>
<td>1.500</td>
<td>0.016</td>
<td>0.043</td>
</tr>
<tr>
<td>SNR by Subject</td>
<td>84</td>
<td>2.877</td>
<td>&lt;0.001</td>
<td>0.130</td>
</tr>
</tbody>
</table>
distribution, however, and the violation of the normality assumption (in this way) is known to have only a small effect for parametric tests using $\alpha > 0.001$ (Glass et al. 1972). The results of the two ANOVAs are very similar and show that SNR, interferer programme, and subject were significant, as well as all two way interactions. The effect sizes of all two way interactions are relatively small except for the interaction between SNR and interferer programme. This interaction simply indicates that as SNR changes the resultant change in intelligibility is not constant across all interferer types.

As table 5.16 shows, the Tukey HSD post hoc test indicates very clearly that the news interferer resulted in much greater intelligibility scores than the other interferer programmes. Though further groupings are less distinct, the Pompeii and SSN tracks also seem to have produced similar intelligibility scores. It is also important to consider the significant two way interaction between SNR and interferer programme, however, and fig. 5.12 shows this interaction.

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>Sig</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>5261.147</td>
<td>&lt;0.001</td>
<td>0.999</td>
</tr>
<tr>
<td>SNR</td>
<td>14</td>
<td>700.688</td>
<td>&lt;0.001</td>
<td>0.992</td>
</tr>
<tr>
<td>Interferer</td>
<td>8</td>
<td>15.912</td>
<td>&lt;0.001</td>
<td>0.726</td>
</tr>
<tr>
<td>Subject</td>
<td>6</td>
<td>5.792</td>
<td>&lt;0.001</td>
<td>0.332</td>
</tr>
<tr>
<td>SNR by Interferer</td>
<td>112</td>
<td>11.717</td>
<td>&lt;0.001</td>
<td>0.448</td>
</tr>
<tr>
<td>Interferer by Subject</td>
<td>48</td>
<td>1.519</td>
<td>0.013</td>
<td>0.043</td>
</tr>
<tr>
<td>SNR by Subject</td>
<td>84</td>
<td>2.213</td>
<td>&lt;0.001</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Table 5.15: An ANOVA of keyword intelligibility scores with all non-significant (at 95% confidence) factors and interactions removed.

<table>
<thead>
<tr>
<th>Interferer</th>
<th>N</th>
<th>group 1</th>
<th>group 2</th>
<th>group 3</th>
<th>group 4</th>
<th>group 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pompeii</td>
<td>210</td>
<td>69.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSN</td>
<td>210</td>
<td>70.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic</td>
<td>210</td>
<td>72.37</td>
<td>72.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explosion</td>
<td>210</td>
<td>73.82</td>
<td>73.82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Casino</td>
<td>209</td>
<td>74.67</td>
<td>74.67</td>
<td>74.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mozart</td>
<td>210</td>
<td>75.43</td>
<td>75.43</td>
<td>75.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skyfall</td>
<td>210</td>
<td>76.76</td>
<td>76.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>209</td>
<td>77.37</td>
<td>77.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80.80</td>
</tr>
</tbody>
</table>

Table 5.16: Homogeneous subsets based on a Tukey HSD post hoc test of the intelligibility scores.
It can be seen that the interaction between SNR and interferer programme slightly complicates the interpretation. For example, although the upwards slope of intelligibility scores for the news interferer begins at a much lower SNR than the curves for the other interferers, some interesting effects are seen at very low and very high SNRs. At very high SNRs the news interferer seems to have much wider confidence intervals than other interferers, indicating a greater number of word errors across subjects than for other interferer programmes. This is likely to be due to temporal sparsity of speech which entails that the target speech never completely masks the interferer, leaving some room for confusions and errors even at high SNRs. At the lowest SNR the mean intelligibility score for the news interferer appears to be higher than the score at $-30$ dB SNR, although the confidence intervals overlap. This is likely to be due to the temporal gaps in the speech, which allowed the subjects to identify occasional correct target words even at very poor SNRs because the brief silences in the interferer programme effectively produce brief instances of ideal SNR for the target speech (i.e. complete masking of speech by another speech programme is very unlikely). Additionally, the upwards slope of the intelligibility scores for the news interferer programme is much more gradual and spans a range of 25 dB whereas, by means of comparison, the upward slope of intelligibility scores for the Pompeii interferer spans only 10 dB.

More generally it seems likely that the news interferer is a specific and extreme case of a group of interferer programmes: those featuring speech. The interferer programmes featuring speech all appear to share certain qualities:

1. the upward slope of intelligibility scores is more gradual, and
2. the intelligibility scores have wider confidence intervals after the mean scores plateau out near 100%.

By contrast, interferer programmes not featuring speech appear to have the inverse
qualities:

1. the upward slope of intelligibility scores is sharp, and

2. the intelligibility scores have small confidence intervals after the mean scores plateau out near 100%.

Figure 5.13 shows these differences more clearly; for each of the non-speech interferer curves almost the entire range of speech intelligibility scores is covered across a range of around 10 dB, while the speech interferer curves achieve this over a range of 20-25 dB SNR. It is noteworthy that the wide variance of intelligibility scores for the news interferer at $-35$ dB is not found for the other three interferer programmes. This is because while the subjects were able to listen in the gaps of the much louder news interferer programme, the same opportunity was not available for the other three interferers which all featured background music or sounds which, for very low SNRs, remove any possibility of listening to the target in the gaps.

The non-speech interferer programmes, although all featuring similar very steep upward slopes, have SRTs up to 10 dB apart. This difference is large and, as a case in point, when the SNR was $-20$ dB the intelligibility for the Pompeii interferer was close to zero while the intelligibility for the Skyfall interferer was above 90%. These differences imply that a very simplistic model of speech intelligibility which simply maps a curve to the data across SNRs is unlikely to make predictions which are more accurate than within 5 dB of the correct SRT, even if separate curves are used for speech interferer and non-speech interferer cases. Even so such a model is worth constructing to provide a
baseline against which to test the predictions of more sophisticated speech intelligibility models in later work.

**Predicting intelligibility scores**

The primary motivation for obtaining the speech intelligibility data was to investigate any possible link between acceptability and intelligibility, and if such a link were present to ultimately test or develop a speech intelligibility model for use. Without reference to existing speech intelligibility models, however, a simple speech intelligibility model can immediately be generated based on the available data by mapping a function to the observed intelligibility scores. This is worthwhile because it provides a baseline against which other, more sophisticated, speech intelligibility models may later be tested.

The most suitable choice of function would be a logistic function since they describe proportion or population growth. In section 5.5.1 logistic functions were fitted to individual interferers using eq. (5.3). Figure 5.14 shows these logistic curves, and it can be seen that while most functions had a steep gradient ($\beta < 1.8$ for word-based logistic fits) a few did not. Those curves with shallower gradients ($\beta > 2.9$) are those which featured speech interferers, namely: News, Interview, Traffic, and Casino interferers. Of these four logistic functions the fit to the Traffic interferer is most similar to the non-speech interferers which is likely because this interferer included pop-like background music. The logistic fits to the interferers featuring speech are also similar to each other in another manner: the variance of $\alpha$ is much greater for these ($\alpha > 0.5$) than for the fits to interferers not featuring speech ($\alpha \leq 0.4$). This is precisely as expected, and is likely the result of both the informational content of the target and interferer programmes interacting and the variabilities in the prosody of the interfering speech. On the basis of this, it seems appropriate to group the interferers by presence of speech. Under this grouping the Traffic interferer fits only approximately into the speech interferer group, and may be better dealt with as a marginal case.

Logistic functions were fitted to the mean intelligibility scores (averaged across subjects, repeats, and the four interferer programmes featuring speech) for each SNR using the Matlab curve fitting tool. Figure 5.15 shows these two logistic functions. Since the 95% confidence intervals overlap there is insufficient evidence that the models are significantly different.

For the function predicting intelligibility in the presence of an interferer featuring speech, the coefficients are $\alpha = -20.54 \pm 0.55$ and $\beta = 4.126 \pm 0.482$. The model fits with $R^2 = 0.8733$ and adjusted $R^2 = 0.873$, and has a RMSE of 0.1316 (i.e. 13.2% error). Since the function only varies significantly between $-35$ and 5 dB, however, a calculation of RMSE and $R^2$ which uses the full SNR is likely to be optimistic; a recalculation for only the data within this range gives RMSE = 0.1646 (i.e. 16.5% error) and adjusted $R^2 = 0.8414$. 

149
Chapter 5: Speech Intelligibility Experiment

Figure 5.14: Logistic curves fitted to intelligibility scores separated by interferer programme.

For the function predicting intelligibility in the presence of an interferer not featuring speech the equation is identical but with $\alpha = -17.7 \pm 0.28$ and $\beta = 0.234 \pm 0.232$. The model fits with $R^2 = 0.9429$ and adjusted $R^2 = 0.9428$, and has RMSE $= 0.1001$ (i.e. 10.0% error). When considering only the data within the range $-35$ to $5$ dB SNR the RMSE is $0.1287$ (i.e. 12.9% error), with adjusted $R^2 = 0.9227$.

Given that the confidence bounds on the two logistic fits overlap it was considered worthwhile investigating a single logistic fit to the data, to see if the performance was very different. The resulting logistic fit was a function with $a = -18.75 \pm 0.3$ and $b = 3.151 \pm 0.263$. The model fits with $R^2 = 0.9022$ and adjusted $R^2 = 0.9021$, and has RMSE $= 0.1244$ (i.e. 12.4% error). When considering only the data within the range $-35$ to $5$ dB SNR the RMSE is $0.1578$ (i.e. 15.8% error), with adjusted $R^2 = 0.8726$.

This singular logistic fit does not perform very differently from the fit to the speech interferer data, but performs slightly worse than the fit to the data for interferers not featuring speech. This is because the variance for the interferers featuring speech is greater. Even so, this logistic fit is not very inferior and will be a useful benchmark against which speech intelligibility models may tested. Thus the current model for predicting intelligibility is:

$$I = \frac{1}{1 + e^{-\frac{18.75 - \text{SNR}}{3.151}}}. \quad (5.6)$$
Figure 5.15: 95% confidence intervals of the logistic functions fitted to the speech intelligibility scores for interferers featuring speech (blue shaded area) and interferers not featuring speech (red shaded area).
5.5.2 Distraction scores

Before an analysis of variance of the distraction scores is conducted it is prudent to remove any outliers and investigate the consistency of the subject’s scoring. In order to investigate the consistency of the subject’s scores for the distraction task the absolute difference was taken between the two data points for each subject, interferer programme, and SNR. Where this difference score was large for specific interferers (across multiple subjects) it may be an indication that it is difficult to judge a distraction score for a particular condition; by contrast where the difference score is large for only a single subject it may be an indication that the subject was not very consistent in the task.

The first, and simplest, insight into the consistency of distraction scoring can be observed in fig. 5.16. The average distraction score (across subjects, interferer programmes, and repeats) is displayed alongside the average difference between repeated scores (across subjects and interferer programmes). As would be expected the average scoring consistency appears to be better for very high and for very low SNRs since in these regions the subjects used values close to the edges of the scale. Specifically for mean distraction scores between 20% (+20 dB SNR) and 80% (−15 dB SNR) the average difference between repeated scores is approximately constant (at around 10%). Average scores, however, do not reveal any information about the consistency of specific subjects or interferer programmes.

Upon inspecting the difference scores across SNRs for each interferer programme and each subject there were no clear trends indicating that any individual subjects were particularly inconsistent or that particular interferer programmes were particularly difficult to consistently judge (see fig. 5.17 for example plots). A single outlier was found for subject 3 with the news interferer programme who, at −35 dB SNR, rated the distraction at 5 and 99 for the two repeats. Since at −35 dB the target programme is very difficult to hear, and considering that all other data points for this condition were above 80, it is likely that the score of 5 for this condition represents a user input error and was therefore removed from further analysis.

ANOVA

As with the intelligibility data, the assumptions of normal distribution and homogeneity of variances were not met for many of the conditions. For reasons discussed in section 5.3.2 the effect of this upon the interpretation of the ANOVA is expected to be negligible.

Table 5.17 shows an initial ANOVA conducted on distraction scores. The analysis indicated that all main effects and interactions were significant with moderate effect sizes. Since the three way interaction between subject, SNR, and interferer was significant with a moderate effect size it is difficult to interpret the data. Upon
Chapter 5: Speech Intelligibility Experiment

Figure 5.16: The black, solid line represents the average (across subjects, interferer programmes and repeats) distraction score for each SNR tested, and the blue, dot-dashed line represents the average (across subjects and interferer programmes) difference between the score given in each repeat. Error bars show the 95% confidence intervals.

Figure 5.17: Differences between distraction scores for each subject and SNR for the explosion interferer (left) and the interview interferer (right)
inspection of the distraction scores across SNRs, subjects and interferers it became apparent that, as with speech intelligibility scores, the interferers might reasonably be grouped according to whether they featured speech. This assessment was supported further by inspection of the Tukey HSD post hoc test carried out on interferer programmes which indicated that the interferers featuring speech were significantly more distracting than those which did not (see table 5.18).

Table 5.19 shows the ANOVA when excluding interferers featuring speech (and excluding SSN, which appeared to show very different distraction scores). While the interferer programme was no longer significant, the interaction between SNR and interferer programme was significant with a small effect size. It is notable that the effect size of the interaction between SNR and interferer programme is less than the effect size of both the subject and the interaction between subject and SNR.

Post hoc Tukey HSD tests were carried out to investigate the effects further. The tests indicated that there was little difference between distraction scores for 30 and

<table>
<thead>
<tr>
<th>Interferer</th>
<th>N</th>
<th>group 1</th>
<th>group 2</th>
<th>group 3</th>
<th>group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>210</td>
<td>47.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explosion</td>
<td>210</td>
<td>48.62</td>
<td>48.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skyfall</td>
<td>210</td>
<td>49.10</td>
<td>49.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mozart</td>
<td>210</td>
<td>50.58</td>
<td>50.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pompeii</td>
<td>210</td>
<td>50.86</td>
<td>50.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>News</td>
<td>209</td>
<td>52.17</td>
<td></td>
<td>52.17</td>
<td></td>
</tr>
<tr>
<td>Casino</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td>53.57</td>
</tr>
<tr>
<td>Traffic</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td>53.74</td>
</tr>
<tr>
<td>Interview</td>
<td>210</td>
<td></td>
<td></td>
<td></td>
<td>54.25</td>
</tr>
<tr>
<td>Sig.</td>
<td></td>
<td>0.870</td>
<td>0.123</td>
<td>0.564</td>
<td>0.200</td>
</tr>
</tbody>
</table>

Table 5.18: Homogeneous subsets based on a Tukey HSD post hoc test of the distraction scores.
35 dB SNR and little difference between distraction scores for $-25$, $-30$, and $-35$ dB SNR. Although this might indicate that there are negligible differences to perceived distraction beyond these SNRs it might also simply be a result of subjects scaling their responses to the SNRs available (i.e. if a wider range of SNRs had been used these distraction scores might have been significantly different).

Figure 5.18 shows the effect of SNR upon distraction scores for each subject across the four interferer programmes which did not feature speech (and excluding SSN). Confidence intervals have not been plotted because each mean score is an average across only two data points (the two repeats per condition per subject), and thus the confidence intervals are very wide. The scores produced by subject 5 were quite different from the other subjects within the range $-10$ to $+30$ dB SNR, however the data was no less consistent than that of the other subjects and were fairly similar across the four interferer programmes. In support of this interpretation of the data fig. 5.19 shows Tucker1 plots for subject’s distraction scores grouped by the presence of speech in the interferer. In both cases the grouping of subjects is fairly close and subjects can be considered to be performing the same task.

Another ANOVA was conducted for the interferer programmes featuring speech (see table 5.20). The results were similar to the ANOVA for interferer programmes not featuring speech, except here a significant but very small interaction was found between interferer programme and subject. This is probably due to the very high distraction scores reported by subject 5 for the interview interferer programme; this is easily explained since during the experiment subject 5 reported finding the vocal characteristic of one of the male speakers in this interferer programme particularly annoying.

A post hoc Tukey HSD test again revealed that there was little difference between distraction scores for $30$ and $35$ dB SNR and little difference between distraction scores for $-25$, $-30$, and $-35$ dB SNR.

Figure 5.20 shows the effect of SNR upon distraction scores for each subject across the four interferer programmes featuring speech. The general trends of the scores across SNRs and subjects are similar, yet the individual data were much less closely grouped than for interferers not featuring speech. The explanation for this more spaced

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>600.386</td>
<td>&lt;0.001</td>
<td>0.990</td>
</tr>
<tr>
<td>SNR</td>
<td>14</td>
<td>216.504</td>
<td>&lt;0.001</td>
<td>0.973</td>
</tr>
<tr>
<td>Subject</td>
<td>6</td>
<td>10.756</td>
<td>&lt;0.001</td>
<td>0.434</td>
</tr>
<tr>
<td>SNR by Interferer</td>
<td>45</td>
<td>2.113</td>
<td>&lt;0.001</td>
<td>0.121</td>
</tr>
<tr>
<td>SNR by Subject</td>
<td>84</td>
<td>4.827</td>
<td>&lt;0.001</td>
<td>0.370</td>
</tr>
</tbody>
</table>

Table 5.19: An ANOVA of distraction scores for non-speech interferers (also excluding SSN) with all non-significant (at 95% confidence) factors and interactions removed.
Chapter 5: Speech Intelligibility Experiment

Figure 5.18: Mean distraction scores separated by subject across SNR for the four interferer programmes not featuring speech (and excluding SSN).

Figure 5.19: Tuckert plots for distraction scores for interferers excluding speech (left) and including speech (right). The SSN interferer was not included in this analysis.
distribution of scores across subjects is rather intuitive: when the interferer did not feature speech subjects found the task of rejecting the interferer programme equally easy because the information being conveyed by the interferer was not of the same kind as the information contained within the target speech. When the interferer featured speech, however, the informational content of the interferer was of the same kind as the informational content of the target; in these cases the subjects are required to identify and separate out the multiple speakers, a task about which subjects disagree much more on the difficulty. Considering this, as well as the significant interactions between SNR and subject and between interferer programme and subject, it is likely that highly accurate predictions of distraction could only be obtained by categorising listeners; unfortunately it is far from clear on what basis such a categorisation should be performed, and it seems likely that a great many factors (which may be difficult to obtain) would be required to inform the categorisation (such as familiarity with target speaker).

Finally it should be noted that subjects may have used the scale differently from one another and in order to test whether this has affected the analysis of distraction scores all data were z-transformed (separately for each subject) and ANOVAs were conducted on these data. The same effects were found in this data as with the untransformed scores.

**Simple predictions of distraction**

While these results are interesting it is sufficient, for the purpose of this work, to predict average distraction scores (especially where there are no clear and simple distinctions between subject groups). Figure 5.21 shows the mean distraction scores (across subjects and repeats) for each of the interferer programmes in their respective groupings.

In fig. 5.22 the average curve of each of these groups are shown alongside the average distraction scores for the SSN interferer. In general the three curves agree fairly well below −5 dB SNR (with the speech interferers being slightly less distracting). Above −5 dB the SSN interferer was least distracting and the interferers featuring speech
Figure 5.20: Mean distraction scores separated by subject across SNR for the four interferer programmes featuring speech.
Figure 5.21: Mean distraction scores, averaged across subjects and repeats, for each interferer programme in their respective groupings. Non-speech interferers are on the left, speech interferers are on the right.

were most distracting. This trend can be explained by considering that for SNRs below $-5$ dB the intelligibility of the target drops rapidly so the distraction scores are similarly high regardless of the interferer type (with a slightly lower distraction score for the speech interferers because they permit some listening in the gaps). By contrast, above $-5$ dB SNR the intelligibility of the target speech is very high and the distraction score likely represents the extent to which listeners believed their alternate audio programme would pull their attention away from the target speech; thus for any fixed SNR SSN, being random and thus containing no information, is uninteresting and would not direct attention away from the speech, non-speech interferers, containing a different type of information from the target speech, will be slightly more distracting, and speech interferers, containing the same type of information as the target speech, are likely to be most distracting.

To fit a logistic function to these curves a translation is applied to the previous logistic function used to account for the negative correlation between SNR and distraction. The formula is therefore:

$$D = 1 - \frac{1}{1 + e^{\alpha - \beta \cdot SNR}}$$

(5.7)

where $D$ is a distraction score ranging from 0 to 1. Fitting the appropriate logistic function to the speech interferer cases gives $\alpha = 2.366 \pm 0.6890$ and $\beta = 12.67 \pm 0.65$, which fits the mean (across all subjects and repeats and across the 4 interferers) distraction scores with RMSE = 0.1337, $R^2 = 0.8461$ and adjusted $R^2 = 0.8459$.  

159
Figure 5.22: Distraction scores for each of three interferer types: with speech, without speech, and SSN.
Fitting a logistic function to the non-speech interferer cases gives $\alpha = -1.415 \pm 0.5584$ and $\beta = 10.9 \pm 0.51$, which fits the mean distraction scores with RMSE = 0.1174, $R^2 = 0.8928$ and adjusted $R^2 = 0.8927$. These curves are shown in Fig. 5.23. Although the subjective distraction data for these groups were significantly different, the fitted logistic functions applied here have overlapping confidence intervals, and it is likely that these fits could be optimized further with more subjective data.

As with the intelligibility scores, it was worth comparing these logistic fits to a fit made on all data (without separating by the presence of speech in the interferer programme). The resulting fit had $\alpha = 0.4548 \pm 0.4527$ and $\beta = 11.87 \pm 0.42$, which fits the mean (across all subjects and repeats and across the 4 interferers) distraction scores with RMSE = 0.1288 (i.e. 12.9% error), $R^2 = 0.8646$ and adjusted $R^2 = 0.8645$.

With $\alpha$ so close to zero, the resulting fit implies that an SNR of 0 dB corresponds very well to 50% distraction. It was therefore considered worthwhile investigating a similar logistic fit which excluded the coefficient altogether. The resulting fit had $\beta = 11.84 \pm 0.42$, which fits the mean (across all subjects and repeats and across the 4 interferers) distraction scores with RMSE = 0.1289 (i.e. 12.9% error), $R^2 = 0.8643$ and adjusted $R^2 = 0.8643$. With such similar results, the simpler model is preferred. Thus the current model for predicting distraction is:
Table 5.21: Correlations between acceptability, masking, and SNRs.

<table>
<thead>
<tr>
<th></th>
<th>Acceptability</th>
<th>Masking</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Acceptability</strong></td>
<td>Pearson Correlation</td>
<td>1.000</td>
<td>−0.158</td>
</tr>
<tr>
<td></td>
<td>Spearman’s rho</td>
<td>1.000</td>
<td>−0.158</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1889</td>
<td>1889</td>
</tr>
<tr>
<td><strong>Masking</strong></td>
<td>Pearson Correlation</td>
<td>−0.158</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Spearman’s rho</td>
<td>−0.158</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1889</td>
<td>1889</td>
</tr>
<tr>
<td><strong>SNR</strong></td>
<td>Pearson Correlation</td>
<td>0.592</td>
<td>−0.399</td>
</tr>
<tr>
<td></td>
<td>Spearman’s rho</td>
<td>0.592</td>
<td>−0.399</td>
</tr>
<tr>
<td></td>
<td>Sig. (2-tailed)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>1889</td>
<td>1889</td>
</tr>
</tbody>
</table>

\[ D = 1 - \frac{1}{1 + e^{\frac{\text{SNR}}{11.87}}} \]  \hspace{1cm} (5.8)

5.5.3 Acceptability and masking data

To investigate masking and acceptability, correlations were first calculated between these and SNR (see table 5.21). As with the pilot experiment significant correlations were found for all of these pairs. The magnitude of the correlations is somewhat less than expected, however, which is likely a consequence of the range of SNRs being considered (i.e. no correlation would be expected between SNR and masking for positive SNRs, thus by including these SNRs the correlation is smaller than expected). The correlation analysis for masking and SNR was repeated including only data between −35 and −15 dB SNR (because −15 dB SNR had all zeros for the binary masking data), and the new correlation was −0.469 for both Pearson’s correlation and Spearman’s Rho with a significance of <0.001 and \( n = 629 \). The correlation analysis for acceptability and SNR was repeated including only data between −10 and 35 dB SNR (because −10 dB SNR had all zeros for the binary acceptability data), and the new correlation was 0.515 for both Pearson’s correlation and Spearman’s Rho with a significance of <0.001 and \( n = 1260 \).

Acceptability

To investigate the acceptability data more thoroughly the binary scores were converted into scalar values representing the proportion likelihood acceptable (calculated by
In section 5.1.6 a hypothesis was stated which expected a difference in acceptability for interferer programmes featuring speech and interferer programmes not featuring speech (relative to their masking thresholds). To consider this, as with previous measures, the acceptability scores were grouped by interferer programme type according to whether the interferer programme contained speech or not. Figure 5.24 shows the acceptability scores for programmes grouped in this way. Close inspection of the acceptability scores for interferer programmes not featuring speech reveals that while the SSN and explosion interferers have acceptability scores similar to the other interferers for most of the SNR range they have higher acceptability scores between 15 and 30 dB SNR. This can be explained by considering that for very high SNRs (e.g. 35 dB and above) the interferer programme is nearly inaudible and therefore very acceptable such that the interferer programme is irrelevant, and for moderate and low SNRs (15 dB SNR and below) the interferer is sufficiently loud that the target is rarely acceptable and thus the interferer programme is again irrelevant; in the range between 15 and 30 dB SNR, however, the interferer is neither perceptually ‘very quiet’ nor ‘far too loud’ but within a range where the informational content of the interferer programme can therefore have a larger effect and since SSN and explosion interferers are relatively random signals they feature much less information than the three music interferer programmes.

Two distinct groups of acceptability scores were found when considering subject differences with subjects 1-4 indicating greater acceptability than subjects 5-7 for the same listening conditions. Splitting by subjects in this way and by interferer programmes as previous indicated produces the curves shown in fig. 5.25. In fig. 5.26 these acceptability scores are averaged for each group and plotted side by side. The data indicates that when the interferer programme did not feature speech it was more likely to be considered acceptable at lower SNRs than when the interferer programme
Chapter 5: Speech Intelligibility Experiment

Figure 5.25: Acceptability scores separated by SNR and interferer programme for interferer programmes separated by whether they include speech. Averages across subjects 1, 2, 3, and 4 are shown on the left, and averages across subjects 5, 6, and 7 are shown on the right. Note that the mean acceptability scores for the news interferer for subjects 5, 6, and 7 were identical to the scores for the traffic interferer for the same subjects.

did feature speech, however this effect appears to be smaller than the difference between acceptability scores for the two groups of subjects (i.e. the ‘tolerance’ of the subject appears to have a larger effect on acceptability). Although the temporal sparseness of speech interferers allows for greater intelligibility at very low SNRs (especially below $-10$ dB SNR), the common modality of information between target and interferer causes the listening scenario to be more distracting at SNRs above $-5$ dB SNR.

**Masking**

As with the binary acceptability scores, the masking data were converted into scalar values representing the proportion likelihood masked. By observing the masking scores averaged across interferers but separated by subject it was possible to investigate whether listeners were performing similarly. Figure 5.27 shows that the data from subject 2 were substantially different from the data from other subjects and closer inspection of data for each interferer confirmed that subject 2 was indeed consistently marking the target programme as inaudible at much higher SNRs than the other
Figure 5.26: Acceptability scores separated into two subject groups and according to the presence of speech in the interferer programme, then averaged across interferer programmes, subjects and repeats within each group. Solid lines indicate acceptability scores for interferer programmes featuring speech and dash-dotted lines indicate acceptability scores for interferers which did not feature speech.
Figure 5.27: Masking scores separated by SNR and subject but averaged across interferer programmes. The black dot-dashed line shows that the data from subject 2 differed greatly from the other subjects.

subjects. Since this subject reported no hearing difficulties, it is possible that the instructions were misunderstood and that this subject was selecting ‘inaudible’ for those occasions where the target speech was audible but where none of the words could be discerned. This data was excluded from further analysis.

Considering the effect of interferer programme upon the masking data from subjects 1 and 3-7 there is no reason to expect that the interferer programmes could be simplistically grouped by whether they contain speech, as with distraction and acceptability scores. It is expected, however, that those programmes comprised primarily of speech would be poor maskers (i.e. the SNR would need to be much lower before the target is masked) compared to those programmes which do not feature similar temporal gaps. Figure 5.28 shows the average masking score, across subjects (excluding subject 2) and repeats, for each interferer programme. The average score for the interview, news, and Mozart interferer programmes was 0 even at $-35$ dB SNR, which indicates that a masking threshold cannot be inferred for those programmes from this data (beyond the statement that the masking thresholds must be below $-35$ dB). The news interferer, being entirely comprised of speech, is unlikely to mask other speech due to the temporal gaps, however the interview interferer included a background
Chapter 5: Speech Intelligibility Experiment

musical track which was expected to mask the target speech at the lowest SNRs. Even more surprising is that the Mozart interferer programme which was comprised of relatively consistent orchestral programme material was never able to mask the speech target; from informal listening it appears that the Mozart interferer programme was unable to mask the sibilance of certain phonemes in the speech.

The SSN was the most effective masker which is easily explained since it has the average frequency spectrum of the target speech and no temporal gaps; this is likely to be close to optimal as a masker for these target speech phrases.

The casino interferer programme was also a very effective masker; although the casino programme was primarily a conversation between a male and female speaker (and therefore included temporal gaps) it also included background noise comprised of casino Foley, effects and low level chatter. When the SNR was sufficiently low the subjects were likely unable to discern this background chatter from the target speech, and thus reported the target as masked even though in the strictest sense it may still have been audible (i.e. the subjects likely succeeded at detection but not discrimination or recognition). This also explains why the curve was shallower than those for most other interferer programmes (90% of the scale was covered over 15 dB compared to 5 dB for the skyfall interferer or 10 dB for the SSN); at −25 and −30 dB SNR some subjects were likely able to discern the target speech from the background chatter while others were not.

The masking curve for the explosion interferer has a slightly unusual shape which, upon informal listening, appears to be somewhat determined by the presence of sibilance in the target speech. Although the Masking score at −35 dB SNR is 50% the results indicate that data was split perfectly across the two different target phrases. The first, “the hogs were fed chopped corn and garbage”, was always marked as inaudible and had little sibilance with the word ‘hogs’ occurring simultaneously with an explosion in the interferer, whereas the second, “the rope will bind the seven books tightly”, featured more sibilance towards the end of the sentence during the sound of helicopter rotors and music in the interferer programme. While this does not significantly affect the capacity to use this data to validate masking threshold predictions (since the predictions would need to be made for each target-interferer programme combination) it does imply that two target phrases were perhaps not sufficient to properly confound the effect of the target phrase for the purpose of investigating average masking threshold curves for target speech. As a result these average masking curves should be considered as indicative only, and it should be carefully noted that the target phrase is an important factor in determining the masking threshold, particularly for interferers which are highly impulsive.
Figure 5.28: Masking scores separated by interferer programme and SNR but averaged across repeats and all subjects excluding subject 2. The masking scores for the interview, news, and mozart interferer programmes were zero for all SNRs.
5.5.4 Interferer intelligibility

It was noted in section 5.1.2 that in addition to the use of target programme intelligibility for setting a lower boundary of acceptability, intelligibility may also be useful where auditory interference scenarios feature a speech interferer if a relationship is found between acceptability and the interferer intelligibility. This is plausible because when both the target and interferer contain speech there may be confusion or informational masking which decreases the acceptability of the listening scenario; where the interferer is unintelligible, however, the listening scenario is likely to improve.

Since the data gathered in this experiment were of the intelligibility of the target programme and the acceptability of the listening scenario, data do not exist to directly investigate this relationship. It is reasonable, however, to infer some general results from those cases featuring only speech in the target and interferer programmes. The news interferer featured only speech, so this case could be selected and mirrored in the line $SNR = 0$ dB to consider the interferer intelligibility. The logistic fit to intelligibility scores for the news interferer cases had an SRT of $-25.19$ dB, thus the inferred SRT for the interferer intelligibility would have a SRT of $25.19$ dB. Figure 5.29 shows the news interferer intelligibility scores, and the logistic fit to this data, reversed in the line $SNR = 0$ dB, along with non-transposed acceptability and distraction scores averaged across all cases.

The intelligibility scores, and the logistic fit to them, do not seem match the distraction scores very well. The distraction scores begin falling from 1 almost immediately after any increase in SNR from -35 dB, whereas the intelligibility scores do not fall until after 0 dB SNR. The gradients of these slopes also seem to differ.

The mean acceptability scores correspond slightly better with the intelligibility scores; both begin changing at approximately 0 dB SNR, and by 35 dB the intelligibility scores drop below 10% while the acceptability scores near 80%. The gradients of the slopes do not match up as well as the range, however. The correlation between the intelligibility scores and the acceptability scores for the news scenario was $R = -0.7886$, and between the logistic fit to these scores and the acceptability scores for the news scenario the correlation is $R = -0.8141$ (i.e. around two thirds of the variance is explained).

These correlations are fairly high, however it may that some of the correlation between intelligibility and acceptability is explained by their mutual relationship with SNR. If this is so, it is likely that the correlation found here was exaggerated by averaging across repeats. The correlation between intelligibility and acceptability scores for the news scenario without averaging across repeats (i.e. 30 data, 2 for each of the 15 SNRs, averaged across the 7 subjects) was $R = -0.7617$.

If this relationship is present simply as a result of covariation with SNR, then both quantities would be expected to correlate as well, or better, with SNR. The correlation
Figure 5.29: Intelligibility scores (blue), and associated logistic fit (black), for the cases featuring the news interferer reflected in the line $\text{SNR} = 0 \text{ dB}$, and mean acceptability (red) and distraction (green) scores across all cases (solid lines) and only for the news interferer (dashed line).
between these 30 intelligibility scores and SNR was $R = -0.7455$, and the correlation between the acceptability scores and SNR was 0.8759. This indicates that some of the correlation between intelligibility and acceptability is likely to be due to covariance with SNR, however the presented data is not sufficient to determine the extent of this.

From the results of this experiment, therefore, there is some evidence indicating that when both target and interferer programmes are speech-based there is an inverse correlation between interferer intelligibility and acceptability, however the degree to which this is caused by their dependence on SNR is unclear. The acceptability and interferer intelligibility tend to vary similarly over the same approximate range of SNRs, and this may be useful both for marking a high acceptability score (i.e. where the interferer is unintelligible but not inaudible).

5.5.5 Overview of results

While the collected data show some different significant effects and trends, for all measurement variables the SNR had the largest effect. The data can therefore be plotted against SNR to show the general relationships between measurement variables, as in fig. 5.30. From this diagram a few interesting trends are revealed.

Firstly, at $-5$ dB SNR the intelligibility scores have reached their effective maximum very close to 1; for the same SNR, however, the distraction score is 0.57 and the acceptability score is 0.08. This strongly indicates that the intelligibility score alone is not sufficient to describe the perceptual quality of listening experience for scenarios where the target is speech. Instead it is more appropriate to consider optimal speech intelligibility to be a prerequisite component of a high quality listening environment.

Secondly, it is interesting to note that at $-10$ dB SNR the acceptability score is 0.02 while the distraction score is close to 0.68. At $+35$ dB SNR the acceptability score is at 0.79 and the distraction score is 0.06. This indicates that, while there is a correlation between the two measures, acceptability is not simply the inverse of distraction. It is reasonable to suggest that, like intelligibility, distraction may be an important and necessary component of acceptability, but not the exclusive factor. Since the distraction score was so well predicted by a logistic fit to SNR, however, that it is unclear whether the correlation between distraction and acceptability is anything more than an expression of covariation with SNR. Put another way, if listeners decided upon their ratings of distraction purely by estimating the SNR, then the correlation between distraction and acceptability would be expected to be approximately as high as the correlation between acceptability and SNR. It is also possible that the acceptability and distraction scores collected differed due to the collection methodology; if subjects were more conservative with their scoring of acceptability (collected in a binary format) than with their scoring of distraction (collected in a scalar format) then this result would
It was noted, however, in section 5.5.3 that the acceptability scores were split across subjects in a way which suggested that subjects belong to one of two groups: ‘high tolerance’ and ‘low tolerance’. One possible interpretation of this data is that the low tolerance group are primarily using audibility as a measure of acceptability. If the masking scores are inverted along the SNR axis the resulting curve is presumed to approximate the masking curve of the target upon the interferer. As fig. 5.31 shows, the low tolerance curve appears to be similar to the inverted masking curve but shifted by 5 dB. Thus it is reasonable to suggest that the low tolerance group interpreted ‘acceptable’ to mean ‘nearly inaudible’. The high tolerance group, it might also be suggested, are interpreting acceptability to mean the inverse of distraction, however the data from this experiment does not strongly support this suggestion. As fig. 5.31 also shows, the distraction curve when inverted along the score axis corresponds well with the high tolerance acceptability curve only between 15 and 25 dB SNR.
Figure 5.31: Scores for distraction, acceptability, and masking, speech intelligibility across SNR with 95% confidence intervals. Acceptability scores are separated into high tolerance and low tolerance subject groups, and masking and distraction curves have been reflected in the SNR axis and Score axis respectively to show how well they correspond to the two acceptability curves.
5.6 Summary and conclusions

The research question posed at the start of this chapter was “what relationships exist between intelligibility, acceptability, and other relevant measures?” Listening tests were carried out to obtain speech intelligibility data for a range of ecologically valid auditory interference scenarios. Additional measures of distraction, acceptability, and masking were also collected. The results indicated that SNR was universally the dominant factor in determining the value of these measures, however the interferer programme also had an effect which, in many cases, was dependent upon the presence or absence of speech in the interferer. Furthermore, the comparison of measures indicated that low acceptability and high distraction scores were possible even for relatively high intelligibility scores, and that acceptability was not simply the inverse of distraction. Acceptability scores, however, were close to zero for all cases where intelligibility was significantly less than 100%.

The results showed that when interferers were primarily comprised of speech, the psychometric functions for speech intelligibility, masking, distraction and acceptability were significantly different to those found when the interferers did not contain speech. The intelligibility psychometric functions for interferers featuring speech were shallower, with temporal sparsity in the interferer producing speech reception thresholds (SRTs) at lower signal to noise ratios (SNRs), and more variable at higher SNRs (likely due to the potential for word confusions). The distraction scores, for interferers featuring speech, were slightly lower for SNRs below −15 dB (since the gaps in the interfering speech allowed for greater word reception) and higher for positive SNRs (since there was greater scope for confusion). Acceptability and masking data revealed a similar split between interferer programmes featuring and not featuring speech, although much larger subject effects were found for acceptability than for distraction. For acceptability scores, the subject effects indicated that listeners might be grouped into distinct high and low tolerance groups, although with relatively few subjects this proposition remains to be confirmed.

In general, intelligibility was required to be near perfect before listening scenarios were rated to be acceptable. Intelligibility therefore marks a useful lower boundary of the acceptability space when the target programme is speech. In the next chapter, existing speech intelligibility models are evaluated for use in auditory interference scenarios, since if intelligibility can be predicted, the lower bound of SNR can be determined.
Chapter 6

Speech Intelligibility Prediction

In the previous chapter an experiment was conducted to find speech intelligibility scores for auditory interference scenarios featuring a range of ecologically valid programmes. Although it was not possible to predict the acceptability of individual listening scenarios based on intelligibility, it was shown that intelligibility described the range of acceptability for speech targets; specifically the acceptability was always zero until the target intelligibility was nearly 100%, and when the interferer intelligibility was less than 10%, the acceptability exceeded 80%.

Since it may be useful to describe such auditory interference scenarios with reference to intelligibility, a research question presents itself: “how can the intelligibility of speech within auditory interference scenarios be predicted?” This chapter addresses this question by evaluating speech intelligibility models using the data gathered in the experiment described in the previous chapter.

A wide range of speech intelligibility models have been developed to account for a variety of scenarios including room acoustics, digital speech enhancement and clinical audiology. In section 6.1 of this chapter four dissimilar, but appropriate, models are outlined, and in section 6.2 they are evaluated for accuracy predicting the intelligibility scores from the experiment described in the previous chapter.

6.1 Intelligibility models

A range of models for the prediction of speech intelligibility exist. Some are based on generalised models of audibility, while others are more specific; in general, however, models differ based on their intended application, which can vary from predicting intelligibility for hearing impaired listeners to measuring intelligibility impairments caused by compression techniques.

An exhaustive list of intelligibility models would be impracticable to evaluate, however, a range of models can be investigated based on differences in model design and based on previously reported performance accuracy.
6.1.1 The speech intelligibility index

Probably the first intelligibility model to be designed was the Articulation Index (AI) described in French and Steinberg (1947). The model was developed at Bell Laboratories and was designed to predict the intelligibility of speech passed through a telecommunications system. There have since been many refinements to the AI and in 1997 it was adopted by the American National Standards Institution (ANSI S3.5 1997) under a different name, the Speech Intelligibility Index (SII). The SII is perhaps the most widely used of the speech intelligibility prediction models, and is therefore a good place to begin searching for an appropriate speech intelligibility model.

Model design

Developed at the same lab, it is perhaps unsurprising that the SII bears resemblance to Fletcher's power spectrum model. The SII works by calculating masking effects for separate frequency bands to estimate which components of the target programme are audible, and then weighting those audible components by the assumed importance of those frequency bands before summing the result. The SII is therefore calculated by:

\[
SII = \sum_{i=1}^{n} I_i A_i,
\]

where \(I_i\) is the band importance function, \(A_i\) is the band audibility function, and where \(i\) indicates the frequency band under analysis. The band importance function \((I_i)\) is, for a given frequency band, a numerical value characterising the relative significance of this frequency band to speech intelligibility and is specified in (ANSI S3.5 1997). The band audibility function \((A_i)\) is, for a given frequency band, a numerical value between 0 and 1 specifying the effective proportion of the speech dynamic range within the band that contributes to speech intelligibility.

It is important to note that the SII, being contingent upon \(A_i\), therefore indicates the proportion of speech cues which are available to the listener and not the proportion of words or sentences which are expected to be understood. While these quantities are strongly related, it has been shown that for many conversational speech scenarios, in fact, an SII of 0.5 corresponds to nearly 100% sentence intelligibility (Fletcher and Steinberg 1952; Kryter 1962). This is largely due to the capacity of listeners to estimate sentences based on contextual information, and implies that mapping function of some kind would be required to translate the output of the SII into intelligibility predictions. (ANSI S3.5 1997) states that, “the transfer function should be developed by the user for the type of speech material whose intelligibility needs to be predicted”, but since for the current application the predictions should be as general as possible, a general transfer function (or range of transfer functions) is required.

ANSI S3.5 (1997) offers four methods of calculating the SII depending upon the
selection of frequency bands. In order of most to least accurate (according to the standard) these are: 21 critical bands, 18 one-third octave bands, 17 equally contributing critical bands, and 6 octave bands. Band importance functions are also specified as relating to various types of speech depending on how they were empirically derived (e.g. using nonsense syllables, or short passages.).

The band audibility function is calculated using:

\[ A_i = L_i K_i, \]  

where \( L_i \) is the speech level distortion factor, which is calculated based on the difference between the speech level in the current frequency band and a predetermined level describing ‘normal’ vocal effort. The term also includes an adjustment for conductive hearing loss. \( K_i \) is a temporary variable which is calculated using the speech level in the current frequency band and the ‘equivalent disturbance level’, which is calculated based on various masking effects (including within-band masking, spread of masking, and self-speech masking).

**Model accuracy**

The accuracy of the SI I is difficult to estimate, because it depends so heavily on the test material. According to Kryter (1962), “Test scores of “percent correct” are essentially meaningless unless the type of material, size-of-message set, and talker-listener training are known.” With the caveat that such factors sway the prediction accuracy, however, the SI I describes the part of the intelligibility which is due primarily to the audibility of the speech very well.

**Summary**

In summary, therefore, the SI I is a widely used measure of intelligibility calculated as a special type of frequency dependent SNR. It is calculated as the product of the audibility function \( (A_i) \), obtained by comparing the speech level within separate frequency bands with empirically obtained values, and an empirically determined band importance function \( (I_i) \).

The requirement for a speech-type dependent transfer function is a disadvantage, as is the highly empirically derived (rather than physiologically inspired) methodology, since these imply that the model may not generalise very well to different types of speech. It is worth noting, also, that the SI I was designed with the assumed application of additive noise or target speech distortion, not with additive meaningful linguistic content (if the interferer is speech).
Chapter 6: Speech Intelligibility Prediction

6.1.2 Short-time objective intelligibility

The Short-Time Objective Intelligibility (STOI) model of (Taal, Hendriks, Heusdens and Jensen 2011) was developed to predict intelligibility where noisy speech is processed by a time-frequency varying gain function. While this is not a description of the listening scenario under consideration, it is similar in that the arbitrary nature of the interferer programme implies that its masking (and partial loudness) effects will differ across time and frequency. For this reason it is worth investigating this model for predicting intelligibility within ecologically valid listening scenarios.

**Model design**

Figure 6.1 shows an overview of the operation of the STOI model.

STOI works by first decomposing the clean and mixed signals into time-frequency units with 50% overlapping, Hann-windowed frames with a length of 256 samples and with frames zero-padded up to 512 samples. Silent regions are identified as those frames in the clean speech with energy at least 40 dB below that of the frame with maximum energy. With the silences removed a total of 15 one-third octave bands are used with centre frequencies ranging from 150 Hz to 4.3 kHz. After a normalisation and clipping stage, the output of STOI is determined as the average of the intermediate intelligibility scores, which are themselves determined as the correlation coefficients between the temporal envelopes of frames for each one-third octave band. STOI finally outputs “a scalar value which is expected to have a monotonic relation with the average intelligibility of the listening scenario (e.g., the percentage of correctly understood words averaged across a group of users)”.

A useful characteristic of the STOI model is that it operates at near real-time speeds, and so could be utilised for a real-time implementation of a sound zone evaluation model.
Model accuracy

In (Taal, Hendriks, Heusdens and Jensen 2011), the predictions of the STOI model correlated with intelligibility scores with between $R = 0.92$ and $R = 0.96$ for a range of time-frequency degraded scenarios. It is far from clear, however, how well the model would apply to sound zone scenarios.

Summary

The STOI model makes very fast, highly accurate predictions for time-frequency degraded listening scenarios. It is unclear how closely such scenarios resemble auditory interference scenarios however.

6.1.3 Time frequency multi-look model

The Time/Frequency Multi-Look Model (T/FMLM) of (Hant and Alwan 2003) was produced in an attempt to account for both the detection and discrimination of speech stimuli in noisy backgrounds: a process which naturally lends itself to predicting intelligibility.

Model design

The general principle is to use an auditory preprocessor to produce an internal representation of the auditory stimuli before considering the ease with which information across adjacent time-frequency units can be combined to give information about the signal. An overview of the T/FMLM model is given in fig. 6.2.

The target and interferer programmes are first windowed into 6 ms windows with 1 ms cosine ramps at the start and finish producing a window with 5ms equivalent rectangular duration. In (Hant and Alwan 2003) each window was then preprocessed using the model of (Dau et al. 1996). The output of the model is taken to be the ‘detectability’. The detectability is calculated with:

$$d' = \sqrt{\frac{1}{N_t N_f} \sum_{i=1}^{N_t} \sum_{j=1}^{N_f} w_j(\mu s_{ij} - \mu m_{ij}))(d'_{ij})^2},$$

(6.3)

where $w_j(\mu s_{ij} - \mu m_{ij})$ acts as a gate based on a preset threshold, and $d'_{ij}$ represents the partial detectability. The partial detectability is calculated with:

$$d'_{ij} = \frac{\mu s_{ij} - \mu m_{ij}}{\sqrt{\frac{\sigma s_{ij}^2 + \sigma m_{ij}^2}{2}}},$$

(6.4)
and the gate is calculated with:

\[
w_j(|\mu s_{ij} - \mu m_{ij}|) = \begin{cases} 
1 & \text{if } |\mu s_{ij} - \mu m_{ij}| > \theta(j) \\
0 & \text{if } |\mu s_{ij} - \mu m_{ij}| < \theta(j)
\end{cases}
\]  

(6.5)

\(i\) and \(j\) represent time and frequency respectively, \(s\) is the intensity of the target plus interferer, \(m\) is the intensity of the interferer level, and \(\theta\) represents a threshold defined by:

\[
\theta(f) = 3.81 + 2.39 \left( -\frac{1}{2} + \frac{1}{2} \left( 1 - \exp \left( \frac{-16.15}{1.19} \right) \right) \right)
\]

(6.6)

In this way, the difference between the level of the target and interferer mixture, and the interferer alone is compared with a frequency dependent threshold \(\theta(f)\). The term \(w_j(|\mu s_{ij} - \mu m_{ij}|)\) forces the sum to zero where this difference is less than \(\theta(f)\). Where the difference is greater than \(\theta\), it is scaled according to \(d'_{ij}\) which is based on the magnitude of the difference and on the variation in intensity across time-frequency units.

Broadly, this approach is similar to the CASP model, in that it is an analysis based on the comparison between the mixed target and interferer with the interferer alone. It differs from CASP in that it is based on the numerical difference between these across time and frequency, rather than their correlation.
Chapter 6: Speech Intelligibility Prediction

Model accuracy

The T/FMLM was evaluated on a wide range of detection and discrimination tasks for phonemes in (Hant and Alwan 2003). The performance was generally very good, predictions were accurate to within a few dB in most scenarios, however the author is unaware of any evaluation of the T/FMLM model with connected speech.

Summary

Broadly, this approach is similar to the CASP model, in that it is an analysis based on the comparison between the mixed target and interferer with the interferer alone. It differs from CASP in that it is based on the numerical difference between these across time and frequency, rather than their correlation.

6.1.4 Coherence speech intelligibility index

The Coherence Speech Intelligibility Index (CSI I) of Kates and Arehart (2005) was developed specifically for application to the problem of nonlinear distortions introduced by hearing aids. As with the STOI model, the intended application is not directly comparable to the listening scenarios under consideration, but the model attempts to account non-linear degradations to intelligibility which may be similarly arbitrary to those found with auditory interference. Three criteria are specified for the CSI I prediction model:

1. The model must be applicable to systems featuring frequency dependent magnitude and phase responses,
2. In the absence of nonlinear distortions, but in the presence of additive noise, performance should be consistent with the SII,
3. The method should be applicable to a speech test signal.

Model design

The CSI I is calculated similarly to the critical band method of the SII, however the speech power spectrum \( \hat{P}(k) \) is replaced with:

\[
\hat{P}(k) = |\gamma(k)|^2 S_{yy}(k),
\]

where \( S_{yy}(k) \) represents the autospectral density of \( k \), and where \( |\gamma(k)|^2 \) represents the mean squared coherence which describes the fraction of a variable that is linearly dependent upon another variable. This calculation can be thought of as being analogous to using the correlation or the coefficient of determination to describe how two variables change together. In this case the coherence under test is that between ideal reference speech and the mixture.
The SII noise power spectrum is replaced by:
\[ \hat{N}(k) = [1 - |\gamma(k)|^2]S_{yy}(k), \]  
and instead of an SNR calculation, a signal-to-distortion-ratio (SDR) is calculated with:
\[ \text{SDR}(j) = \frac{\sum_{k=0}^{K} W_j(k) |\gamma(k)|^2 S_{yy}(k)}{\sum_{k=0}^{K} W_j(k) [1 - |\gamma(k)|^2] S_{yy}(k)}. \]

The CSI is therefore an extension to the SII which accounts for both additive noise and nonlinear distortions, by incorporating the coherence function.

**Model accuracy**

In an evaluation described in Kates and Arehart (2005), the CSI model was found to have a high accuracy with predictions having correlation \( R = 0.94 \) with intelligibility scores. The CSI model was also one of the top performing models tested in the evaluations described in Taal, Hendriks and Heusdens (2011) for time-frequency weighted distortions.

**Summary**

The CSI is an extension to the SII which incorporates the coherence function in order to account for nonlinear distortions.

### 6.2 Model evaluations

Before evaluating models to see which produces the most accurate results, it is beneficial to use a simple benchmark model against which the results can be more meaningfully interpreted. Extremely simple benchmark models can be constructed by predicting the same speech intelligibility score for every trial, based on which ever outcome is most common. In the case of the intelligibility data gathered from the experiment conducted in chapter 5, a prediction of 1 for every trial produces an RMSE of 47.21\%. Every model, therefore, should have an error lower than this if it is to be considered to be performing meaningful predictions.

#### 6.2.1 SII predictions

For the reported predictions the band importance function for 21 critical bands and short passages was selected for use. SII predictions were produced for each of the 270 trials of the intelligibility experiment described in section 5.4. The accuracy of these predictions was poor with RMSE = 66.85\% and \( R = 0.5159 \). As previously
discussed, however, the SII would not be expected to predict intelligibility directly, but would likely require a mapping function. The predictions were therefore mapped onto a logistic function of the form described in section 5.5.1 but with the SII prediction replacing the SNR term. The $\alpha$ and $\beta$ parameters were tested in the ranges $-10 < \alpha < 10$ and $0 < \beta < 10$ in increments of 0.1, and the search was repeated with ranges sequentially focused on the minimum RMSE with reducing increments until the optimum RMSE was found to four significant figures. The starting range of values for $\alpha$ were selected because the SII term can only vary between 0 and 1, and the range of values tested for $\beta$ were selected because negative $\beta$ values offer no gradients which could not already be accounted for with positive $\beta$ values and because $\beta = 10$ produces a gradient far shallower than the subjective data. The values producing the lowest RMSE were $\alpha = 0.000036$ and $\beta = 0.000040$, and had accuracy of RMSE = 22.25% and $R^2 = 0.8290$. This does not improve upon the accuracy of the logistic function fitted directly to the SNR in section 5.5.1 (RMSE = 22.3% compared with RMSE = 12.4%).

When this process was repeated using the 18 third-octave band method instead of the 21 critical band method, however, the accuracy of the unfitted SII was improved with RMSE = 49.74% and $R = 0.6602$ (see fig. 6.3). Fitting the data to a logistic function in the same way resulted in $\alpha = 0.0042$ and $\beta = 0.0022$ with an accuracy of RMSE = 13.66% and $R = 0.9386$. This accuracy, while still not improving upon that of the logistic function fitted to SNR, is within a few percentage points (RMSE = 13.7% compared with RMSE = 12.4%), and has similar correlation ($R^2 = 0.8809$ compared with $R^2 = 0.9023$).

**Simple adaptation to the SII**

The SII is based on the long-term spectrum of the target and interferer programme and, as a result, does not consider temporal variations. A simple solution to this problem...
is to break up the programmes into temporal windows and take the mean of the SII predictions. One problem with this method is that the silent periods between words and phonemes in the target programme will always have a SII of 0, thus the mean SII will always be less than unity even in ideal conditions. In order to account for this, temporal windows in which the target programme is silent are removed prior to SII prediction. This method, as well as the code used to achieve this, was adapted from (Donohue 2013).

Predictions were made for each of the 270 trials using the third-octave band SII model with window lengths of 100ms, 200ms, 400ms, 600ms, and 800ms. Windows were considered to be silent if the values of the envelope exceeds half the median value in the whole target programme sample (This parameter can be adjusted as necessary based on the characteristics of the background noise of the recorded target programme). The resultant SII predictions produced R = 0.6855, 0.7004, 0.7138, 0.7212 respectively. When these data were fitted to logistic functions the resulting predictions had R= 0.9720, 0.9778, 0.9762, 0.9708, and 0.9679 with RMSE = 0.0942, 0.0843, 0.0873, 0.0961, and 0.1003 (based on $\alpha = 0.0370, 0.0364, 0.0347, 0.0344, 0.0344$ and $\beta = 0.0120, 0.0122, 0.0135, 0.0145, 0.0147$).

Since the intelligibility for higher SNRs is consistently 1, these measures of accuracy should also be recalculated for the range between $-35$ and $+5$ dB SNR. When considering only those data points the scores were RMSE = 12.11, 10.81, 11.21, 12.35, and 12.91% with R = 0.9624, 0.9704, 0.9681, 0.9607, and 0.9568.

Notably, some of the patterns found in the subjective data were predicted here. Firstly, the speech-based interferer trials were predicted with the highest intelligibility scores (especially for the lower SNRs), as was found in the subjective data. Secondly the Speech Shaped Noise (SSN) interferer was predicted to have a very steep gradient which was also reported within the subjective data. In contrast to the subjective data, however, the predictions for the Mozart interferer appear not to bear the correct relation

![Figure 6.4: SII scores mapped onto a logistic function with silences removed and windowing at 200 ms (left) and subjective intelligibility data (right)]
Although the prediction accuracy is very high (around 10% error), a systematic error is visible in that the minimum prediction possible was 0.048. This occurs because when the SII prediction is 0 the logistic translation is equal to:

\[ P = \frac{1}{1 + e^{(0.0364 - 0.0122)}} = \frac{1}{1 + e^{2.98}} = \frac{1}{1 + 19.69} = 0.048. \]  

(6.10)

That it is appropriate to use such derived logistic fits is supported by the similarity between these curves and that reported in (Fletcher and Steinberg 1952), as shown in fig. 6.5, which relates discrete sentence intelligibility to the articulation index.

In general, SII predictions ought to be better correlated with the subjective intelligibility scores than the logistic fit of SNRs otherwise the SII model cannot be said to be actively predicting more effectively that fitting the SNR to the scale. While this was not achieved using the SII in its original form, when time windowing the signals, removing silences, and taking the mean SII the accuracy was greatly improved. The correlation of the time windowing SII model was \( R^2 \geq 0.9 \) for all window sizes tested between 100 and 800 ms (compared with \( R^2 = 0.8726 \) for a logistic fit to SNR).

### 6.2.2 STOI

From this description of STOI it was expected that some type of calibration would be needed, however to begin with the scores were obtained all 270 trials making no adjustments to this procedure. The resulting predictions are shown alongside the subjective intelligibility scores in fig. 6.6, and the predictions were found to have RMSE
6.2.3 TFMLM

In (Hant and Alwan 2003) the T/FMLM model is specified using the auditory preprocessor of (Dau et al. 1996). In this work the more recent CASP model had already been implemented in section 4.2 for the use of predicting masking thresholds, however, and this therefore represented an opportunity for neatly integrating masking threshold and speech intelligibility predictions into a single model. Predictions of this modified T/FMLM would not be identical to those using the Dau et al. (1996) auditory preprocessor, yet they should be very similar. The DRNL filter, the inner hair cell envelope extraction, and the expansion stage of the CASP model were therefore

\[ = 24.35\% \text{ with } R = 0.8048. \]

A logistic fit of the 270 STOI predictions to the intelligibility data is shown in fig. 6.7. the logistic fit has \( \alpha = 0.3895 \) and \( \beta = 0.0265 \), and represents a very significant improvement in prediction accuracy with RMSE = 14.13\% with \( R = 0.9338 \). For data points within the range -35 to +5 dB SNR, the accuracy was RMSE= 18.23\%, and \( R = 0.9089 \).

Figure 6.6: Uncalibrated STOI predictions (left), and subjective intelligibility scores (right).

Figure 6.7: Logistic fit to STOI predictions (left), and subjective intelligibility scores (right).
used before applying logarithmic compression and adding internal noise. Finally the mean and standard deviations of the mixture and interferer programmes were used to calculate the ‘detectabilities’.

Running the T/FMLM model in this way produce 270 speech intelligibility predictions which approximated the subjective data with RMSE = 59.86 and R = 0.5460. The resulting predictions are shown alongside the subjective intelligibility scores in fig. 6.8. These output values are, however, simply $d'$ values intended to relate monotonically to subjective scores and must therefore be fit to a logistic function. The optimal logistic fit (minimising the RMSE) has $\alpha = 9.6742\beta = 0.6218$ with RMSE = 15.45% and $R = 0.9201$. This is a very significant improvement, and as fig. 6.9 shows, the predictions match the trends found within the subjective data fairly well.

6.2.4 CSII

When the CSII is calculated for the 270 trials the accuracy of the output is poor with RMSE = 39.20% and $R = 0.6895$. Notably, however, this accuracy is superior to the
standard output of the SII (R = 0.6685). Figure 6.10 shows these predictions, and it is notable that trials featuring the traffic interferer at −30 dB SNR were greatly misidentified.

A logarithmic function fit to the CSI improves scores somewhat with RMSE = 0.1501 and R = 0.9272 (for $\alpha = 0.0369, \beta = 0.0025$). This is less accurate than the optimal fit gained by the SII when using 200 ms time windowing (10.8% error), and less accurate than the SII without time windowing (13.7% error), which implies that the time windowing and silence removal implemented for the SII predictions was not the primary factor involved in the difference in accuracy between these models.

### 6.3 Model comparison

Table 6.1 shows a comparison of the intelligibility models tested. While all models (with an appropriate logistic fit) are accurate to within approximately 15% RMSE, the SII was the only model which had lower error than a logistic fit to the SNR.
Chapter 6: Speech Intelligibility Prediction

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic fit to SNR</td>
<td>12.4%</td>
<td>0.9023</td>
</tr>
<tr>
<td>SII</td>
<td>10.8%</td>
<td>0.9704</td>
</tr>
<tr>
<td>STOI</td>
<td>14.1%</td>
<td>0.9338</td>
</tr>
<tr>
<td>TFMLM</td>
<td>15.5%</td>
<td>0.92</td>
</tr>
<tr>
<td>CSII</td>
<td>15.0%</td>
<td>0.9272</td>
</tr>
</tbody>
</table>

Table 6.1: Prediction error for intelligibility models

Although performing with less accuracy, the T/FMLM predictions identified the approximate correct order of interferer programmes with respect to intelligibility, whereas the STOI model produced more accurate predictions overall yet misidentified the order of interferer programmes. It is noteworthy that the T/FMLM model correctly identified these patterns, since it is partially based upon the CASP model and could be integrated into the masking model in a more cohesive way than STOI or the SII.

6.4 Summary and conclusion

At the start of this chapter the following research question was posed: “how can the intelligibility of speech within auditory interference scenarios be predicted?” In order to address this question, a range of speech intelligibility models were implemented and utilised to predict the data obtained in the speech intelligibility listening test described in the previous chapter and the sets of predictions were compared with each other and with a logistic fit to SNR.

All of the models tested were able (with a suitable logistic fit) to predict mean intelligibility scores to within around 15% RMSE, but only the SII produced predictions with accuracy exceeding that of a logistic fit to SNR (10.8% error compared with 12.4%). The time-windowed SII is therefore recommended as a method for describing intelligibility for auditory interference scenarios featuring speech. These intelligibility predictions could be used to determine the range of SNRs over which acceptability would vary, and in doing so estimating the low and high boundaries. Since it is not possible to use these intelligibility predictions to directly predict acceptability, the next chapter focuses on how to predict acceptability directly when the target programme is speech.
Chapter 7
Acceptability Training and Validation Experiments

It was shown in chapter 3 that masking thresholds can be used to predict acceptability scores for interfering audio programmes, and in chapters 5 and 6 intelligibility was investigated to narrow down the range of SNRs within which acceptability scores vary when the target programme includes speech. Chapter 5 also indicated that listener acceptability scores might be divided into two definite groups, rather than being unimodal. Two questions therefore remain. First, “how can the acceptability of auditory interference scenarios featuring a speech target be determined?” And second, the related question “what is the general distribution of listener acceptability responses?”

In order to answer the first question, it was deemed necessary to obtain a large set of acceptability data with which to train and validate a model predicting acceptability for audio interference with a speech target programme. In order to consider the second question, the experiment should include sufficiently many subjects to investigate the possible bimodality of acceptability scores indicated by the results in chapter 5.

This chapter describes two experiments conducted to gather the acceptability data for a wider range of subjects and ecologically valid stimuli than the previous experiments. Section 7.1 describes an experiment to gather data for training the acceptability model whilst addressing the differences between subjects. Section 7.2 describes a smaller experiment designed to gather acceptability scores for stimuli processed using a sound zoning system, with the goal that this data set could be used to validate the applicability of the acceptability model to new listening scenarios (rather than just new stimuli).

7.1 Training experiment

The primary motivation for the training experiment was to gather data which can be used to construct the acceptability prediction model. The acceptability prediction model aims to predict the proportion of acceptable scores produced by listening within auditory interference scenarios featuring a speech target programme.

The findings of the speech intelligibility experiment suggested that listeners might belong to one of two groups: a ‘high tolerance’ or ‘low tolerance’ group (see section 3.2.2). It is also possible, however, that this grouping was an artefact caused by subjects reporting acceptability scores whilst also transcribing speech; for example
Chapter 7: Acceptability Training and Validation Experiments

if listeners used the intelligibility of the target speech as a cue to deciding upon their ratings of acceptability. If, however, the result was not an artefact of the test methodology and listeners in the general population really are divided into two groups it would be appropriate to construct two distinct acceptability models which report the acceptability for each listener type.

The training experiment was therefore designed to meet two goals:

1. a variety of ecologically valid data should be gathered with which to construct an acceptability model, and

2. sufficient subjects should be tested to confirm or refute that listener judgements are grouped as either ‘high’ or ‘low’ tolerance.

7.1.1 Methodology

Subjects were seated near the centre of a ITU-R BS.1116 (1997) conformant listening room with a Genelec 1032 loudspeaker positioned two metres away at 0 degrees azimuth, and 1.2 m above the floor (approximately head height for seated subjects). The loudspeaker was calibrated such that pink noise of equivalent Root Mean Square (RMS) level to the stimuli was measured at 65 dB SPL at the listening position. A laptop computer provided a simple two button interface for subjects with one button (a green tick) representing acceptable and another (a red cross) representing unacceptable. Judgements of acceptability were therefore binary, resulting in a task which subjects reported to be very simple (although not all trials were considered easy). The user interface is shown in fig. 7.1

For each trial of the test the listening scenario automatically replayed the combined target and interferer stimuli. A countdown was displayed, along with a status message, during the first three seconds of each trial indicating that the interferer programme was currently muted; this helped subjects to identify the target programme distinctly within the mixture. During playback the user interface input was disabled, forcing the subjects to audition the entire 10 seconds before making a judgement. After the combined stimuli were replayed the subject was allowed as long as necessary to make a judgement about acceptability and report it via the appropriate button; subjects commented that they usually completed this step immediately since for most trials the judgement was made during audition. After reporting the judgement, the next trial began automatically, starting with another 3 second countdown.

A familiarisation stage comprised of 20 (unique) trials was required to ensure that subjects properly understood the task and to give them practice with the range of stimuli in use. In the subsequent main experiment 200 stimuli combinations were utilised, with a further 20 anchors giving 220 trials in total. This decision was arrived
Hello and welcome to this speech acceptability listening test. This test is designed to investigate listening scenarios featuring interfering audio programmes. In each case you should listen to the audio sample and imagine that you are relaxing by listening to the radio at home or in the car.

At the beginning of each trial there will be 3 seconds where only the 'target' radio programme will be audible; after this time, the interfering programme will be introduced and both programmes will replay for a further 7 seconds (10 seconds in total). During each trial you should listen to the audio sample and consider whether you find the listening scenario to be acceptable or not. After the audio sample has finished playing you may report your judgement by clicking on either the tick (if you found it acceptable) or the cross (if you found it unacceptable).

When you are ready to begin the first trial go ahead and click the button on the left marked 'Begin'. Each subsequent trial will begin automatically after you have made your judgement. Please note that the buttons will be disabled while the audio sample is playing.

There are 220 trials altogether and we ask that you take a short break after every 55 trials (at trials 55, 110, and 165).

Thanks for your participation!

Figure 7.1: The user interface used by subjects to provide binary acceptability data.
Table 7.1: The probability of appropriately scoring 8 or more anchors when picking at random.

<table>
<thead>
<tr>
<th>Number of correct trials</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.04395</td>
</tr>
<tr>
<td>9</td>
<td>0.00977</td>
</tr>
<tr>
<td>10</td>
<td>0.00098</td>
</tr>
<tr>
<td>8 or more</td>
<td>0.0547</td>
</tr>
</tbody>
</table>

at as a compromise between the opposing goals of including as many stimuli as possible (in order to represent a wide range of ecologically valid stimuli) and including sufficiently few stimuli that the test could be completed relatively quickly (in order to get as many subjects as possible). The full experiment therefore required a minimum of 40 minutes per subject (240 trials \(\times\) 10 seconds), and was usually completed in less than an hour.

The twenty anchor trials were comprised of ten high anchors (featuring no interferer programme) and ten low anchors (mixture set to \(-10\) dB SNR). Although a lower SNR could have been used for the low anchor, it was considered possible (even with the three second introduction) that subjects might misidentify the target and interferer programme if the target programme was too quiet. It was decided that the acceptability scores provided by subjects who failed to identify at least eight of the anchors correctly in each group of ten should be excluded from further analysis; the exclusion criterion was set at eight to keep the probability of scoring the anchors appropriately at around 5%. This is calculated by summing the probabilities of correctly rating eight, nine, or ten anchors, (see table 7.1) with each probability calculated as follows:

\[
P(k \text{ correct}) = \binom{n}{k} \times 0.5^n \times (1 - 0.5)^{n-k}
\]

(7.1)

where \(n\) indicates the number anchors (for each anchor type), in this case 10, and where \(k\) indicates the number of anchors marked correctly. Since there were both high and low anchors, however, the probability that a subject’s data would fail to be excluded even if guessing at random is equal to the product of the probability that eight or more trials were guessed correctly for the high anchor and the probability that eight or more trials were guessed correctly for the low anchor. Mathematically this can be expressed as:

\[
P(\text{not excluded}) = P(8 \text{ or more})_{High} \cap P(8 \text{ or more})_{Low} = 0.0547^2 = 0.003
\]

(7.2)

With the exclusion criterion of eight trials per anchor, therefore, the probability of failing to exclude a subject who guesses at random is 0.3%.

Subjects did not have the opportunity to listen to each stimulus more than once. It was considered preferable to obtain the initial reaction of the listeners to the stimuli than
to solicit considered judgements based on repeated listening since the former is more representative of ecologically valid listening scenarios. Disallowing repeated listening is also consistent with the methodology for gathering acceptability results utilised previously in the speech intelligibility experiment discussed in chapter 5, wherein subjects would have been able to improve their score on the intelligibility aspect of the work had repeats been allowed.

No trials were repeated since this would have reduced the number of unique stimulus combinations possible in a fixed experiment duration and it was considered more important to gather as much unique training data as possible than to examine listener consistency. It is also possible that repeated trials might have altered subject response behaviour (e.g. repeated identical stimuli could cause the listener to become more bored or annoyed with the test which may affect their judgements). Trial order was also randomised across subjects to minimise any potential presentation order effect.

Since subjects did not have the opportunity to listen to each pair of stimuli more than once it was important that they pay attention during each trial. To facilitate this subjects were required to take short (approximately two minute) breaks after each 55 trials, dividing the experiment into four ten minute sessions.

7.1.2 Stimuli

In order to select appropriate stimuli it was of primary importance that the stimuli should be as ecologically valid as possible. Of secondary, but also considerable, importance was the stipulation that stimuli should represent a broad range of possible stimulus types within the set of all possible ecologically valid stimuli. Finally it was also desirable to minimise any potential stimulus selection bias.

Ecologically valid stimuli likely to be found in living rooms and cars would be those such as commercially available music programmes, radio programmes, television and film programmes, as well as telecommunications. Since radio programmes play popular music, read out the news, advertise products (often involving a wide range of sound effects), and conduct telephone interviews, they include excerpts which cover a broad range of stimuli likely to be similar to all ecologically valid stimuli of interest. While it may be argued that excerpts from television programmes, films, or telecommunications might be somehow characteristically different from the radio programmes obtained in this way, any such differences are likely to be small since the same broad categories of programme types are involved (i.e. music, speech, sound effects). Additionally, the data gathered in the experiment discussed in chapter 5 suggested no substantial differences between acceptability scores for those stimuli extracted from films and those extracted from radio shows; instead, the differences between acceptability scores across interferer programmes was the partly explained by the presence (or absence) of linguistic content.
Chapter 7: Acceptability Training and Validation Experiments

within the interfering programme.

In order to obtain stimuli meeting the above stated stipulations radio excerpts were obtained, at randomly generated times, from some of the most popular radio shows in the United Kingdom. This method resulted in a range of music, speech, and mixed stimuli with a variety of different characteristics (e.g. genres of music, talker accents, etc.). For the target programmes, ten second excerpts were recorded from Talksport and from BBC Radios 1, 2, 4, 5, and BBC World. These were selected because they were the most popular radio stations, according to (Radio Joint Audience Research 2013) listening figures, featuring speech dominated programme segments with websites allowing streaming of up to a week of prior broadcasts. Interferer programmes were gathered from TalkSport, Heart London, Kiss 100 FM, Capital London, Classic FM, and BBC Radios 1, 2, 3, 4, and 5. These were selected for the same reason as the target programmes with the exception that radio stations which are not primarily speech were included. Where target stimuli recorded at the randomly generated times were not primarily speech-based (which occurred mostly for BBC Radios 1 & 2) new random times were generated for that day until primarily speech-based stimuli were present.

The random times were generated (eight per day per radio station over a five day period for the target, four per day per station over a 6 day period for the interferers) over an extended period of time including both weekdays and weekends to get a variety of programme types. Ten second excerpts were recorded at each time from the respective Internet streaming service.

Stimuli were combined at randomly selected SNRs ranging from $0 - 45$ dB. Selecting a range of SNRs too narrow would not produce scores which cover the full range of variance of acceptability, while selecting a range wider than necessary would limit the resolution of data gathered (for a fixed number of trials). The selected limits were based on the results from the speech intelligibility experiment, presented in section 3.2.2, which showed close to 0% average acceptability at 0 dB SNR, and around 80% average acceptability at 35 dB.

7.1.3 Subjects

21 subjects were recruited to take part, 9 female & 12 male, ranging from 22 to 65 years in age. Subjects included both naïve and expert listeners (in terms of musical ability, technical listening skills, and familiarity with listening tests). With 21 binary ratings of acceptability per trial the mean acceptability scores had a resolution of 4.5%.
7.1.4 General results and analysis

Subject comments

Subjects reported that the task was simple to understand and execute although it was, on some trials, not easy to make a judgment. A few subjects reported that the presence of speech interference with target speech was less likely to be acceptable than interfering music; which was reported in previous listening tests. Some subjects also reported an initial bias due to their familiarity with the voice of one of the radio presenters, although this effect would be present for very few trials. In any case such biases would be present in the listening scenarios under consideration and should be averaged out across many subjects.

Some subjects also speculated that their responses may be biased by the context in which they ordinarily listen to the radio; for example some subjects reported that they ordinarily listen to the radio while completing chores at home, others tended to listen to the radio while driving, while some rarely listen to the radio.

Anchors

All subjects rated at least 8 of the 10 high anchors as acceptable, and the mean score across subjects for the high anchors was 9.5. For the low anchors, all subjects again rated at least 8 of the 10 low anchors as unacceptable, and the mean score across subjects was 0.2. Since all subjects exceeded the minimum threshold for the anchors none of the subject’s scores were removed from further analysis, and the subjects appear to have been performing the correct task.

SNR range

The stimuli had SNRs ranging from 0 to 45 dB. These were selected to fully cover the range of acceptability. The total acceptable responses was 119.5 acceptable and 80.5 unacceptable, indicating that the SNR range selected was appropriately centered. For the 44 trials with a SNR in the top 10 dB (i.e. between 35 and 45 dB SNR) the 21 subjects reported that 885, of a possible 924, were acceptable (i.e. over 95% of cases). For the 43 trials with a SNR in the bottom 10 dB (i.e. between 0 and 10 dB SNR) the 21 subjects reported that 53, of a possible 903, were acceptable (i.e. fewer than 6% of trials). This indicates that the full range of SNRs were utilised. The range of SNRs selected was therefore appropriate.

Subject agreement

Only 49 of the 200 trials had a mean acceptability score between 0.25 and 0.75. Of the remaining 151 trials, 86 had a mean acceptability score above 0.75 and 65 had a mean acceptability score below 0.25. Thus for approximately three quarters of all trials, three
quarters of the subjects were in agreement about whether the listening scenario was acceptable. In general, therefore, there was good agreement between subjects. The disagreement largely occurred, as expected, in the middle range of SNRs tested, where subjects neither considered the interferer to be clearly unacceptably loud, nor clearly acceptably quiet.

One way to evaluate individual subjects for agreement is to consider, for each subject, the proportion of the 200 trials in which a subject’s acceptability judgement agreed with the majority opinion of the other subjects. Table 7.2 shows these quantities for all 21 subjects, alongside the number of trials each subject found acceptable. Subject 17 had the lowest proportion, with subjects 3 and 15 having slightly higher agreement with the majority.

Another way to investigate the deviation from the behaviour of the other listeners is to calculate the phi coefficient. The phi coefficient (sometimes known as Matthews correlation coefficient) describes the correlation between two sets of binary data, and is calculated with:

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1}\cdot n_{0}\cdot n_{1}\cdot n_{0}}}$$

where $n_{ij}$ represents the number of trials reported as either acceptable (1) or unacceptable (0) by the test subject (i) and the majority of other subjects (j), and where $n_{1\cdot}, n_{0\cdot}, n_{\cdot1},$ and $n_{\cdot0}$, represent $n_{11} + n_{10}, n_{01} + n_{00}, n_{11} + n_{01},$ and $n_{10} + n_{00}$ respectively.

**Distribution of subject scores**

One of the aims of this experiment was to verify or refute the previous indication that listeners may belong to distinct ‘high’ or ‘low’ tolerance groups. The histogram presented in fig. 7.2 suggests a unimodal, rather than bimodal, distribution. Subjects 3 and 17 reported a larger number of trials to be acceptable than other listeners; these were 166 and 187 respectively, compared with the mean 119.5. This is also reflected in their disagreement proportions (0.735 and 0.645 respectively) and their phi coefficients (0.5314 and 0.3828). It seems possible, from observing the histogram, that these two listeners could represent the high tolerance group. This explanation is weak, however, because subject 15 has a similar disagreement proportion (0.765) and phi coefficient (0.6103) to subject 3, yet subject 15 had the lowest number of acceptable trials (69). There is no evidence, therefore, refuting the claim that subjects 17, 3, and 15 simply occupy the extremes of a unimodal distribution of listeners.

**Subject effects**

Subject scores were separated into two groups according to whether the subject has received any musical or technical ear training. When grouped in this way the data for
### Table 7.2: The table reports, for each subject, the proportion of trials in which the subject agreed with the majority opinion of the other subjects, the number of trials considered acceptable by the subject, and the phi coefficient.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Proportion of consistent trials</th>
<th>Number of trials acceptable</th>
<th>phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.895</td>
<td>107</td>
<td>0.7936</td>
</tr>
<tr>
<td>2</td>
<td>0.845</td>
<td>109</td>
<td>0.6917</td>
</tr>
<tr>
<td>3</td>
<td>0.735</td>
<td>166</td>
<td>0.5314</td>
</tr>
<tr>
<td>4</td>
<td>0.890</td>
<td>127</td>
<td>0.7819</td>
</tr>
<tr>
<td>5</td>
<td>0.875</td>
<td>130</td>
<td>0.7530</td>
</tr>
<tr>
<td>6</td>
<td>0.965</td>
<td>119</td>
<td>0.9300</td>
</tr>
<tr>
<td>7</td>
<td>0.895</td>
<td>110</td>
<td>0.7919</td>
</tr>
<tr>
<td>8</td>
<td>0.860</td>
<td>130</td>
<td>0.7259</td>
</tr>
<tr>
<td>9</td>
<td>0.880</td>
<td>99</td>
<td>0.7743</td>
</tr>
<tr>
<td>10</td>
<td>0.820</td>
<td>143</td>
<td>0.6564</td>
</tr>
<tr>
<td>11</td>
<td>0.920</td>
<td>115</td>
<td>0.8396</td>
</tr>
<tr>
<td>12</td>
<td>0.865</td>
<td>90</td>
<td>0.7612</td>
</tr>
<tr>
<td>13</td>
<td>0.855</td>
<td>127</td>
<td>0.7128</td>
</tr>
<tr>
<td>14</td>
<td>0.900</td>
<td>108</td>
<td>0.8027</td>
</tr>
<tr>
<td>15</td>
<td>0.765</td>
<td>69</td>
<td>0.6103</td>
</tr>
<tr>
<td>16</td>
<td>0.855</td>
<td>145</td>
<td>0.7368</td>
</tr>
<tr>
<td>17</td>
<td>0.645</td>
<td>187</td>
<td>0.3828</td>
</tr>
<tr>
<td>18</td>
<td>0.880</td>
<td>96</td>
<td>0.7774</td>
</tr>
<tr>
<td>19</td>
<td>0.850</td>
<td>90</td>
<td>0.7328</td>
</tr>
<tr>
<td>20</td>
<td>0.890</td>
<td>125</td>
<td>0.7819</td>
</tr>
<tr>
<td>21</td>
<td>0.935</td>
<td>118</td>
<td>0.8693</td>
</tr>
</tbody>
</table>
the anchors showed little difference between the groups, see fig. 7.3. This implies that both types of subjects were performing the task correctly.

It is also interesting to see whether trained and untrained subjects differed in their general disposition towards interference; for example it is possible that trained listeners, in being capable of discerning slighter distortions, may be more easily annoyed by interference. Figure 7.4 presents histograms of the number of acceptable trials for trained and untrained listeners. No obvious clustering at either end is apparent; in fact the two subjects with the highest number of acceptable responses were in opposite groups.

In general, therefore, there were no strong subject effects.

**SNR and acceptability**

As expected, a strong correlation was found between SNR and acceptability. Figure 7.5 shows acceptability for the 200 trials plotted against the SNR of each trial. The correlation coefficient between SNR and mean acceptability score was $R^2 = 0.83$.

The trial with mean acceptability close to 0.6 and SNR close to 45 dB SNR seems to be an outlier caused by subjects making a target-interferer misidentification. In the final 1.5 seconds of the stimulus the target programme changes from male speech to music, and it is likely that subjects misattributed this music to the interferer.
Chapter 7: Acceptability Training and Validation Experiments

Figure 7.3: The proportion of responses marked as acceptable for subjects with (left/blue) and without (right/red) musical or technical ear training.

Figure 7.4: The proportion of responses marked as acceptable for trained and untrained listeners.
Chapter 7: Acceptability Training and Validation Experiments

Figure 7.5: Mean acceptability scores reported by the 21 subjects plotted against SNR for each of the 200 trials. The black dash-dotted line is presented only to aid reading of the plot, whereas the red dash-dotted line represents a linear regression to the data.

7.2 Validation experiment

Another experiment was carried out to gather a smaller quantity of data to be used for the validation of the acceptability model. This experiment was conducted to gather data which differed from that produced by the acceptability experiment described in section 7.1 due to both the stimuli and listening environment. While a cross-validation could be used to investigate the robustness of an acceptability model to new stimuli, this validation data set would provide an opportunity to investigate the robustness of acceptability models to use with a sound zoning method.

7.2.1 Differences in experiment design

The experiment methodology was very similar to that used in section 7.1.1, and so the differences are outlined here. Subjects were seated in a small listening room with a computer and a pair of Sennheiser 600 HD headphones. The monitor of the computer displayed a single page featuring all 24 trials. Subjects were given as much time as they required to listen to each trial and decided whether to mark it with a tick (for acceptable) or a cross (for unacceptable). The user interface is shown in fig. 7.6.

The buttons were not disabled after one click so repeat listening was possible. All button clicks were recorded, however, and the records indicated that subjects rarely listened to a trial more than once.
Hello and welcome to the speech acceptability listening test, demonstrating binaural recordings of radio programmes replayed through a sound zone system.

Before you begin, please click the reset button on the left. Now, please listen to each sample (by clicking on the play icon) and imagine you are relaxing by listening to the radio at home or in the car. For each sample consider whether you find the listening scenario to be acceptable. Click on the cross to toggle between acceptable (a tick) and unacceptable (a cross).

Please note that each sample begins with 3 seconds of the ‘target’ programme alone before the interfering programme is introduced.

When you are finished, press the button in the lower right to submit your results and see how they compared to our preliminary panel! Don’t forget to sign a consent form if you want to allow us to analyse your data for further research.

Thanks for your participation!

Figure 7.6: The user interface used by subjects to provide binary acceptability data.
14 subjects participated in the listening test which was ordinarily completed within 5 minutes.

7.2.2 Stimuli

The stimuli were radio programmes collected using the methodology described in section 7.1.2. Once randomly paired and level normalised, however, they were not mixed as before; instead the programmes were replayed through a 60 loudspeaker circular array after being processed using an Acoustic Contrast Control (ACC) sound zoning method. This sound zoning method aims to maximise the difference in the average energy across two physical zones by delaying and filtering the stimuli.

The circular array was positioned at a height of 1.62 metres in an acoustically treated room with dimensions 6.55 by 8.78 by 4.02 metres. The circular array had radius 1.68 metres with circular target and interferer zones with radii 0.18 metres and centered 0.7 metres either side of the centre of the loudspeaker array. This sound zoning system was calibrated using 192 microphones per zone to produce a replay level of 76 dB SPL of white noise in the target zone. The contrast achieved by the ACC method varied across frequency, but was approximately 20 dB at 100 Hz, rising to 30 dB at 1.5 kHz, and falling thereafter to around 10 dB at 7 kHz. A head and torso simulator was positioned in the centre of the target zone such that its ears were at the same height as the loudspeakers. The binaural head dummy was used to make recordings of the stimuli pairs reproduced by the sound zoning system.

As a result of this stimuli processing method, the SNR of each trial was determined by a combination of the variable contrast over frequency of the ACC method implemented, and the spectral content of the stimuli processed. Because of this, the SNR across trials was neither constant, nor randomised (as in the training experiment).

The trials also, as a result of the sound zoning method, were perceived by listeners to have a degree of spatial separation. This is a distinct difference from the trials in the training experiment (which were monophonic). The ACC method effectively selects the direction of arrival of the target programme to optimise contrast across zones and, in this case, the target frequencies were reproduced primarily from 160 and 200 degrees (i.e. approximately 20 degrees in front and behind of the right axis) with lower frequencies spreading further around to the front and back of the listener. The interferer frequencies appear to be arbitrarily distributed with little cohesion (i.e. apparently coming from everywhere). As a result, the target programme was perceived to be coming from the right hand side, whilst the interferer programme was perceived to be coming from the left and front (probably because the right side components were largely masked by the target).

Another artefact which can occur when using ACC is pre-ringing. This occurs because
long impulses responses result in complex filters produced in the frequency domain. When these filters are implemented, a pre-ring occurs in the stimuli which produces the percept of temporally smeared programmes. This can seriously degrade programme quality, so in order to minimise this pre-ring the microphone calibration impulse responses were truncated to 30 ms. This effect was described in Cai et al. (2013).

7.2.3 General results and analysis

The 336 acceptability ratings were averaged into 24 mean acceptability scores ranging from 0.0667 to 0.8667. For each of the 24 trials the left and right ears were summed and the SNR of the resultant mono recording was calculated. Figure 7.7 shows the mean acceptability scores plotted against SNR.

The correlation between SNR and the mean acceptability scores were much poorer, $R = 0.1268$ than in previous experiments. The data point in the top left of fig. 7.7 is apparently an outlier, however, since the target programme featured a radio presenter introducing a musical artist and the interfering music was likely misattributed to the target programme as the artist in question. When this outliar is removed the correlation is greatly improved with $R = 0.3485$. 

![Figure 7.7: Mean acceptability scores plotted against signal to noise ratio.](image)
7.3 Summary and conclusion

At the start of this chapter two questions were posed: “how can the acceptability of auditory interference scenarios featuring a speech target be determined?”, and “what is the general distribution of listener acceptability responses?” In order to address both of these questions it was considered necessary to obtain a large quantity of acceptability data for a wide variety of ecologically valid stimuli and many subjects. Two experiments were conducted: the first designed to gather more acceptability data about a wider range of stimuli for the purpose of training a model to predict the acceptability of speech target programmes with audio interference, and another, smaller, experiment designed to gather a validation set which can be used to estimate the extent to which the acceptability model is robust to new listening environments and sound zoning methods.

The second question was addressed by noting that for the 21 subjects taking part in the first experiment, there was no evidence to suggest that acceptability scores were not unimodal. The first question, however, cannot be answered only by collecting the subjective data but also requires an in-depth model training stage. The next chapter answers this question by describing the application of the data sets gathered in these experiments to the construction, testing and validation of a model of the prediction of the acceptability of speech with auditory interference.
In chapter 4 a method was shown for the prediction of the acceptability of auditory interference scenarios using predictions of masking thresholds, but it was also noted that the method may not be reliable when the target was speech. In chapters 5 and 6 the intelligibility of speech was investigated for use predicting acceptability thresholds and while intelligibility can be used to describe the range of acceptability, it was not able to assist the prediction of the acceptability of specific cases.

In this chapter, the primary research question is therefore “How can the acceptability of auditory interference scenarios featuring a speech target be predicted?” This chapter describes the construction of a model to answer this question. Section 8.1 introduces model construction in general, and outlines the data sets available for use and the metrics by which models can be compared. Next, in section 8.2 a first set of models is constructed, of which one is selected, by using features based on the internal representations of the CASP model. A benchmark model is also constructed, and the prediction accuracy and generalisability of the models are compared. Subsequently, in section 8.3 the process is repeated after generating further features based on stimuli levels and spectra, and manually coded features based on subject comments. In section 8.4 the process is repeated once more, further including features derived from the Perceptual Evaluation methods for Audio Source Separation (PEASS) model. The various selected models are compared in section 8.6, and the findings are summarised and conclusions drawn in section 8.8.

8.1 Modelling approach

Before constructing and evaluating various models of acceptability, it is important to introduce some general principles regarding the construction of models. It is also necessary to identify the available data for training and testing the models, as well as outlining the metrics by which models will be evaluated and compared.

8.1.1 Model complexity

A range of possible acceptability models can be constructed, from a simple linear regression using one feature (such as SNR) to complex, hierarchical, multi-dimensional
models. Some models will be more accurate, but at the cost of robustness to new listening scenarios or stimuli. In general when building models of prediction it is useful to include as many features as possible as long as this does not diminish robustness. This is because complex attributes, such as whether a listening scenario will be perceived as acceptable, depend upon a wide array of disparate contributing factors. Using too few features will result in a model with inaccuracies which fail to account for significant effects acting upon the attribute (in this case, acceptability). Conversely, a model including too many features may have an increased accuracy for the data upon which the model is trained, but fail to replicate this improved accuracy when tested upon new data. This latter error, known as 'overfitting', occurs because a regression simply fits the feature coefficients to the data in the optimal manner so a greater number of features will tend to improve prediction accuracy even if some features do not genuinely describe the prediction attribute. Overfitting can therefore be detected by comparing the accuracy of the model at predicting the training set with the accuracy of the model at predicting the test set. A reasonable compromise, therefore, needs to be achieved between the selection of sufficient features to accurately model the attribute and the selection of sufficiently few features to maintain the robustness of the model to a new data set (and to new test scenarios if desired).

For the prediction of acceptability, a very simple model using only the SNR of the listening scenario as a feature can be constructed. A model based on SNR is a sensible starting point because, as discussed in chapter 1, the acceptability of auditory interference scenarios is clearly bounded by the audibility of the target and interferer programmes. It has already been shown in the analysis of previous experiments that such models are capable of predicting acceptability scores with reasonable accuracy, and the robustness both to new data and to new listening scenarios would be expected to be high due to the simplicity of the model. Conversely, however, such a model would be unlikely to represent the optimal accuracy of all models of the acceptability of sound zoning scenarios because the acceptability of auditory interference scenarios is likely to be a multi-faceted problem, dependent on multiple characteristics of both the target and interferer audio programmes. Put another way, it seems very likely that the sound zoning method and stimuli involved have some effect upon the acceptability of the listening scenario beyond the resultant SNR. If this assertion is true, a multi-feature model of acceptability will be capable of greater prediction accuracy (while maintaining robustness), and this will allow for a deeper understanding of those aspects which affect the listening scenarios under consideration. If the assertion is false, however, a single-feature model utilising SNR should represent the optimum model of acceptability, and there would be no reason to assume that any other features have a significant effect upon acceptability.

On the foundation of this argument, then, the appropriate benchmark against which to
judge the performance of the constructed acceptability model would be the performance of the single-feature SNR based model of acceptability.

8.1.2 Data sets for training and testing models

Two experiments were described in chapter 7, the first of which was designed to produce acceptability data for the training of an acceptability model and the second of which was designed to produce a smaller quantity of acceptability data for the validation of the acceptability model. In addition to these, the acceptability data procured during the speech intelligibility experiment described in chapter 5, and the masking and acceptability experiment described in chapter 3 could potentially be utilised for either training or testing. The latter data set recorded acceptability thresholds rather than binary data, however, and would therefore be difficult to compare with the other data sets. The acceptability data gathered from the speech intelligibility experiment, however, could be utilised as an additional validation set.

The justification for this application of data sets is as follows: the widest range of ecologically valid stimuli were used in the acceptability experiment, making it ideal for training. The data gathered from the speech intelligibility experiment (hereafter referred to as ‘validation 1’) were produced using a methodology and stimuli fairly similar to that of the training data, which makes it ideal for validating that the model extrapolates well to new stimuli. The remaining data set (hereafter referred to as ‘validation 2’), having been gathered using stimuli processed through a sound zoning system and auditioned over headphones makes it better suited to an extremely challenging type of validation: simultaneous validation to new stimuli and reproduction methods. Therefore, a model which validates well to validation 1 would be considered robust to new stimuli, whereas a model which validates well to validation 2 would be, to some degree, considered robust to sound zone processing techniques. The key differences between the three data sets are outlined in table 8.1.

8.1.3 Model metrics

In training and testing the models constructed, metrics are required to describe both the accuracy and robustness of the predictions. This section describes the selected metrics, the justification of their selection, and their calculation. Subsequently, the schema for the application of these metrics across data sets is laid out.

Accuracy

To evaluate the prediction accuracy of the model there are broadly two groups of metrics which may be used: the error and the correlation. The error is a measure of the distance between the model predictions and the subjective data, whereas the
correlation is a measure of the extent to which these two quantities vary in the same manner. These two approaches to describing accuracy are very similar and generally produce similar trends. It is possible, however, to have two sets of predictions with the same correlation yet with different error (or vice versa); for example, if all the predictions have a constant offset from the subjective data the correlation will be very high, even though the error may be high. When such occasions arise it is usually an indication that by making an appropriate adjustment to the model (or some of its features) the predictions can also have low error. Fundamentally, however, a model which makes accurate predictions is one which has low error, and this should therefore be the ultimate metric of importance.

Multiple metrics describing error and accuracy exist. In this work R is used for correlation whereas RMSE, and RMSE* are used to describe error. These metrics were previously described in eqs. (3.1), (3.3) and (4.3) on pages 71, 72, and 95 respectively. A few additional points should be noted however.

Firstly, the denominator of the RMSE equation, $n - k$ inherently penalises models with a greater number of features; this is useful when building multi feature models because as the number of features increases a regression is more closely able to map the predictors to the response data. However even if the predictors are entirely random, the inclusion of a greater number of features will allow a regression to more closely map the predictors to the response data. When tested on new data, however, the model is unlikely to generalise well because the features did not actually describe the phenomenon being modelled in a meaningful way. This is an example of overfitting and, in the extreme example, if $k$ is equal to $n$ the RMSE score will be calculated to be infinity.
Secondly, in order to calculate the RMSE*, the confidence interval is required. The data under investigation here is binary, however, so the confidence intervals were calculated using the normal approximation to the binomial distribution calculated with:

\[ CI_{95} = 1.96 \times \sqrt{\frac{p(1-p)}{n}} \]  

where \( p \) is the proportion of subjects describing the trial as acceptable Nas et al. (2010). It should be noted that when using the normal approximation to the binomial distribution, confidence intervals will have width 0 for trials in which all subjects agree. As a result, the RMSE*’s (see eq. (4.3)) calculated during model training will be artificially inflated, and so should be interpreted with caution. However, since this bias is related to the subjective scores it will affect all constructed models equally and thus does not obstruct model training.

**Robustness**

It is important that the model should be robust to new stimuli, and one way to help ensure this is to minimise the extent to which multiple features are utilised to describe a single cause of variance in the training data. For example, if SNR, target level, and interferer level are all found to correlate well with the subjective data, it may be wise to avoid using all three features in one model since the SNR is entirely contingent upon upon the target level and interferer level. In some cases it may be less clear when multiple features describe the same phenomena, and it is therefore useful to have an objective method for estimating this. One way to achieve this is to calculate the multicollinearity of the features in the model, i.e. the degree to which the actual feature values vary together. When multicollinearity is high, it is likely that both features are describing the same, or similar, characteristics of the data. The multicollinearity can be estimated using the Variance Inflation Factor (VIF), which is calculated once for every feature in the model. The VIF calculation for a particular feature is carried out by first performing a linear regression to all other features, e.g.:

\[ X_n = \lambda_1 X_1 + \lambda_2 X_2 + \ldots + \lambda_{n-1} X_{n-1} \]  

where \( X_n \) is the \( n \)th feature of the model, and \( \lambda_n \) is the coefficient for the \( n \)th feature. For each linear regression model the coefficient of determination \( (R^2) \) is calculated, and the VIF is then given by the following equation:

\[ VIF_n = \frac{1}{1 - R_n^2} \]
Therefore, if two features have no correlation with one another the VIF will be 1, and if two features are perfectly linearly correlated (negatively or positively) the VIF will be infinity. A search for multicollinearity within a regression model can therefore be conducted by calculating the VIF for every feature, and noting that particularly high values may indicate redundancy in the feature selection.

Hair and Anderson (2010) recommend that the features in a model should have VIF no higher than 10, which corresponds to the standard errors of the model features being ‘inflated’ by a factor of three (√10 = 3.2), but they also warn that for small training sets a more stringent threshold should be enforced. O’Brien (2007), however, cautions that such thresholds are arbitrary and that some contexts permit VIFs much higher than 10 whereas for other contexts a VIF of 10 represents extreme multicollinearity. Instead, they argue, it is important to identify the cause of the multicollinearity and make a contextually informed decision about the validity of the model.

In this work, therefore, consideration is given to the causes of high VIFs without imposing an arbitrary threshold. Although many of the features which would be ultimately excluded were known to describe similar phenomena, it is sometimes worthwhile to include multiple similar features to find which offers the best performance.

8.2 A model using CASP based features

One place to start modelling the prediction of acceptability would be to use features derived from the CASP model. This is a convenient starting point since CASP has already been used earlier in this work to model the human auditory system in a physiologically inspired way.

8.2.1 Features

The final step before model training is the construction of a list of features (sometimes called ‘predictor variables’). In order to construct a model to predict acceptability features must be identified and combined into a cohesive model. The identification of features requires contextual understanding of the problem and is therefore difficult to entirely automate. As a result a ‘complete’ list of possible features is unachievable. Instead, a large number of features which might reasonably be expected to relate to the listening scenario are tested. It can never be guaranteed, therefore, that every relevant feature has been identified, but with a sufficiently large number of plausible candidate features there may be a reasonable degree of confidence that the relevant avenues of investigation have been considered.

The CASP model was used (excluding the final modulation filterbank stage) to produce internal representations of the target, interferer, and mixed stimuli. From these
representations a wide range of features was derived. The stimuli were divided into 400 ms frames stepping through in 100 ms steps and each frame was processed using the CASP model. Three groups of features were derived from the resulting frames: standard framing (SF), no overlap (NO), and 50 ms no overlap (50MS). SF features were obtained by time framing in the way previously found to be optimal for masking threshold predictions in chapter 4, NO features were based on the signals reconstructed by using only every fourth frame (i.e. as if the stimuli had been processed by CASP in 400 ms non-overlapping frames), and 50MS features were derived using the NO internal representation signal broken into 50 ms frames. The 50MS condition was included because short-time analysis is sometimes worthwhile for speech in order to capture information over one or several phonemes. It is important to note that the NO internal representation would not be identical to an internal representation produced by using CASP to process the entire signal in one chunk. This latter scenario, however, is not considered since for a real system it is likely that the stimulus duration would be indefinite, and thus some type of framing schema would be necessary.

From these internal representations a large number of features were derived. In each case, the features were constructed due to an expected relationship with acceptability.

**Time-level features**

One set of features was based on the level of the internal representations across time, which is related to the perception of the level of the programmes, and thus would likely relate to acceptability. Three minimum level features were derived for the target, interferer, and mixture programmes: TMinLev, IMinLev, and MMinLev respectively. These features were calculated by summing across all time-frequency units of the internal representation within each 400 ms frames. The resulting vector indicates the total level of each 400ms frame, and of these the lowest value was selected for use as a feature. These features therefore describe, for the target, interferer, and mixture programmes, the energy of the 400ms frame with the least energy. By recording the level of the frame with the highest level three more features, TMaxLev, IMaxLev, and MMaxLev, were constructed. A further six features were constructed by taking the ranges and standard deviations of these frame vectors. These features indicate the variation of frame level over time and therefore describes the dynamic range of the programmes; these features are referred to as TRanLev, IRanLev, and MRanLev, TStdLev, IStdLev, and MStdLev. In total, there were 12 features in this group.

**Frequency-level features**

Another set of features was based on the relative intensities across frequency. For these features the same process was followed as for the previous 12 features however instead of summing across frequency bands and samples within each frame (and then calculating quantities based on the frame vector), the internal representations were summed across
samples within each frame and across frames (but not across frequency bands). In this way the TMinSpec, IMinSpec, and MMinSpec, and the TMaxSpec, IMaxSpec, and MMaxSpec features represent the minimum and maximum level in any frequency band respectively. In addition to these 6 features, it may also be useful to record the frequency band which had the highest and lowest intensities. Thus the TMinF, IMinF, and MMinF, and the TMaxF, IMaxF, and MMaxF are represented as the number of the frequency bin (i.e. 1-31) which had the highest level (averaged across and within all frames).

For the range and standard deviation, the TRanSpec, IRanSpec, and MRanSpec, and the TStdSpec, IStdSpec, and MStdSpec, represent the change in level across frequency bands. To better understand the meaning of these features, it can be noted that broadband white noise, having equal energy across all frequencies, would have a StdSpec of 0, and a sine tone would have a fairly high StdSpec.

In section 3.2.3, subjects reported that sibilance from interfering speech and cymbals in interfering pop music were particularly problematic, so it may be that high frequencies in the interferer programme are particularly noticeable. One way to account for this would be to record the ratio of energy in the higher frequencies to that in the lower frequencies. At precisely which frequency to draw the boundary between ‘high’ and ‘low’, however, is unclear. In the musical information retrieval toolbox of Lartillot and Toivainen (2007), one of the many available features is similar to the process described here and is referred to as ‘brightness’. Two cut-off thresholds are suggested in the toolbox: one at 1kHz based on the work of Laukka et al. (2005), and one at 3kHz based on the work of Juslin (2000). For this work, therefore, both cut-off points were recorded and the cutoffs at 1kHz are referred to as TSpec1, ISpec1, and MSpec1, and the cutoffs at 3kHz are TSpec3, ISpec3, and MSpec3.

If both programmes have substantial high frequency content, the interfering high frequencies may be obscured by the target. To consider this, features were calculated by subtracting IMaxF from TMaxF, and by subtracting MMaxF from TMaxF; these are referred to as SpecFDiff and SpecFChange. The first gives an indication of the distance (in frequency bands) between the peak level of the interferer and the target, and the second gives a similar indication for the mixture and target. Another two features, AbsSpecFDiff and AbsSpecFChange, were calculated by taking the absolute value of each of these features, such that whichever programme had the higher frequency peak level was no longer relevant - merely the distance between the peaks.

In total this constitutes a further 28 features.

**Correlation features**

A cross-correlation feature, based on the $\mu$ value of the CASP model, is calculated by multiplying each time-frequency unit in the target programme by the corresponding
unit in the mixture programme and summing across time and frequency (for each frame). These values are then divided by the number of elements in the matrix, and the resulting vector describes the similarity between the programmes over time. The mean and standard deviation of this vector were taken as features, ‘XcorrMean’ and XcorrStd.

In Huber and Kollmeier (2006), a model of audio quality is described based on a similar approach to the XcorrMean feature. The Dau et al. (1997) model, a variant of the CASP model, is used to process the reference and test signal, before the cross correlation is used to produce the Perceptual Similarity Measure (PSM) which is taken to be a measure of the audio quality because low correlation indicates that severe degradations are present in the test signal. In addition to the PSM, a measure called the PSMt is calculated, which is based on taking the 5% quantile of multiple cross correlations for the signals processed in 10ms frames (but subsequently weighted using a moving average filter). In the paper, no explanation is given for the selection of the 5% quantile, but it seems reasonable that this choice produces a metric which describes the worst degradations in a way which balances both the severity of the degradations with the frequency of them.

In order to consider the possibility that this may be a more powerful feature than the mean cross correlation, a feature was produced based on both the 5% and 95% quantiles: Xcorr5per and Xcorr95per.

Thus, based on cross-correlation, a further four features were included in the pool.

**SNR based features**

Features based on the SNR of the internal representations were also calculated. For each frame, the target programme internal representation was summed across time and frequency and divided by the equivalent sum for the interferer programme. The mean and minimum of this vector were taken as the features, MeanSNR and MinSNR, to describe the relative intensities over time.

Another way of investigating this internal SNR was also considered. Each unit in the time-frequency map of each frame of the target programme was divided by the equivalent time-frequency unit in the mixture internal representation. The resultant time-frequency map therefore gives the perceptual level ratio, where the level of each unit relates to perceived prominence of the target therein. Each time-frequency unit was then replaced with a 1 if it exceeded a fixed threshold, and a zero if it did not. The proportion of units marked with a 1 can then be used as a feature to describe the proportion of the mixture programme which is dominated, by at least a given threshold, by the target programme. The threshold then represents the percentage which must be dominated by the target; i.e. a threshold of 0.9 indicates that at least 90% of the level in the mixture is due to the presence of the target programme. Since the
threshold is somewhat arbitrary, 10 thresholds were used in steps of 0.1 from 0 to 0.9. These features are named DivFrameMixT0–DivFrameMixT9. The types of features were also calculated for the interferer programme divided by the mixture, these are DivFrameMixI0 – DivFrameMixI9.

A further 22 features were thus added to the feature pool based on SNR.

**Summary**

66 features were calculated based on the 400ms framed internal representations, and equivalent features were calculated for the NO and 50MS conditions. In total, therefore, 198 features were calculated to describe relevant characteristics of the target, interferer and mixture.

### 8.2.2 Model building approach

With the training and validation data sets, the evaluation metrics, and the feature list described, the approach to constructing, evaluating and refining the models can now be outlined.

Assuming that sufficient high quality data has been obtained for model training and validation, there are two primary considerations in the construction of a model: feature combination and hierarchy. The feature combination relates to which features, from a pre-specified list, should be selected, and the hierarchy relates to the way those features should be combined to form the model. A common, and powerful, hierarchy is multi linear regression. The key advantage of multi linear regression is its simplicity. A multi linear regression model is one which is of the form:

\[
y = \lambda_0 + \sum_{i=1}^{n} \lambda_i x_i
\]  

(8.4)

where \(\lambda_i\) is the linear coefficient applied to each feature \(x_i\), and \(\lambda_0\) is a constant bias. When the features are normalised, the coefficients give an indication of the relative importance of each feature to the prediction accuracy of the model; for this reason the coefficients are sometimes referred to as ‘weightings’. Additionally, feature coefficients can be used to identify poor feature selection; if two features are selected describing similar phenomena yet are assigned opposite coefficient signs this can imply that the model could be reconstructed replacing the two features with a single feature which captures the relevant information appropriately. One disadvantage to multi linear regression is that the resultant model is capable of producing predictions outside the range of acceptability scores (in this case less than zero and greater than one). While other, more sophisticated hierarchies do not suffer this disadvantage, a multi linear
regression model is more easily justified because it allows for a meaningful interpretation of the selected features (by considering the feature coefficients).

Linear regression has been used for a wide range of audio modelling tasks including timbre (Lembke et al. 2013), emotion (Eerola et al. 2009), and soundscape quality (Brocolini et al. 2012). While more sophisticated, non-linear approaches (such as neural networks) are sometimes capable of improved accuracy, in many cases (e.g. (Zhongzhou et al. 2005) and (Brocolini et al. 2012)) the difference between model training approaches is very small. However, these more sophisticated approaches often come with the cost of a model architecture which makes it very difficult to understand the reasons behind the final feature selection and weighting, whereas linear regression describes clearly the importance and direction of effect of each feature when the features are normalised.

The feature combination problem can be optimally solved by an exhaustive search (brute-force), i.e. by combining every possible combination of features in the list and choosing the model which best meets the performance criteria. A serious practical limitation of this approach, however, is expressed in (Korf 1998): “The problem with all brute-force search algorithms is that their time complexities grow exponentially with problem size. This is called combinatorial explosion, and as a result, the size of problems that can be solved with these techniques is quite limited”. To give an example of the combinatorial explosion within context, the total number of models to construct for a list of length \( n \) features is equal to:

\[
N = \sum_{k=1}^{n} \binom{n}{k}
\]  

(8.5)

For a short list of only 5 features, therefore, \( N = 31 \). For a longer feature set comprising of 30 features, however, \( N = 1.0737 \times 10^9 \). If models could be constructed at a rate of 100 per second, it would still take around 1243 days to complete this processing. It is clear, therefore, that for feature lists of the order described in the previous section (i.e. hundreds of features) the problem quickly becomes intractable. In such cases a search algorithm is required to find the most accurate solutions within a reasonable time frame.

**Training**

In this work the Matlab ‘Stepwisefit’ function was used as an implementation of a stepwise search algorithm for training an initial multi-linear regression model. This recursive algorithm works similarly to a greedy algorithm, in that on every step it may pick the next best feature, but it may also remove features if they become redundant. In order to do this it uses two threshold values: p-enter, and p-remove. By default, these
have values of 0.05 and 0.10. The p-enter threshold indicates that, with the significance level set to 95%, the addition of a new feature will only occur when the regression to the new model which includes the new feature would result in the new feature having a coefficient which can be said to be statistically significantly different from zero (i.e. where the new feature contributes to the model). The other threshold value, p-remove, allows for the algorithm to discard features which were previously selected when the addition of multiple new features renders them redundant (i.e. when they no longer have a coefficient significantly different from zero). The entrance threshold is stricter than the exit threshold, and this helps avoid nesting in the algorithm (where infinite steps are taken sequentially adding and removing the same feature).

The Stepwisefit algorithm works by taking ‘steps’, wherein each step comprises of calculating the p-value of an F-statistic for every single new feature in the feature pool, with the null hypothesis being that the coefficient of the new feature is equal to zero (i.e. that the feature is irrelevant). When this has been completed for every candidate feature, the feature with the lowest p-value is picked (provided that this p-value is lower than the p-enter threshold of 0.05), and the next step begins. When there are no remaining candidate features to add with a p-value below p-enter, subsequent steps test whether any currently selected features have a p-value above p-remove (0.10), and removes these starting with the highest p-value. After the removal steps are complete (when no currently selected features have a p-value above p-remove), steps continue by adding any features with p-value below p-enter again. These two processes alternate until there are no more feature to add or remove. The pseudocode algorithm shown overleaf elucidates this procedure.

In this way the algorithm selects, one by one, new features which are most likely to improve the prediction accuracy of the model. The removal stage allows for the removal of features which have subsequently become obsolete; this can occur when the combination of two or more features describes the variance which was also already (less accurately) described by a single feature in the model. According to (Bowerman and O’Connell 1986), it is common practice to use 0.05 for the entry criterion and this, along with the default removal criterion of 0.10, were the values used in this work.

After training a model using this algorithm, the history of steps was manually investigated to search for instances of multicollinearity and other signs of poorly selected features. From here, further manual adaptations could be made (e.g. by omitting or adapting features).

It was likely that the process would result in an over fitted model (since many features on the list were similar). This possibility of over fitting necessitates a cross validation step.
Chapter 8: Building a Model to Predict Acceptability

Algorithm 1 An outline of Matlab Stepwisefit

1: an existing model is selected, such as $y = \text{some constant } C$.
2: \textbf{while} at least one candidate feature has $p < 0.05$, or at least one selected feature has $p > 0.10$ \textbf{do}
3: \hspace{0.5em} \textbf{while} at least one candidate feature has $p < 0.05$ \textbf{do}
4: \hspace{1.0em} \textbf{for} $i \leftarrow 1$ to All candidate features \textbf{do}
5: \hspace{1.5em} $ptest(i) = \text{GetPValue}(i)$
6: \hspace{1.0em} \textbf{end for}
7: \hspace{1.0em} \textbf{if} $\min(ptest) < 0.05$ \textbf{then}
8: \hspace{1.5em} $\text{AddFeature}(\min(ptest))$
9: \hspace{1.0em} \textbf{end if}
10: \hspace{0.5em} \textbf{end while}
11: \hspace{0.5em} \textbf{while} at least one selected feature has $p > 0.10$ \textbf{do}
12: \hspace{1.0em} \textbf{for} $i \leftarrow 1$ to All features in model \textbf{do}
13: \hspace{1.5em} $ptest(i) = \text{GetPValue}(i)$
14: \hspace{1.0em} \textbf{end for}
15: \hspace{1.0em} \textbf{if} $\min(ptest) > 0.10$ \textbf{then}
16: \hspace{1.5em} $\text{AddFeature}(\max(ptest))$
17: \hspace{1.0em} \textbf{end if}
18: \hspace{0.5em} \textbf{end while}
19: \hspace{0.5em} \textbf{end while}

Cross validation

During training, the robustness of the models to new stimuli was considered by using a 2-fold cross validation method on the training data. This involves randomly shuffling the 200 trials and splitting the data set into two 100 trial ‘folds’. The model was trained on one fold and the RMSE of the predictions for the other fold was calculated, before swapping the folds and repeating. The two RMSEs were averaged and this mean RMSE was reported. In this way a 2-fold cross validation procedure gives an indication of how the trained model is likely to perform when used to predict new test data. It is worth noting that even for models which generalise very well, the 2-fold cross validation procedure will tend to produce RMSEs which are higher than calculated on the full model, since the procedure has only half the data on which to train.

It is possible that when the stimuli are randomly shuffled, each fold may contain stimuli with very different characteristics for one or more of the features in the model. When this happens, the cross validation accuracy will be artificially diminished, and the scores will give an unreasonably pessimistic indication of the generalisability of the model. The optimal solution to this problem involves exhaustively evaluating every possible pair of stimulus-fold assignments. This process, however, is subject to a similar type of combinatorial explosion as in the model training stage. In this case an exhaustive search would require $\binom{200}{100} = 9.0549 \times 10^{58}$ combinations to be evaluated. This is impractically large, however a simple solution exists to the practical problem of obtaining the mean cross validation score: the mean score can be estimated by taking a random sample
of all possible combinations. In this work, ten thousand 2-fold cross validations were performed for each model under test, and the mean RMSEs (across folds and samples) was compared with the RMSE reported in the training stage to give an indication of the robustness of the model to new data.

Validation

Finally, for each model of interest, predictions were made for the data sets from validation 1 and validation 2. These predictions were tested for error and correlation, giving an indication of the robustness of the model to new contexts, listening scenarios, and sound zone processing.

8.2.3 A benchmark model

When evaluating a model it can be useful to have a benchmark against which to compare the model to aid in interpreting the performance. For a complex model, a good benchmark will often be a simple model, for if a simple model can achieve equal accuracy and robustness the additional complexity becomes unwarranted. In this work a simple benchmark model can be constructed based on a linear regression of SNR, since SNR is known to correlate well with acceptability scores and is a quantity which is fundamental to the sound zoning problem. As discussed in section 7.2.3 the correlation between SNR and subject scores was $R = 0.91$ and $R^2 = 0.83$. Training a linear regression model to SNR based on the 200 trials in the acceptability experiment produces the model:

$$A_p = (0.0264 \times \text{SNR}) - 0.0492 \quad (8.6)$$

Where $A_p$ is the predicted acceptability. The predicted acceptability scores have RMSE = 15.97% and RMSE* = 9.45%. It is worth noting that 34 of the 200 predictions fall outside the range $0 < A_p < 1$.

Cross validation

Ten thousand 2-fold models were produced and the mean RMSE was 16.22%. The standard deviation of the RMSEs was 0.2%, although a histogram shows that the scores were skewed towards lower RMSEs (see fig. 8.1). The skew occurs because on some repeats many of the data points least well predicted by SNR are clustered into the same fold, whilst in most repeats the less well predicted data points are more evenly split between the folds. Because of the skewed distribution, the standard deviation describes a slight over-estimate of the variation across the repeats.

The cross validation shows that, as expected, the simple benchmark model which utilises
only SNR generalises well to new stimuli.

Validation

The benchmark model was used to produce predictions for validation 1. The predictions had correlation $R = 0.7839$ and $R^2 = 0.6145$, with RMSE = 19.95\% and RMSE* = 6.55\%. Figure 8.2(a) shows the model predictions and acceptability scores.

Since there were only seven listeners the mean acceptability scores are coarse (8 steps spaced by 12.5\%); as a result the model is likely to be more accurate than indicated by the correlation and error statistics given. The speech intelligibility listening test, from which the validation 1 data set was derived, featured ‘repeat’ trials, across which the target sentence differed but all other characteristics (e.g. SNR, target speaker, interferer programme) were identical. By averaging across these trials the number of data points may be halved, as is the spacing between mean acceptability scores. It should be noted that this approximation, while increasing the resolution of mean acceptability scores, does not increase the number of listeners (although the number of judgements per mean acceptability score is doubled).

The benchmark model predictions were repeated for the mean acceptability scores averaged across subjects and repeats. Figure 8.2(b) shows the model predictions and acceptability scores for these cases. These predictions had correlation $R = 0.8634$ and $R^2 = 0.7455$, with RMSE = 16.69\% and RMSE* = 5.20\%. The improved correlation and reduced error imply that the large steps in the mean acceptability scores are at
Chapter 8: Building a Model to Predict Acceptability

Figure 8.2: Predicted acceptability scores plotted against mean acceptability scores for validation 1. The black dash-dotted line represents a perfect positive linear correlation. In plot (a) the mean acceptability scores are averaged across seven subjects for 144 trials, whereas in plot (b) the mean acceptability scores are averaged across seven subjects and repeats for 72 trials.

least partially responsible for the reduced correlation and increased error obtained before averaging.

The RMSE (16.69%) was very similar to that obtained in the cross-validation (16.22%), implying that this model is stable and robust to new stimuli with only a small decrease in accuracy compared with the training data (15.97%).

The benchmark model was subsequently used to produce predictions of acceptability for validation 2. The predictions had correlation $R = 0.1268$ and $R^2 = 0.0161$, with RMSE = 21.56% and RMSE* = 10.9%. Figure 8.3 shows the model predictions and acceptability scores. When the data point identified as likely to be an outlier, is excluded the predictions for the remaining 23 data points have $R = 0.3485$ and $R^2 = 0.1215$, with RMSE = 17.32% and RMSE* = 6.44%.

Since the SNR was dictated by the sound zoning method for validation 2 the range of SNRs was much smaller than in the training data set. The SNRs ranged from 2.7 to 18.7 dB with a mean of 11.4 and a standard deviation of 4.9, whereas the training data set had SNRs ranging between 0 and 45 dB with a mean of 22.7 and a standard deviation of 13.2. Since the range of SNRs was relatively small for the validation experiment, it is likely that listeners weighted other characteristics of the listening scenario as being more important to their judgement of acceptability than in the training set. It is also possible that the impression of spatial separation, or new artefacts introduced by the sound zoning method, are partly responsible for the poor validation.

Summary

In summary, therefore, the benchmark model based on SNR performs well with correlation $R = 0.91$ and RMSE= 16% for training and cross-validation. For validation 1
the performance was very similar with $R = 0.86$ and RMSE = 16.6% for data averaged across subjects and repeats. For validation 2, however, the correlation was small to moderate ($R = 0.35$) and the RMSE was reduced to 17% (with one outlier excluded). These scores imply that the model generalises very well to new stimuli, but rather less well to the listening scenario featuring the sound zoning method.

As previously stated, the benchmark model is unable to distinguish between sound zoning systems and programme items which result in identical SNRs. More complex models of acceptability would need to exceed the accuracy of this model, and match the robustness in cross-validation and validation, in order to be considered superior.

### 8.2.4 Constructing the model

A series of models were constructed using the procedure described in section 8.2.2, and the features described in section 8.2.1. Figure 8.4 shows the accuracy and generalisability of the models produced in each step compared with the benchmark model discussed in section 8.2.3. From steps 2 until 15 the RMSE, RMSE*, and 2-fold RMSE are lower for the constructed acceptability model than for the benchmark model.

Generally when adding more features, if the RMSEs and RMSE*'s decrease while a cross-validation metric (such as the 2-fold SNR) increases this is a good indication that further improvements in accuracy to the prediction of the training data are simply over-fitting (and should therefore not be considered generalisable). In this case, the 2-fold RMSE increases from 14.06% on step 8, to 14.67% on step 9, however the 2-fold
Figure 8.4: A plot showing the accuracy and generalisability of the acceptability model constructed in each step of the stepwise regression procedure compared with the benchmark model. The solid lines represent measurements for the constructed acceptability model and the dot-dashed lines represent measurements for the benchmark model. In each case the blue line represents the RMSE, the black line represents the RMSE*, and the red line represents the 2-fold RMSE.
RMSE subsequently continues falling after this on every step. This, alone, is therefore insufficient to exclude any of the models. An investigation into the features selected is therefore worthwhile.

Table 8.2 shows the selected features, their ascribed coefficients, and the calculated VIF for steps 1 – 3. On step 2 the highest VIF is 2.61, whereas on step 3 the highest VIF is 29.71, one order of magnitude greater. The reason for the sudden increase in multicollinearity seems to be due to the inclusion of the DivBadFrameMixI9 (NO) feature, which correlates very well with the already included DivBadFrameMixT8 (NO) feature ($R = -0.97$). This is unsurprising because one feature describes the proportion of time-frequency units in which the target programme accounts for more than 80% of the level, whereas the other feature does the same but for the interferer programme and with the threshold set at 90%. Considerable overlap would therefore be expected. This, in itself, may not be sufficient grounds for the exclusion of the model (or either feature), however since the coefficients of the two features are both positive (yet describe opposed phenomena) it is reasonable to suggest that the model is an overfit to the data. The model produced in step 2 was therefore selected as a candidate model since it was prior to any coefficient reversals and prior to inflated VIFs, as well as being prior to a divergence between RMSE and 2-fold RMSE. The model features include:

1. $x_1$: DivBadFrameMixT8 (NO), and
2. $x_2$: IStdLev (NO)

As discussed in section 8.2.1, the first of these features describes the proportion of time-frequency units in the internal representation of the mixed programmes can said to be accounted for by more than 80% by the equivalent time-frequency unit in the internal representation of the target programme. Specifically, this was for internal representations with no time frame overlaps, with time-frequency units calculated as samples by frequency bins. The second feature represents the standard deviation of...
the level of the internal representation of the interferer programme, averaged across frequency; thus this feature describes the constancy of the overall level of the interferer programme over all samples. The positive coefficient for the first feature, and the negative coefficient for the second feature indicate that as more of the mixture can be accounted for by the target programme, and as the interferer level varies less over time, the likelihood that the listening scenario will be considered acceptable increases.

The correlation between the training acceptability data and the DivBadFrameMixT8 feature for these programmes was $R = 0.8952$. For the IStdLev (NO) the correlation was $R = -0.8367$. Scatter plots of the feature values and the training acceptability data are presented in fig. 8.5. The linear regression model for the raw (without normalising) features is given by the equation:

$$A_p = 1.7565 x_1 - 0.0002 x_2 - 0.3477. \quad (8.7)$$

On the training data, the model had $R = 0.9208$ and $R^2 = 0.8478$, with RMSE = 15.03% and RMSE* = 8.91%. The 2-fold RMSE was 15.47%. For all of these metrics, therefore, the model predicted the training data more accurately than the benchmark model.

### 8.2.5 Validation

The model was used to produce predictions for validation 1. The predictions had correlation $R = 0.7584$ and $R^2 = 0.5752$, with RMSE = 23.80% and RMSE* = 8.94%. As before the data were averaged across repeats and predictions were made for these new data. The predictions had correlation $R = 0.8420$ and $R^2 = 0.7090$, with RMSE = 20.89% and RMSE* = 8.43%. All of these metrics of model accuracy were poorer than those of the benchmark model, which had $R = 0.8634$, with RMSE = 19.95% and
Chapter 8: Building a Model to Predict Acceptability

Figure 8.6: Predicted acceptability scores plotted against the mean acceptability scores of validation 1. The black dash-dotted line represents a perfect positive linear correlation. In plot a the mean acceptability scores are averaged across seven subjects for 144 trials, whereas in plot b the mean acceptability scores are averaged across seven subjects and repeats for 72 trials.

RMSE* = 5.20% for the averaged data. The original and average predictions are shown in fig. 8.6.

The model was subsequently used to produce predictions of acceptability for validation 2. The predictions had correlation R = 0.0333 and R^2 = 0.0011, with RMSE = 83.85% and RMSE* = 64.81%. These predictions were extremely poor because all predicted values were below 0. The correlation, however, was also very low and this seems to be due to the poor correlation between the IStdLev (NO) features and the validation 2 acceptability scores of R = −0.0091. By contrast, the DivBadFrameMixT8 feature had a small to moderate positive correlation with acceptability scores of R = 0.3997.

8.2.6 Summary

A stepwise regression method was utilised to identify 18 possible models for predicting acceptability, each producing greater accuracy on the training data. The multicollinearity, coefficients, and features were carefully examined and there was good evidence to exclude models 3 – 18. Model 2 was therefore selected for validation testing because it did not include features describing similar phenomena with opposed coefficients. The model performance exceeded the accuracy of the benchmark model for the training and cross-validation data, but generally performed poorer than the benchmark model for the two validation data sets. The good cross-validation performance, and the reasonably high correlation with validation 1 indicate that the features may be useful to include in a more extensive model.
8.3 A search for further features

In the previous section the construction of a model was described using features derived from the CASP model. While it is clear that some of these features offer promising initial results, it is also clear that these features alone were not able to produce a model superior to the benchmark linear regression to SNR. In this section a range of additional non-CASP based features are introduced to the feature pool and the model training procedure is repeated to see if more accurate and robust predictions are possible.

8.3.1 Features

In addition to the previously discussed CASP based features, a further set of features was produced by considering the level, loudness, and spectra of the programmes (without any auditory processing). This section gives an overview of these additional features.

**Level and loudness based features**

A range of features were calculated to describe the level of the stimuli. Simplistic features based on the RMS level of the items were obtained including the target level (RMS-TarLev), the interferer level (RMS-IntLev), and the SNR (RMS-SNR). In addition to these, a range of features were produced describing the proportion of the stimuli for which the SNR fell below a fixed threshold. These were calculated by dividing the programmes into 50 ms frames, and calculating the RMS SNR for each frame. The features were then taken as the proportion of frames in which the SNR did not exceed a fixed threshold. Thresholds ranged from 0 dB to 28 dB in steps of 2 dB. These features therefore incorporate some time-varying information, and since they describe the proportion of frames which had a poor SNR, were referred to as RMS-BadFrame0, RMS-BadFrame2, ... RMS-BadFrame28.

It was considered possible that psychoacoustically based loudness features might perform better than standard measures of level. For this reason a range of features were obtained using the loudness model in the Genesis toolkit (Genesis-Acoustics 2013) based on the model proposed by Zwicker and Fastl (1990). These included TLoud, the loudness level exceeded during 30 ms of the signal (the default duration), TLoud50, the loudness level exceeded during 50 percent of the signal, Tmax, the maximum instantaneous loudness of the target, and the equivalent features for the interferer (ILoud, ILoud50, and IMax). In addition to these, the loudness ratio LoudRat (TLoud − ILoud), the peak loudness ratio LoudPeakRat (TMax − IMax), and the peak to loudness target and interferer ratios TMaxRat and IMaxRat (TMax − TLoud, and IMax − ILoud), were calculated.
A further 28 level and loudness based features were therefore added to the total feature pool.

**Spectral centroid features**

It was also considered likely that the relative frequency spectra of the stimuli would be relevant to the acceptability. In line with this, subjects had occasionally commented that higher frequency interference was especially problematic. The mean and standard deviation of the spectral centroid for each stimulus was therefore calculated using the Matlab code of Nagel (2013). By default, a vector of spectral centroids is produced based on frames of 2048 samples (4.6 ms at 44100 Hz) with 80% overlap. The means and standard deviations of these were taken to be used for the features TSpecMean, ISpecMean, TSpecStd, ISpecStd, as well as the euclidian distances of these quantities SpecMean (TSpecMean - ISpecMean), and SpecStd (TSpecStd - ISpecStd).

A further 6 features were therefore added to the pool.

**Manual features**

Although in principle all features can be derived from the stimuli directly, in practice the acquisition of some 'higher level' (i.e. cognitive) features from the stimuli are very difficult problems which represent fields of study in their own right. Instead, such high level features can be produced by a human listener identifying the relevant traits.

In this work, three such features were directly coded by the author based on auditioning the stimuli. Since during the experiments in chapters 3, 5 and 7 subjects commented that speech is a more problematic interferer than music when the target programme is also speech, it was deemed worthwhile to obtain features describing this aspect of the target and interferer programmes. No computational models are known to the author capable of accurately detecting whether an arbitrary audio sample contains speech, whereas humans are adept at this process. The task is somewhat complicated by defining the boundaries of the feature (e.g. do musical vocals count as speech?). Three manual features were therefore coded to describe the extent to which a target or interferer programme is 'speech-like'. These were: ManSpeech, ManSpeechOnly, and ManInst. The first feature was coded as a 1 when the interferer contained speech (excluding musical vocals), and 0 otherwise, the second feature was coded as a 1 when the interferer contained only speech (e.g. with no background music), and 0 otherwise, and the third feature was coded as a 1 when the interferer contained only instrumental music (i.e. did not contain any linguistic content), and 0 otherwise.

The correlation between the mean acceptability scores and the manually coded features were R= 0.0048, R= −0.0099 and R= 0.0057 respectively. The correlations were very low, and so these features are unlikely to be very important for predicting acceptability. Nonetheless, subjects occasionally reported that interfering speech was
more problematic than interfering music, so it is possible that a covariate exists which predicts acceptability well yet occurs more commonly with speech interferers than with musical interferers (e.g. high frequency content or temporal sparsity).

Three manually encoded features were therefore included in the feature pool.

**Summary**

A further 37 features were therefore collected describing the level, loudness, and spectra of the stimuli, as well as accounting for subjective comments about speech-speech interactions. These were added to the CASP based features producing a total feature pool of size 235.

### 8.3.2 Constructing the model

Again, a series of models were constructed using the procedure described in section 8.2.2, and the features described in both section 8.2.1 and section 8.3.1. Figure 8.7 shows the accuracy and generalisability of the models produced in each step compared with the benchmark model discussed in section 8.2.3. This time all steps had lower RMSE, RMSE*, and 2-fold RMSE than the benchmark model. For this new set of models, the cross validation error increased from 13.00% on step 7 to 13.03% on step 8.

Table 8.3 shows the selected features, their ascribed coefficients, and the calculated VIF for each of the first 7 steps. On step 6, the highest VIF is 5.65 whereas on step 6 the highest VIF is 17.83: more than three times as high. On step 7 the DivBadFrameMixT7 feature is included, which is very similar to the DivBadFrameMixT9 feature already included. While similar features may itself not be reason for exclusion, the coefficients of these two features have opposed signs, and thus step 6 is a more appropriate choice of model. On step 5 the IStdLev feature is included, when on step 2 the IStdLev (NO) feature was already introduced. These two features describe the standard deviation of interferer level across 400ms frames and across samples respectively. Though the features seem similar, the change in time frame over which they operate constitutes an important difference between them. For the training data, these features had only a weak positive correlation of $R = 0.3410$. Furthermore, the coefficients for the normalised features are not opposed, so there is no strong evidence to suggest that the introduction of the IStdLev (NO) feature is an overfit to the training data.

The correlation between the training acceptability data and the first three features selected, RMS-BadFrame18, IStdLev (NO), and DivBadFrameMixT9, was very high with $R = -0.9252$, $R = -0.8366$, and $R = 0.8065$ respectively. The remaining three features, TSpecMean, IStdLev, and TMaxSpec, had lower correlations with $R = 0.1570$, $R = -0.1638$, and $R = 0.2669$ respectively.

The model features therefore include:
Figure 8.7: A plot showing the accuracy and generalisability of the acceptability model constructed in each step of the stepwise regression procedure compared with the benchmark model. The solid lines represent measurements for the constructed acceptability model and the dot-dashed lines represent measurements for the benchmark model. In each case the blue line represents the RMSE, the black line represents the RMSE*, and the red line represents the 2-fold RMSE.
<table>
<thead>
<tr>
<th>Step</th>
<th>Features</th>
<th>Coefficients</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RMS-BadFrame18</td>
<td>-0.3546</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>RMS-BadFrame18</td>
<td>-0.2787</td>
<td>3.0118</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0929</td>
<td>3.0118</td>
</tr>
<tr>
<td>3</td>
<td>RMS-BadFrame18</td>
<td>-0.2187</td>
<td>4.9594</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0955</td>
<td>3.0155</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0721</td>
<td>2.8129</td>
</tr>
<tr>
<td>4</td>
<td>RMS-BadFrame18</td>
<td>-0.2136</td>
<td>4.9836</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0947</td>
<td>3.0160</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0759</td>
<td>2.8258</td>
</tr>
<tr>
<td></td>
<td>TSpecMean</td>
<td>0.0332</td>
<td>1.0118</td>
</tr>
<tr>
<td>5</td>
<td>RMS-BadFrame18</td>
<td>-0.2283</td>
<td>5.3223</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0684</td>
<td>4.0969</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0822</td>
<td>2.8875</td>
</tr>
<tr>
<td></td>
<td>TSpecMean</td>
<td>0.0338</td>
<td>1.0123</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0298</td>
<td>1.3911</td>
</tr>
<tr>
<td>6</td>
<td>RMS-BadFrame18</td>
<td>-0.2225</td>
<td>5.3862</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0513</td>
<td>4.6554</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.1204</td>
<td>5.6521</td>
</tr>
<tr>
<td></td>
<td>TSpecMean</td>
<td>0.236</td>
<td>1.2088</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0402</td>
<td>1.5945</td>
</tr>
<tr>
<td></td>
<td>TMaxSpec</td>
<td>-0.0367</td>
<td>2.5492</td>
</tr>
<tr>
<td>7</td>
<td>RMS-BadFrame18</td>
<td>-0.2142</td>
<td>5.5468</td>
</tr>
<tr>
<td></td>
<td>IStdLev (NO)</td>
<td>-0.0420</td>
<td>4.8539</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.1928</td>
<td>17.8300</td>
</tr>
<tr>
<td></td>
<td>TSpecMean</td>
<td>0.0217</td>
<td>1.2177</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0404</td>
<td>1.5946</td>
</tr>
<tr>
<td></td>
<td>TMaxSpec</td>
<td>-0.0417</td>
<td>2.6076</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT7</td>
<td>-0.0612</td>
<td>8.7087</td>
</tr>
</tbody>
</table>

Table 8.3: Features, coefficients, and VIF for the first seven steps of model construction. For clarity, the intercepts have been excluded.
Chapter 8: Building a Model to Predict Acceptability

1. \(x_1\): RMS-BadFrame18,
2. \(x_2\): IStdLev (NO),
3. \(x_3\): DivBadFrameMixT9,
4. \(x_4\): TSpecMean,
5. \(x_5\): IStdLev, and
6. \(x_6\): TMaxSpec

RMS-BadFrame18 indicates the proportion of 50 ms frames within which the SNR of the target and interferer programme was less than 18 dB. IStdLev (NO) and IStdLev indicate the standard deviation of the level of the internal representation of the interferer across samples and frames respectively. DivBadFrameMixT9 indicates the proportion of the time-frequency units in the internal representation of the mixture programme of which at least 90% of the level can be accounted for by the equivalent time-frequency units in the internal representation of the target programme. TSpecMean indicates the mean spectral centroid of the target programme. Finally, TMaxSpec indicates the maximum level of the target programme in any frequency bin across 400 ms frames.

The linear regression model for the raw (without normalising) features is given by the equation:

\[
A_p = - (6.13 \times 10^{-1} x_1) - (5.84 \times 10^{-5} x_2) + (4.55 \times 10^{-1} x_3) \\
+ (6.86 \times 10^{-4} x_4) - (1.53 \times 10^{-8} x_5) - (9.61 \times 10^{-9} x_6) + 9.57 \times 10^{-1} \tag{8.8}
\]

The model predicts the training data with \(R^2 = 0.9505\) and \(R^2 = 0.9035\), with RMSE = 12.09% and RMSE* = 5.65%. The mean 2-fold RMSE was 13.03%. For all of these metrics, this model was more accurate than the benchmark model.

8.3.3 Validation

The model was used to produce predictions for validation 1. The predictions had correlation \(R^2 = 0.7785\) and \(R^2 = 0.6061\), with RMSE = 17.02% and RMSE* = 8.93%. As before the data was averaged across repeats and predictions were made for these new data. The predictions had correlation \(R^2 = 0.8564\) and \(R^2 = 0.7335\), with RMSE = 13.09% and RMSE* = 5.94%. In comparison with the benchmark model, the RMSEs for the original (19.95%) and averaged (16.69%) data were lower, yet the RMSE*s for the original (6.55%) and averaged (5.20%) data were slightly higher. Correlations for the original (0.7839) and averaged (0.8634) data were slightly lower than the benchmark as well. The original and average predictions are shown in fig. 8.8.
Chapter 8: Building a Model to Predict Acceptability

The model was subsequently used to produce predictions of acceptability for validation 2. The predictions had correlation \( R = 0.4294 \) and \( R^2 = 0.1844 \), with \( \text{RMSE} = 27.55\% \) and \( \text{RMSE}^* = 11.52\% \). While the correlation and accuracy are fairly poor here, the correlation is nonetheless much higher than the benchmark model which had \( R = 0.1268 \) and \( R^2 = 0.0161 \), indicating that some of the features are likely to be generalisable. Figure 8.9 shows the predictions for validation 2.

8.3.4 Summary

A stepwise regression method was utilised to identify 8 possible models for predicting acceptability, each producing greater accuracy on the training data. The multicollinearity, coefficients, and features were carefully examined and there was good evidence to exclude models 6–8. Model 5 was therefore selected for validation testing. The model performance exceeded the accuracy of the benchmark model for the training and cross-validation data. For validation 1, the correlations and RMSE*’s were slightly poorer, although the RMSE was improved. For validation 2, the performance was greatly improved over the benchmark model.

While the model did not produce more accurate scores for every metric on every data set, it did produce some more accurate scores on all data sets, and large improvements for validation 2 (the data set including a sound zoning method). It seems, therefore, that this extended model is more generalisable than the simpler SNR based benchmark model.
Chapter 8: Building a Model to Predict Acceptability

Figure 8.9: Predicted acceptability scores plotted against mean acceptability scores reported by the 20 subjects. The black dash-dotted line represents a perfect positive linear correlation.

8.4 Including PEASS

PEASS (Emiya et al. 2011) is a toolkit for analysing source separation algorithms. The source separation problem, which entails separating two streams of audio which have been mixed together, can be considered to be a similar problem to the sound zoning problem. The PEASS toolkit, which may be used to evaluate the overall perceptual quality of separated audio after running a source separation algorithm, is therefore a potentially useful approach to evaluating the effectiveness of a sound zoning system which, rather than separating two streams of audio, aims to keep two streams of audio from mixing.

In contrast, however, it is worth noting that the types of artefacts introduced by a sound zoning system may be quite different from those introduced by source separation methods. For example, the so-called ‘musical noise’ (Hummersone et al. 2014) that is introduced by separating via an ideal binary mask is not introduced by any of the more prominent sound zoning methods, such as those discussed in (Choi and Kim 2002; Poletti 2008; Coleman et al. 2013). Despite the differences between the source separation and the sound zoning problems, it may still be useful to include features based on PEASS.

The PEASS model works ordinarily takes a (clean) target, a (clean) interferer, and the extracted target programme (via some source separation algorithm) as an input. For this work the extracted target programme is substituted with the mixture of
programmes. The model works broadly in three steps:

1. First, a time-frequency decomposition and resynthesis method is utilised to estimate the error in the extracted target programme due to errors in the target, errors in the interferer, and artefact errors (which are defined as any further errors not yet accounted for).

2. Second, the perceptual similarity measure of Huber and Kollmeier (2006) is calculated, using the Dau et al. (1997) auditory model, for each of four pairs of signals. The first pair is the target programme and the extracted target programme (in this context the mixture programme); this is used to determine the perceptual salience of the overall error. The second pair is the extracted target programme and the error in the target programme; this determines the perceptual salience of the error due to the target. The third pair is the extracted target programme and the errors in the interferer; this determines the perceptual salience of the error due to the interferer. The fourth pair is the extracted target programme and the artefact error; this determines the perceptual salience of artefacts.

3. Finally, these four perceptual saliences are used as inputs to a neural network which, using subjective data gathered in listening tests, nonlinearly maps the perceptual saliences to four desired outputs: the Interferer Perceptual Score (IPS), the Overall Perceptual Score (OPS), the Artefact Perceptual Score (APS), and the Target Perceptual Score (TPS).

It is worth noting that for auditory interference scenarios the TPS and APS would be expected to be somewhat irrelevant and the IPS would be expected to be particularly important (since the differences are due only to the presence of an interfering programme). When including a sound zoning method, however, the TPS and APS may be more relevant. In general, the OPS would be expected to be important in all cases.

The IPS, the OPS, the APS, and the TPS were utilised as four additional features within the previous pool of features, resulting in a feature pool of 239 features describing aspects of the stimuli, their relation to one another, subjective comments, and the internal representations of the stimuli.

### 8.4.1 Constructing the model

Once again the method outlined in section 8.2.2 was used to construct models of acceptability, this time including all features discussed thus far. Figure 8.10 shows the accuracy and generalisability of the models produced in each step compared with
Chapter 8: Building a Model to Predict Acceptability

Figure 8.10: A plot showing the accuracy and generalisability of the acceptability model constructed in each step of the stepwise regression procedure compared with the benchmark model. The solid lines represent measurements for the constructed acceptability model and the dot-dashed lines represent measurements for the benchmark model. In each case the blue line represents the RMSE, the black line represents the RMSE*, and the red line represents the 2-fold RMSE.

For all steps the RMSE, RMSE*, and 2-fold RMSE are lower for the constructed acceptability model than for the benchmark model with one exception: the 2-fold RMSE for step eight was 385.27% (and therefore could not fit on the plot within a reasonable scale). The 2-fold RMSE increased from 11.93% in step 5 to 11.94% in step 6, and then fell to 11.81% in step 7 before rising steeply to 385.27% in step 8. Step 5 therefore seems to be an initially appropriate model to select pending further examination of the selected features, their multicollinearity, and the feature weightings.

Table 8.4 shows the selected features, their ascribed coefficients, and the calculated VIF for each step for steps 1 – 6. Prior to step 6 all VIFs remain below 6, but on step 6 the VIFs for two of the features exceed 70. The very high multicollinearity is explained by noting that these two features were describing the proportion of time frames with SNRs under 18 and 20 dB respectively. These two features are assigned coefficients with opposing signs, and so it seems likely that from step 6 onwards the regression is over fitting to the training data.
<table>
<thead>
<tr>
<th>Step</th>
<th>Features</th>
<th>Coefficients</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RMS-BadFrame18</td>
<td>-0.3546</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>0.1664</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>RMS-BadFrame18</td>
<td>-0.2066</td>
<td>4.796</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>0.1664</td>
<td>4.796</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>-0.2016</td>
<td>4.809</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>0.1692</td>
<td>4.800</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>-0.0452</td>
<td>1.004</td>
</tr>
<tr>
<td>3</td>
<td>RMS-BadFrame18</td>
<td>-0.1806</td>
<td>5.477</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>0.1514</td>
<td>5.278</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0488</td>
<td>1.024</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0455</td>
<td>3.153</td>
</tr>
<tr>
<td>4</td>
<td>RMS-BadFrame18</td>
<td>-0.1620</td>
<td>5.974</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>0.1427</td>
<td>5.389</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0532</td>
<td>1.052</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0614</td>
<td>3.517</td>
</tr>
<tr>
<td></td>
<td>MSpecMax</td>
<td>-0.0310</td>
<td>1.389</td>
</tr>
<tr>
<td>5</td>
<td>RMS-BadFrame18</td>
<td>-0.3271</td>
<td>76.419</td>
</tr>
<tr>
<td></td>
<td>PEASS-OPS</td>
<td>0.1510</td>
<td>5.568</td>
</tr>
<tr>
<td></td>
<td>IStdLev</td>
<td>-0.0544</td>
<td>1.056</td>
</tr>
<tr>
<td></td>
<td>DivBadFrameMixT9</td>
<td>0.0560</td>
<td>3.593</td>
</tr>
<tr>
<td></td>
<td>MSpecMax</td>
<td>-0.0321</td>
<td>1.3913</td>
</tr>
<tr>
<td>6</td>
<td>RMS-BadFrame20</td>
<td>0.1699</td>
<td>74.584</td>
</tr>
</tbody>
</table>

Table 8.4: Features, coefficients, and multicollinearity for the first 6 steps of model construction. For clarity, the intercepts have been excluded.

It is also worth noting that there is a small jump in VIF from step one to step two, and it seems likely that the RMS-BadFrame18 and PEASS - Overall Perceptual Score (PEASS-OPS) features may be describing similar phenomena. PEASS-OPS is primarily determined by the cross-correlation between a reference and degraded signal which, in this context, are equivalent to the target and mixture programmes respectively. RMS-BadFrame18, on the other hand, is determined by the time-varying SNR of the target and interferer programmes. For the training data the correlation between the features is $R = 0.89$. In this case, however, the model coefficients have opposite signs, yet they are also describing related phenomena in the opposite manner (i.e. the BadFrame feature describes the proportion of frames which fails to exceed a particular SNR). For this reason, therefore, it is not clear that the features are mutually redundant. Given this, the model produced in step 5 was selected as a candidate model. The model is defined as:
1. $x_1$: RMS-BadFrame18
2. $x_2$: PEASS-OPS
3. $x_3$: IStdLev
4. $x_4$: DivBadFrameMixT9
5. $x_5$: MSpecMax

The features in this model were selected in the prior models with the exception of PEASS-OPS, the overall preference score of the PEASS model. The linear regression model for the raw (without normalising) features is given by the equation:

$$A_p = - (4.46 \times 10^{-1} x_1) + (3.52 \times 10^{-3} x_2) - (2.02 \times 10^{-8} x_3) + (2.32 \times 10^{-1} x_4) - (1.01 \times 10^{-8} x_5) + 0.82.$$  (8.9)

The model predictions had accuracy with $R = 0.9583$ and $R^2 = 0.9183$, with RMSE = 11.09% and RMSE* = 4.99%. the mean 2-fold RMSE was 11.93%. As with the previous models, on these metrics the model exceeds the accuracy of the benchmark model.

### 8.4.2 Validation

The model was used to produce predictions for validation 1. The predictions had correlation $R = 0.7678$ and $R^2 = 0.5894$, with RMSE = 17.47% and RMSE* = 8.14%. As before the data was averaged across repeats and predictions were made for these new data. The predictions had correlation $R = 0.8462$ and $R^2 = 0.7161$, with RMSE = 13.56% and RMSE* = 5.69%. In comparison with the benchmark model, the RMSE for the original (19.95%) and averaged (16.69%) data were lower, and the RMSE* for the averaged data (5.20%) and the original data (6.55%) were slightly higher. The correlations were slightly lower than those of the benchmark model. The original and average predictions are shown in fig. 8.11.

The model was subsequently used to produce predictions of acceptability for validation 2. The predictions had correlation $R = 0.5743$ and $R^2 = 0.3298$, with RMSE = 17.83% and RMSE* = 5.00%. Figure 8.12 shows the predictions for validation 2. The RMSE and RMSE* of the predictions was lower than the benchmark model. The correlation scores were also much higher than benchmark model.

### 8.4.3 Summary

A stepwise regression method was utilised to identify eight possible models for predicting acceptability, each producing greater accuracy on the training data. The
Figure 8.11: Predicted acceptability scores plotted against the mean acceptability scores of validation 1. The black dash-dotted line represents a perfect positive linear correlation. In plot (a) the mean acceptability scores are averaged across seven subjects for 144 trials, whereas in plot (b) the mean acceptability scores are averaged across seven subjects and repeats for 72 trials.

Figure 8.12: Predicted acceptability scores plotted against mean acceptability scores reported by the 20 subjects. The black dash-dotted line represents a perfect positive linear correlation.
multicollinearity, coefficients, and features were carefully examined and there was good evidence to exclude models 6 - 8. The accuracy of model five was examined on the training, cross-validation, and validation data sets. In most cases the model had greater accuracy than the benchmark model, and where it did not the accuracy was approximately equal.

The feature selected in step one was RMS-BadFrame18. The second feature selected was PEASS-OPS. The multicollinearity between these scores was VIF = 4.796. For only two features this is relatively high. It may be that, if only one of these features is included a better solution may exist among the array of features.

8.5 SNR based hierarchy and model adjustments

Upon observing the scatter plot of acceptability scores against SNRs in section 7.1.4, it may be argued that for relatively low SNRs the acceptability scores were generally determined by the SNR and would therefore usually be close to 0, and for relatively high SNRs the acceptability scores were generally determined by the SNR and would therefore usually be close to 1. For cases in between, however, other features played a larger role, and so the variation was greater.

Under this hypothesis, a more powerful model architecture could involve first identifying whether the SNR fell below a fixed low threshold, or above a fixed high threshold. If either threshold were exceeded, the acceptability score would be set at 0 or 1 appropriately; where neither threshold is exceeded, other features, selected by a model training procedure, would be used.

This approach was implemented, selecting 12.5 and 29 dB SNR as the low and high thresholds respectively. These were selected since acceptability scores in the training data below 12.5 dB SNR never exceeded 0.2, and acceptability scores in the training data above 29 dB SNR never fell below 0.7. Upon constructing a model using the previously discussed procedure training on the middle 76 data points, the first two selected features were DivBadFrameMixI0 (NO) and IStdLev (NO); features describing very similar phenomena to those selected in the previous models. The correlation with the training data for all steps of the model construction procedure fell below $R = 0.91$ (the benchmark correlation), and so this approach was not developed further.

Although this modelling approach did not produce a more successful acceptability model, it does highlight a small improvement which can be made to the previous acceptability models. Since the previous three acceptability models were constructed using multiple linear regression, it is possible to produce predictions of acceptability which exceed 1 or fall below 0. Such predictions are not meaningful because an acceptability score of 1 indicates a probability of 100% that a listener selected at random will find the listening scenario to be acceptable, and an acceptability score
of 0 indicates a probability of 0%. Acceptability predictions can therefore be improved, and meaningless results avoided, if predictions exceeding 1 are set to 1, and predictions below 0 are set to 0. Expressed mathematically this is:

\[
A'_p = \begin{cases} 
1 & A_p > 1 \\
A_p & 0 < A_p < 1 \\
0 & A_p < 0
\end{cases}
\]  

(8.10)

where \(A_p\) and \(A'_p\) represent the acceptability prediction and adjusted acceptability prediction respectively. This modification would not be likely to make large differences to the accuracy of well trained models, however the the modification is worth implementing for the sake of more meaningful results in practical applications.

For the CASP-based model, this modification reduced the prediction error on the training data from \(\text{RMSE} = 15.03\%\) to 14.16\% and \(\text{RMSE}^* = 8.91\%\) to 8.19\%, while increasing the correlation from \(R = 0.9208\) to 0.9321. For validation 1 the prediction error reduced from \(\text{RMSE} = 23.80\%\) to 22.27\% and from \(\text{RMSE}^* = 8.94\%\) to 8.42\%, yet the correlation slightly reduced from \(R = 0.7584\) to 0.7546. When averaged across repeats the error reduced from \(\text{RMSE} = 20.89\%\) to 19.19\% and from \(\text{RMSE}^* = 8.43\%\) to 7.27\%, while again decreasing the correlation from \(R = 0.8420\) to 0.8377. These decreases in correlation reflect the reduction in linearity of correlation between predictions and observations which are caused by bounding the predictions at 0 and 1, even though this reduces the prediction error. For validation 2 the prediction error reduced substantially from \(\text{RMSE} = 83.85\%\) to 35.36\% and from \(\text{RMSE}^* = 64.81\%\) to 18.51\%, however since the unmodified predictions were all negative values these metrics describe the accuracy of predicting 0 acceptability in all cases.

For the extended acceptability model, this modification reduced the prediction error on the training data from \(\text{RMSE} = 12.09\%\) to 11.75\% and from \(\text{RMSE}^* = 5.65\%\) to 5.36\%, while increasing the correlation from \(R = 0.9505\) to 0.9536. For both validation data sets none of the predictions exceeded 1 or fell below 0 therefore these scores were unaffected.

In the case of the PEASS-based acceptability model, the modification reduced the error of the predictions for the training data from \(\text{RMSE} = 11.09\%\) to 11.05\% and from \(\text{RMSE}^* = 4.99\%\) to 4.96\%, and increased the correlation from \(R = 0.9583\) to 0.9587. The difference in model accuracy is so small because only 13 of the 200 predictions exceeded 1 or fell below 0, and all of these fell within the range \(-0.0559\) and 1.0433. Since for the PEASS-based acceptability model for both validation data sets the predictions did not included any values exceeding 1 or below 0 these scores were unaffected.

Since the latter two models performed reasonably well for all data sets, the effect of
Chapter 8: Building a Model to Predict Acceptability

this modification to predictions was very small.

8.6 Model selection

Table 8.5 shows a comparison of metrics for the benchmark model with the three models produced, including the model adjustments described in section 8.5. All three models performed better than the benchmark on the training data and cross-validation. The importance of this result should be considered, however, noting that a better model fit is often possible when more features are available, even if the features are not the best possible features with which to build a model. Generally speaking, however, when multiple poorly selected features are used in regression the accuracy of the cross-validation will be low.

For validation 1, the CASP based model performed poorly, failing to surpass the accuracy of the benchmark model in terms either of correlation or error. The other two models, however, performed similarly to the benchmark, with superior RMSEs, yet with marginally inferior RMSE*s and correlations. This trend was consistent regardless of whether the data was averaged across repeats.

For validation 2, the CASP based model again performed poorer than the benchmark. The extended model represented a large improvement over the CASP based model, and the predictions had much better correlation with the data than the benchmark predictions. The RMSE was higher than the benchmark, however, because the predictions ranged from -0.1 to 0.3; this can be explained by a linear offset caused by only a partial agreement between feature weights in the training and validation data sets. The PEASS based model performed markedly better on all metrics than the benchmark, and had improved scores compared with the extended model as well.

The PEASS based model had the best overall performance, although its performance only exceeded the extended model for the validation 2 data set. This indicates that the sound zone processing was better accounted for when using the PEASS based model. For the validation 1 data set, none of the models performed substantially better than benchmark SNR based model. The benchmark model predictions for the validation 2 data were very poor. The PEASS based model is therefore selected as the best combination of accuracy and generalisability.

8.7 Model comparison

It is worth comparing the model with existing computational models which might be brought to bear on the problem. The two most likely groups of models to apply are those which assess speech quality, and those which assess source separation. The
### Table 8.5: A side-by-side comparison of the performance of two acceptability models. Scores are highlighted in green and red by indicating performance metrics which exceeded or fell short of those of the benchmark model.

<table>
<thead>
<tr>
<th></th>
<th>Benchmark</th>
<th>CASP Model</th>
<th>Extended Model</th>
<th>PEASS Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of feature pool</td>
<td>1</td>
<td>198</td>
<td>235</td>
<td>239</td>
</tr>
<tr>
<td>Number of features</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Training Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>15.97%</td>
<td>14.16%</td>
<td>11.75%</td>
<td>11.05%</td>
</tr>
<tr>
<td>RMSE*</td>
<td>9.45%</td>
<td>8.19%</td>
<td>5.36</td>
<td>4.96%</td>
</tr>
<tr>
<td>R</td>
<td>0.91</td>
<td>0.93</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>mean 2-fold RMSE</td>
<td>16.22%</td>
<td>15.46%</td>
<td>13.03</td>
<td>11.93%</td>
</tr>
<tr>
<td>Validation 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>19.95%</td>
<td>22.27%</td>
<td>17.02%</td>
<td>17.47%</td>
</tr>
<tr>
<td>RMSE*</td>
<td>6.55%</td>
<td>8.42%</td>
<td>8.93%</td>
<td>8.14%</td>
</tr>
<tr>
<td>R</td>
<td>0.78%</td>
<td>0.75%</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>Validation 1 with re-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>peats</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>16.69%</td>
<td>19.19%</td>
<td>13.10%</td>
<td>13.56%</td>
</tr>
<tr>
<td>RMSE*</td>
<td>5.20%</td>
<td>7.27%</td>
<td>5.94%</td>
<td>5.38%</td>
</tr>
<tr>
<td>R</td>
<td>0.86%</td>
<td>0.84%</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Validation 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>21.56%</td>
<td>35.36%</td>
<td>32.81%</td>
<td>17.98%</td>
</tr>
<tr>
<td>RMSE*</td>
<td>10.90%</td>
<td>18.51%</td>
<td>15.41%</td>
<td>5.15%</td>
</tr>
<tr>
<td>R</td>
<td>0.13</td>
<td>0.03%</td>
<td>0.45</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Chapter 8: Building a Model to Predict Acceptability

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Validation 1</th>
<th>Validation-Av 1</th>
<th>Validation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ</td>
<td>0.94</td>
<td>0.75</td>
<td>0.83</td>
<td>−0.28</td>
</tr>
<tr>
<td>POLQA</td>
<td>0.91</td>
<td>0.77</td>
<td>0.84</td>
<td>−0.17</td>
</tr>
<tr>
<td>Extended Acc Model</td>
<td>0.95</td>
<td>0.78</td>
<td>0.86</td>
<td>0.45</td>
</tr>
<tr>
<td>Peass based Acc Model</td>
<td>0.96</td>
<td>0.77</td>
<td>0.85</td>
<td>0.57</td>
</tr>
<tr>
<td>OPS</td>
<td>0.91</td>
<td>0.68</td>
<td>0.76</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 8.6: Correlation scores for PESQ and POLQA predictions

PEASS model, which has already been considered as a useful resource from which to draw features, offers the OPS metric which can be considered a reasonable prediction from a source separation model. For speech quality models, Perceptual Evaluation of Sound Quality (PESQ) is the most likely choice although it is worth considering the more recent Perceptual Objective Listening Quality Assessment (POLQA) (which assesses audio quality, rather than simply speech quality) as well.

8.7.1 PEASS comparison

The PEASS OPS scores correlated with the training data with $R = 0.9136$. By way of contrast the extended and PEASS based models had correlation $R = 0.95$ and $R = 0.96$ respectively. For validation 1, the OPS had correlation $R = 0.6814$, whereas the extended and PEASS based models had correlation $R = 0.78$ and $R = 0.77$ respectively. When the data was average across repeats the OPS correlation increased to $R = 0.7597$, whereas the extended and PEASS based models had correlation $R = 0.86$ and $R = 0.85$ respectively. Finally, for validation 2, the OPS had correlation $R = 0.5462$, whereas the extended and PEASS based models had correlation $R = 0.45$ and $R = 0.57$ respectively.

The PEASS OPS performed poorer than the extended model on all but the validation 2 data set, and performed poorer than the PEASS based model on all data sets. The prediction of acceptability, therefore, benefits from including OPS as a feature, but can be made far more accurate and generalisable by the inclusion of the other features discussed.

8.7.2 PESQ and POLQA comparison

The PESQ and POLQA models were utilised to make predictions about the acceptability data sets via the PEXQ audio quality suite of tools provided by Opticom. The accuracy of the predictions are shown in table 8.6.

For the training data, the extended and PEASS based acceptability models had better correlation than the PESQ and POLQA model predictions. The OPS metric alone had
slightly higher correlation than the POLQA predictions, but lower correlation than the PESQ scores.

For validation 1, the extended and PEASS based models again had higher correlation than the PESQ and POLQA model predictions. When the data were averaged across repeats the PESQ and POLQA correlations increased to $R = 0.83$ and $R = 0.84$; these relationships are shown in fig. 8.13. The averaged extended and PEASS based models still had higher correlations however. For these data the OPS did not correlate as well with the mean acceptability scores as either the PESQ or POLQA scores.

Figure 8.13 shows an apparent outlier in both the PESQ and POLQA predictions, where for an acceptability score of 1 the predictions are only 2.4 and 3.8 respectively. These scores refer to the same trial. Since the data shown are based on averaged scores, it is first worth noting that the outlier is not due to an averaging of disparate scores; the PESQ predictions for the two trials were 2.25 and 2.51 individually. With further inspection, however, one can see that the same outlier exists for the trained acceptability models and can be seen in fig. 8.11(b). Since these two trials, upon auditioning, do not appear to differ drastically from the pairs of trails with similar SNRs, it seems that this outlier is a case of listener inconsistency. Finally, for validation 2, the PESQ and POLQA scores had very poor correlations with $R = -0.2808$ and $R = -0.1704$ respectively. Here the extended and PEASS based models had correlation $R = 0.45$ and $R = 0.57$, and the OPS had correlation $R = 0.55$.

Since PESQ performed better than POLQA on the training data, and POLQA performed better than PESQ on validation 1, and since both performed very poorly on validation 2, neither model is clearly more appropriate for use in the prediction of acceptability. The OPS scores did not correlate consistently higher than either model, yet they correlated well with validation 2 and so represent a more generalisable
Chapter 8: Building a Model to Predict Acceptability

measure than either PESQ or POLQA. In all but one case, the predictions of the extended acceptability model correlated with the acceptability scores better than PESQ, POLQA, and OPS. If OPS is included in the feature training, however, the PEASS-based acceptability model can be constructed which outperforms all the other predictors for all data sets. Thus the PEASS-based acceptability model had the greatest accuracy and generalisability of all the models tested.

8.8 Summary and conclusion

This chapter began by posing the question “How can the acceptability of auditory interference scenarios featuring a speech target be predicted?” To answer this question several models of acceptability were constructed. In doing so, training and validation data sets were prepared, an objective method for constructing models of acceptability was detailed, and a benchmark model based on a linear regression to SNR was established. An initial model was constructed by selecting features from a large pool produced by analysing internal representations produced by processing the target, interferer, and mixture through the CASP model.

The acceptability models were compared with the benchmark model and all models exceeded the accuracy of the benchmark for the training data. Over a range of validation data the extended model had equal or better correlation with acceptability scores than the benchmark predictions, although the error was higher in some cases. The PEASS based model, however, performed similar to or better than the benchmark in all cases, and was therefore selected as the most accurate and robust of the produced models.

A small adjustment was introduced to all of the models. Since all of the models are based on linear regressions, it is possible for the predicted acceptability scores to exceed 1 or fall below 0, yet such predictions are not meaningful. In such cases, therefore, predictions are capped at 1 or 0.

Finally, the produced model was compared with existing state of the art models of audio and speech quality (POLQA and PESQ), and with the overall preference score produced by the source separation toolkit PEASS. Between PESQ, POLQA, and PEASS a best model could not be easily selected; when sound zone processing was applied the PESQ and POLQA models performed very poorly, for the training data PESQ performed very well, whereas for the validation 1 data set POLQA performed best. In all cases the PEASS-based acceptability model produced predictions with greater correlation to the mean acceptability scores than any of these existing models.

With a model for the prediction of the acceptability of speech in auditory interference scenarios established, it remains only to piece together this work with the models described in the previous chapters to produce an overall strategy for the prediction of
acceptability. In the next chapter this is discussed, along with example applications and notes for practical implementation.
In the previous chapter a model was carefully designed for prediction the acceptability of speech with auditory interference. Prior chapters considered the acceptability of interfering audio programmes more generally. In this chapter, the findings presented in the previous chapters are drawn together to answer the more general research question, “How can the acceptability of listening scenarios featuring auditory interference be predicted?” Subsequently, a discussion is presented of various applications of this research.

9.1 General method of acceptability prediction

Chapter 1 introduced the research question and framed the problem in terms of acceptability, audibility and SNR. In chapter 2 masking was more carefully examined and it was shown that a range of masking phenomena exist, and that while SNR is extremely important there exist a variety of other known factors including temporal and spectral characteristics, loudness, spatial perception, comodulation and informational content. The first three of these factors were considered explicitly in the model training procedure in chapter 8. Spatial information was not considered in this work, since its effect is expected to be smaller than the factors investigated. Comodulation is somewhat considered implicitly within the modulation filterbank of the CASP model. Informational content is particularly difficult to account for, since it is often context dependent, yet some account has been made of this by considering speech-based target programmes separately.

Chapter 3 described a masking and acceptability experiment conducted to gather data about the listening scenarios under consideration, draw some initial audibility boundaries, and investigate any potential relationship between masking and acceptability. A fairly strong positive correlation was found between thresholds of masking and acceptability, and it was therefore desirable to predict masking thresholds in order to indirectly predict acceptability. With such a wide range of factors affecting masking phenomena a physiologically inspired model was selected, in the hope that such a model would generalise well to scenarios on which it was not trained. It was shown that, after some alterations, a model for the prediction of masking thresholds could be re-purposed.
to predict thresholds of acceptability.

In the close of chapter 4 it was noted that the adapted masking threshold prediction model is unlikely to be well suited to predicting the acceptability of listening scenarios featuring speech. This is because spectro-temporally sparse programmes, such as speech, have quiet or silent gaps throughout for which a predicted masking threshold would be equal to the threshold of audibility, yet subjects generally did not require such stringent conditions to report that the listening scenario was acceptable. In addition to this, informal listening revealed that when the target and interferer programme were both speech the listening scenario was very confusing (because of the similar informational content), and the masking threshold based model did not account for this difference in any satisfactory way. Even so, the masking threshold based model predictions were fairly good for the spectro-temporally non-sparse programmes tested, and the model can additionally be used to bound the range of acceptability scores with predictions of SNRs at which either programme would be inaudible.

The work therefore focused on cases wherein the target programme was primarily speech. The previous chapter shows a method for predicting the acceptability of listening scenarios where the target programme was speech based. With both models completed, it is therefore possible to predict the acceptability of a wide range of auditory interference listening scenarios. Figure 9.1 shows an algorithmic approach for doing this.

The algorithm shows that the first stage to making a prediction about acceptability should be to discern whether either programme is inaudible. This step should be performed first because if either programme is inaudible the acceptability score can be determined with no further processing. If the interferer programme is completely masked the acceptability must be equal to 100% by definition, and if the target is completely masked the acceptability must be 0% by definition. If neither programme is found to be inaudible, the next step depends upon whether the target programme is speech based. If the target programme is not primarily speech, the acceptability prediction can be made by using the masking prediction model described in chapter 4. If it is desirable to find the threshold of acceptability the model can be continually operated using the binary search algorithm procedure described in chapter 4. If the target programme is primarily speech based, the speech acceptability model described in chapter 8 can be operated to produce an acceptability score. If the boundaries of acceptability are required the windowed SII model can be operated to give the lower boundary and the masking threshold prediction model can be operated to give the upper boundary. If a threshold of acceptability is required, the model can be operated in a binary search loop as described in chapter 4 until the SNR for the desired threshold is obtained.
Figure 9.1: A flowchart describing the prediction of acceptability for various auditory interference scenarios.
9.2 Assumptions and limitations

The algorithm described in section 9.1 shows how acceptability scores, as well as thresholds and boundaries, can be predicted using the models produced in the previous chapters. The algorithm (and associated models) have certain assumptions and limitations which are detailed here.

Firstly, the algorithm takes, as an input, the target, interferer, and mixed audio programmes. In some applications, the original target and interferer programme may not be available; such problems are known as "blind" problems, and are much more difficult to solve because less information is available to analyse. Blind problems have not been considered in this work, and the approaches investigated in this research are unable to solve this more difficult type of problem in their current form. Blind problems might be solvable, however, by including a front-end which identifies and estimates the target and interferer programmes. The accuracy of such a model might suffer due to compounded errors beginning in the estimation stage, however, and even where such errors are negligible the source identification problem is so difficult that even the human listeners occasionally misattributed programme items in the listening tests (see section 7.1.4).

Secondly, it was noted in section 2.5 that the audibility of target and interferer programmes will sometimes depend upon spatial factors, and the phenomenon of binaural unmasking can be an important aspect of this. As such, it is likely that acceptability will also be affected by spatial factors. This research has, in order to focus on a few key topics, not incorporated stimuli with a wide range of spatial characteristics, but instead has largely focused on mono sources positioned on the frontal axis. As such, the models discussed in this work do not explicitly account for any spatial effects and there may be cases where these effects are important.

Thirdly, the algorithm presented in the previous section requires the identification of speech in the target programme, yet no computational method for determining whether the target programme is primarily speech-based or not has been provided in this work. Literature on this problem often refers to it as 'voice activation detection' (Ramirez et al. 2007), hinting at the true motivation of the work: the detection of the presence or absence of speech within speech communication channels. This problem is different from that of predicting whether a programme musical, speech-based, or of some other type in an arbitrary audio channel. This research problem is significant and, to the best of the authors knowledge, there exist no models which can reliably make such a prediction. In this work these distinctions were made manually, yet a manual prediction method generally limits the application of a model. Here, contextual information (in the form of meta data) can be extremely useful. In a radio broadcast, for example, meta data could be used to encode sections of the show according to whether the
radio host is speaking, music is playing, or whether advertisements are playing. As an example, this type of metadata could be incorporated into the 'typeLabel' variable of the 'AudioTrack' field of the metadata schema described in (Media Information Management 2013). Even more simply, it may be known that a specific radio show features very little speech, or mostly speech, and one mode of acceptability prediction can be used by default. In other cases, such as with telecommunications, this type of information is known a priori: telecommunications will be speech dominated by their very nature. With this type of information the problem of automatic speech detection may often be side-stepped.

Finally, contextual knowledge about the specific application will be relevant in adapting the models for use in applications others than those for which they were trained. The training data is based on listeners instructed to imagine that they “are relaxing at home or in the car”, and it is possible that a listener’s tolerance for interfering audio programmes will differ in other environments. Though it is speculative, it seems reasonable to suggest that listeners in purpose built installations such as cinemas and auditoria, will be highly intolerant of auditory interference because their expectations of the acoustic environment will be very high; in contrast, for lower quality audio environments such as telecommunications or public address in noisy spaces, listeners would likely be far more tolerant of auditory interference. Such contextual differences should be considered carefully when applying the models described in this work to new listening environments. In some cases the model may be adjusted with a linear bias to account for new listener expectations, whereas in other cases the model may need retraining completely with features re-weighted or new features selected entirely. From the research presented, it is not possible to be certain about which applications will require significant retraining of the acceptability model and which applications will require little or no modifications.

9.3 Practical implementation of the acceptability prediction algorithm

The acceptability algorithm shown in fig. 9.1 was constructed to show, as clearly as possible, the linear steps required to predict acceptability. In practice, however, a computational model of acceptability prediction would not be implemented in precisely this way because there is considerable redundancy in the structure. In addition, there may be some applications where prediction is time critical. In the extreme case, predictions may be in, or even faster than, real-time. The proposed algorithm, using the existing implementations of these computational models, generally runs at the order of 5–10 times slower than real-time (for an Intel Core i3 laptop running non-optimised code). The problem, however, can be addressed via a host of strategies for realising a
fast and parsimonious solution.

Firstly, the prediction of the audibility of the target and interferer programme as well as the masking-based prediction of acceptability are all calculated using the CASP model, with minor modifications appropriate to each stage. For a practical implementation, therefore, it would be most appropriate to integrate all three of these stages into a single step.

Secondly, it was pointed out in section 9.1 that the initial prediction of target and interferer audibility can render further processing unnecessary because an inaudible interferer is equivalent to 100% acceptability whereas an inaudible target is equivalent to 0% acceptability. In such cases, no further processing is required thus naturally reducing the time required to complete processing. Since these circumstances are rare, this processing saving will also be rare and, more importantly, unpredictable. In time critical implementations it is likely to be important that predictions can be made both swiftly and within a known duration; since it is unclear in advance whether further processing would be required, this processing saving has limited value in such cases.

Thirdly, acceptability does not generally increase linearly from the lower boundary to the upper boundary. The results of the experiment conducted in chapter 5 for example showed that acceptability generally does not exceed 0 for SNRs below zero, and the results of the experiment conducted in chapter 7 showed that acceptability is generally below 20% for SNRs below 10 dB. At the other end of the scale, for SNRs over 30 dB the acceptability was always rated above 70%. A consistent time saving could be made by excluding audibility calculations based on the assumption that one or both programmes will rarely be masked; such an assumption could be valid depending on the context of the application. A related alteration which does not require excluding audibility predictions altogether, would be to calculate the SNR and use the previously mentioned thresholds, instead of audibility calculation, to produce preliminary acceptability scores. If the preliminary score is not equal to 0% or 100% the ordinary processing algorithm can be continued.

Finally, where there is an extreme dearth of time or processing power available to make predictions, a large and consistent processor saving can be made by implementing linear or logistic fits to SNR, as discussed in chapters 5 and 7, albeit at the cost of reduced accuracy.

9.4 Research applications

The POSZ project was originally conceptualized with consideration to the specific application of generating and evaluating sound-zoning systems in automotive and domestic environments. The research has been conducted such that assumptions about the types of programme material were made with reference to these environments.
Chapter 9: Predicting Acceptability

The range of programmes, however, consumed in these environments is extremely diverse, and this does little to limit the applicability of these approaches to predicting acceptability in other domains.

Some other potentially interesting applications of these methods of the prediction of acceptability include:

- evaluating sound-zoning systems in public or professional environments,
- designing and evaluating zone-based public address systems,
- designing and evaluating auditory interference in acoustic spaces, and
- evaluating source separation algorithms.

The following discusses the potential application of the research presented in the previous chapters to each of these domains.

9.4.1 Evaluating sound-zoning systems

One of the stated goals of this project was to produce a computational model to evaluate sound-zoning systems, yet the context of the listening environment can play an important role in perception.

Automotive and domestic environments

This application was the original purpose of this research, and it is not surprising, therefore, that the models for predicting acceptability can be directly applied for this purpose. Using a bank of test stimuli recorded after processing through a sound-zoning method (as was done for the validation set described in section 7.2), a sound-zoning system can be evaluated in terms of the average acceptability score produced across a range of stimuli, and these average scores could be compared among various implementations of sound-zoning systems to discern which systems perform best.

This application is useful because it greatly increases the rate at which sound-zoning systems can be improved. Instead of running a costly and lengthy listening test for every minor adjustment to or new implementation of a sound-zoning system, the acceptability prediction model can be run. The time required for evaluating sound-zoning systems is reduced from an order of days or weeks (depending on the number of participants and stimuli), to an order of minutes or hours (depending on the number of stimuli).

A caution worth noting with all objective model of prediction, however, is that in the absence of a perfect model only listening tests can provide a truly definitive evaluation. This does not, however, render objective models obsolete, but rather contextualizes their output as indicative. As a result, objective models are extremely useful for
speeding up cycles of design and implementation, but a listening test is usually worth conducting before final production to confirm predictions.

While the use of an objective model to evaluate and compare sound-zoning methods can greatly improve the rate of the design cycle, an even more drastic improvement could be obtained by building the acceptability prediction model directly into the sound zoning filter construction process. This would be done in such a way that the features of the acceptability prediction model instruct the selection of parameters in the sound-zoning model. For example, a search algorithm could be implemented to optimise the filter weightings across loudspeakers based on the known coefficients of the features of acceptability. The search algorithm is likely to be very complex because the relationship between each loudspeaker filter weight and the acceptability features is not obvious and is likely to be confounded with the effects of other loudspeaker filter weights; yet such a search algorithm is possible and would inevitably lead to improvements in acceptability.

Further improvements to long-term acceptability could be achieved by building an acceptability prediction model directly into a real-time “active” sound zoning system (i.e., a system which adjusts its parameters to maximise acceptability for changing programme items). If the system replay buffer is sufficiently large the acceptability model could notice upcoming events, such as intense transients, and suggest temporary adjustments to the active sound-zoning system. For example, if the compression of a programme, to reduce a transient, only mildly reduces target quality for the listener but greatly diminishes the probability that the listening scenario will be unacceptable for the listener in the alternate zone, then this constitutes a good option for an active sound-zoning system.

Public or professional environments

Controlling the interference of auditory programmes could be useful in public or professional environments. For example, some work suggests that even low level background noise, such as those in call-centres or aircraft, could be stressful and damaging to cognitive processes (Trimmel et al. 2012). In such environments, however, it is likely that listeners already have much lower expectations about the listening scenario, and so may be much more tolerant of interference. This would have the effect of pushing acceptability scores up, relatively to the data gathered in the experiments described in previous chapters. As a result, the acceptability models may need to be recalibrated to suit the change in listener expectations.

9.4.2 Designing and evaluating zone-based public address systems

Many public address systems are used in large public spaces and comprise of a large number of loudspeakers, all of which deliver the same message. In some cases this is desirable, such as when giving a general message applying to all members of the
public, while in other cases it may be more appropriate to give more targeted messages to specific zones. For example, messages describing changes to the scheduled arrival of a train on one platform may not be desirable on every platform, or messages describing the exhibits at a museum may be best provided in a zone specific manner. A system which gives selective public address messages like this should be designed with consideration of the overlap between messages in different zones; here a model of the prediction of acceptability could be of use.

Similarly, smaller electroacoustic public systems might also benefit from quantifying the acceptability of auditory interference. For example, it might be preferable if messages produced from ticket booths at a train station were not audible when standing in front of adjacent ticket booths, however when total inaudibility is not practicable, it may be sufficient to design the loudspeaker system such that the level of interference is within a certain threshold of acceptability. Likewise on passenger aircraft, earphones are often provided for every passenger to facilitate the use of individual entertainment systems; it would likely be preferable if earphones were unnecessary and a sound-zoning system was implemented to provide personal audio for each passenger. The degree of interfering audio from adjacent entertainment systems would be considered, as well as the degree of masking provided by the aircraft noise. Again a model of the prediction of acceptability could be implemented to aid the evaluation, and therefore design, of such systems.

9.4.3 Evaluating and designing acoustic spaces

In public spaces another opportunity for the application of a model of the prediction of acceptability would be in the design of acoustic spaces. In shopping centres, cafeterias, and public transport stations an open-plan design is common or in some cases unavoidable. Such designs have notoriously poor acoustic environments, because without walls or plenty of acoustic absorption the level of background noise, especially of speech, tends to be high. When designing such spaces acoustics consultants are likely to consider the background noise levels, reverberation time, and the speech transmission index of Houtgast and Steeneken (1971) (a useful predictor of intelligibility). While each of these describes a facet of the overall acceptability of the acoustic environment, a more comprehensive approach could be taken by using a model of the prediction of acceptability. The models of acceptability and intelligibility described in this work likely be appropriate for predicting the acoustic environment of such designed spaces, although they would likely require some recalibration, and may require additional features for considering reverberation.
9.5 Summary and conclusions

At the start of this chapter the following question was posed: "How can the acceptability of listening scenarios featuring auditory interference be predicted?" In this chapter the question was answered by proposing an algorithm which draws upon the models of prediction discussed in previous chapters. The approach involves identifying whether target or interferer programmes are masked before, implementing an acceptability models based on whether the target programme is speech-based. Assumptions and limitations of the algorithm (and associated models) were discussed, specifically highlighting the lack of consideration of spatial and contextual cues. Some methods for practical implementation were discussed and it was noted that these are likely to depend upon the degree of computational power available and the time required in which to make predictions. Finally, some practical applications of the acceptability prediction model were discussed, noting that besides the evaluation of sound zoning systems, this type of acceptability model could be used to design and evaluate the acoustics of open spaces and public address systems.

In the next chapter, the thesis is concluded by restating the general research questions and findings presented throughout the previous chapters of this thesis.
Chapter 10 Conclusion

This chapter summarises the research described in this thesis. The primary findings of the experiments are outlined, and the scope, limitations, and implications of these are discussed in the wider context of the field. Subsequently, the future work expected to be most consequential is proposed. Finally, the novel contributions to the field are highlighted.

Research questions were posed at the start of each chapter of this thesis outlining the goals of the work. In sequence, these questions are milestones underpinning the course of investigations conducted in this work, and they are restated here to provide a summary of the work conducted.

In chapter 1 the concept of sound zones was introduced, and it was noted that, while SNR was clearly an important measure, auditory masking and the perception of speech might play special roles in the perception of the auditory interference scenarios produced. Chapter 2 therefore began with an investigation of auditory masking by asking two questions: “what are the factors which determine whether an auditory stimulus will be masked by the presence of a second stimulus?”, and “what is the relevance and importance of each factor for sound zones?”. A review of the literature of auditory masking described simultaneous, forward, and backward masking, as being contingent upon absolute level, SNR, frequency spectrum, and relative onset and offset times. These were the most relevant and important factors for masking within sound zones since their effects are so large. Binaural unmasking, stimulus uncertainty, and CMR were also described, and it was shown that masking is also mediated by these underlying factors, but that they were less important to the work because they were likely to be less prevalent. Chapter 3 then aimed to investigate acceptability within auditory interference scenarios, and its relation to masking with two research questions posed: “what is the range of SNRs over which acceptability primarily varies?”, and “is there a relationship between masking acceptability?”. A pair of experiments were conducted to answer these questions and most acceptability thresholds fell within the range 0–35 dB SNR. A linear correlation was found between average masking and acceptability, and it was clear that predicting acceptability using auditory masking would be possible. Since for practical applications the masking data is likely to be unavailable, chapter 4 then set out to investigate existing models for the prediction of auditory masking, with a view to using predictions of masking to predict acceptability.
Chapter 10: Conclusion

The research question: “how can auditory masking be predicted?”, was answered by comparing a variety of auditory masking models. For this work the CASP model was selected and implemented with small modifications to convert the probability of audibility into a masking threshold and to interpret time-windowed long duration programmes.

It was noted in chapter 3 that acceptability scores may differ when both target and interferer programmes are speech-based, and the implementation of the CASP model would be unlikely to properly account for such differences; chapter 5 therefore began focusing on speech intelligibility by posing the research question “what relationships exist between intelligibility, acceptability, and other relevant measures?” The question was addressed by conducting experiments gathering data of intelligibility, acceptability, masking, and distraction. The findings revealed that acceptability and distraction were not simply inverted measures of the same subjective quantity, and also that intelligibility could be used to mark a lower boundary of acceptability which could be used to broadly describe the range over which acceptability mostly varied. Since intelligibility would therefore be a useful measure to describe the perception of auditory interference scenarios, chapter 6 posed the research question: “how can the intelligibility of speech within auditory interference scenarios be predicted?” To answer this question, models for the prediction of intelligibility were compared and evaluated with the data gathered in the previous chapter. A time-windowed implementation of the SII was found to produce the most accurate predictions, and could be used to describe the limits and range of acceptability in auditory interference scenarios.

Based on the work of the previous chapters it was therefore possible to predict the range and limits of acceptability scores for scenarios featuring speech as a target programme, but not to predict acceptability directly. Chapter 7 thus posed the research question: “how can the acceptability of auditory interference scenarios featuring a speech target be predicted?” Additionally, the results of the experiments conducted in chapter 5 indicated that subjects might produce acceptability scores which were bimodally distributed, and the following research question was posed: “what is the general distribution of listener acceptability responses?”. To consider these questions two experiments were conducted: one to gather a wide variety of acceptability data, with many subjects, to investigate whether data was bimodal and to obtain training data for a model of the prediction of acceptability with a speech target programme, and a second to produce a small validation set in which stimuli had been processed using a sound zoning system, in order to validate the trained model. The results of the first experiment showed that acceptability scores were distributed unimodally across subjects. In chapter 8 all the acceptability data were used to train and validate models for the acceptability of speech with interfering audio programmes, and a model was constructed and evaluated. The proposed model at the end of this chapter
can therefore be used to predict the acceptability of auditory interference scenarios featuring a speech target. Finally, in chapter 9, the work from previous chapters was drawn together to answer the more general research question “how can the acceptability of listening scenarios featuring auditory interference be predicted?” The proposed algorithm utilises a masking prediction model to determine whether both the target and interferer programme are audible; if the target or interferer programme is inaudible, the acceptability must be equal to 0 or 100 respectively. If both programmes are audible, one path is taken if the target programme contains speech, and another is taken otherwise. In the latter case, the boundaries of acceptability can be predicted using the modified CASP model, and acceptability can be predicted by additionally using the linear regression model described in chapter 4. In the former case, the lower acceptability boundary may be determined using the time-windowed implementation of the SII, the upper acceptability boundary may be determined using the modified CASP model, and the acceptability score may be determined using the model described in chapter 8.

This brief overview of the work conducted is expanded in more depth over the following sections 10.1 to 10.5.

10.1 Framing the problem

In chapter 1 the concept of sound zoning, as the diminishment of auditory interference for a listener positioned within a physical zone, was introduced. The most fundamental aspect of the problem was considered to be driven by SNR. This suggestion was made because for a sufficiently high SNR the interferer would be inaudible and for a sufficiently low SNR the target would be inaudible. On this basis, a framework was suggested with endpoints capped at these two inaudible markers. In the range between these two endpoints there would likely be a continuum of better and worse listening scenarios, and it was considered reasonable to suggest that there would be some SNR below which a listener would consider the listening scenario to be unacceptable for consumption at home or in an automobile. For populations of listeners, one could map this range with a metric called ‘acceptability’ which describes the probability that a listener picked at random would find a listening scenario to be acceptable.

Mapping out the boundaries of this range, and predicting the location within it therefore constitutes the focus of this work. In order to begin mapping the boundaries of the range of acceptabilities, it is first necessary to understand something about audibility and auditory masking.
10.2 Auditory masking in auditory interference scenarios, and the use of masking prediction for the prediction of acceptability

Chapter 2 reviewed the literature on the phenomena relating to auditory masking. This included discussion about the temporal, spectral, and spatial aspects of masking, as well as the phenomena known as ‘informational masking’ and combodulation masking release. Chapter 3 described a pair of listening tests to gather acceptability and masking data. By using the masking data directly, one can draw general conclusions about the SNRs which mark the upper and lower boundaries of the acceptability space. Using a linear regression, however, the masking threshold data was also used to predict acceptability thresholds. A good correlation was found between masking and acceptability for the test data and three quarters of the variance was explained by the model.

The model required masking thresholds as an input, however, and for practical implementation this will often not be known. To avoid this problem, chapter 4 considered existing models for the prediction of masking thresholds, with a view to predicting both acceptability and the boundaries of audibility for the target and interferer programme. A review of models for the prediction of masking thresholds revealed that the CASP model was a good starting point for a model of the prediction of acceptability. Modifications to the implementation of the CASP model were made to allow predictions to be made for the data presented in the previous chapter. Prediction accuracy was good with RMSE < 3 dB, and the model was further customised to use this data to predict acceptability. The predictions had correlation with the acceptability data equal to the correlation between the masking and acceptability data; so the model performed well.

Some problems, however, were still present within this framework of predicting acceptability. Firstly, subjects had commented that when the target and interferer programmes were both speech the listening scenario was generally less acceptable, however the model does not consider this. Secondly, when a target programme is speech (without additional background programmes) the silent gaps are liable to render the notion of a masking threshold invalid, i.e. the masking threshold of the interferer programme is equivalent to the threshold of audibility, however this does not seem to agree with subjective experience of auditory interference scenarios (e.g. some, audible but low, level of interferer is often acceptable). It was therefore considered necessary to focus on cases where the target programme was entirely, or primarily, speech.
10.3 Speech intelligibility and acceptability in auditory interference scenarios

Due to the considerations raised at the end of chapter 4 about the special cases of speech target programmes, chapter 5 focused on cases where the target was based on speech. The general goals of speech communication were considered, and while characteristics such as the timbral quality of the speech may be important, it was considered that the principal criterion upon which almost any speech programme would be assessed was intelligibility: the capacity for the intended communicable information to be received and appropriately comprehended. In order to investigate any relationship between acceptability in auditory interference scenarios and speech intelligibility, an experiment was conducted gathering data on acceptability and intelligibility. In addition data regarding masking and distraction were also obtained.

The experiment showed that intelligibility and acceptability data did not linearly correlate. Acceptability did not, however, exceed zero until speech intelligibility was near perfect, so intelligibility could be used to draw finer boundaries in the acceptability space when the target programme contains speech. Distraction in cases involving a speech target was shown to be almost entirely explained by a logistic fit to SNR. A trend was found between the masking data and the acceptability data for only half of the listeners, which seems to support the suggestion that masking is less useful for predicting acceptability when the target programme is speech.

In chapter 6 some models for the prediction of speech intelligibility were investigated and compared. Various models were tested to find how accurately predictions could be made of the intelligibility data obtained from the experiment described in chapter 5. The most accurate predictions were produced by a modification to the simplest model: the SII. By time windowing, rather than considering only the long-term SII, accurate predictions were obtained for the intelligibility of the target programme. The adapted SII therefore can be used to mark the lower boundary of the acceptability space for auditory interference scenarios including target speech. Additionally, where the interferer features speech it is possible to determine most of the range of acceptability scores with the interferer intelligibility, although the absolute upper boundary of acceptability may not be marked in this way because audible, yet unintelligible, interferers may still diminish acceptability to some degree.

With the boundaries of the acceptability space for scenarios featuring target speech determined, it remained only to model acceptability within these boundaries more clearly.
Chapter 10: Conclusion

10.4 A model for predicting the acceptability of speech in auditory interference scenarios

To build a model predicting the acceptability of auditory interference and sound zoning scenarios featuring target speech, two more experiments were carried out: one to gather a large training corpus, and a small experiment to gather validation data using a sound zoning system. Features were generated and a rigorous model training procedure, optimising for accuracy and robustness while considering model parsimony, was carried out. The resulting model predicted acceptability more accurately than a benchmark model based on a linear regression to SNR for the training, cross validation, and both validation data sets. The model consisted of five features: one based on SNR, one based on the PEASS model, and three based on analysis of the internal representations in the CASP.

10.5 A method for predicting acceptability

With a model for predicting the acceptability of speech complete, in chapter 9 the various strands of research were brought together to prescribe a method for predicting acceptability more generally. An algorithm was outlined showing how acceptability could be predicted for arbitrary auditory interference scenarios using predictions of audibility and intelligibility, as well as the model for predicting the acceptability of speech depending on the target programme content. Details were given regarding a practical implementation of the model and suggestions were made for ways to reduce the processing time required for predictions. Finally a discussion was provided of applications for the model to real world problems noting where modifications or retraining would be necessary. These included considerations of evaluation and real-time optimisation of sound-zoning systems, as well as evaluating the acceptability of auditory interference in public and professional spaces, or aiding in the design of personal audio systems for aeroplanes or at transport stations.

10.6 Scope, limitations, and further work

The first limitation to the work was identified in chapter 1 when the problem was initially framed: a distinction was made between effects caused by the presence of an interferer and effects caused by degradations to target quality. It was decided that the presence of an interferer would ordinarily be more instrumental to the degradation of the acceptability of the listening scenario than target quality degradations, although there may be some degree to which these degradations are fungible. At any rate, the
effect of the interferer is more fundamental to the problem of sound zones and was a reasonable place to start. In practical applications of sound zoning systems, some degree of target quality degradation is rarely avoidable because a system controlling a sound field to maximise contrast between spatial zones will inherently modify both programmes reproduced. This work has focused on the effects of auditory interference for a target programme with no quality degradations (although the second validation data set in chapter 8 did include stimuli produced using a physical implementation of a sound zoning system), so the effect of target quality degradations, and any interactions between target quality degradations and the presence of an interferer is largely unknown and represents an interesting field for future study. Specifically, it would be very useful to obtain curves describing the relative importance of target quality degradations to acceptability for various interferer levels; this would allow a more sophisticated acceptability model to be produced, and in turn sound zoning systems could be carefully calibrated to optimise these parameters.

Another limitation of the work is the possibility of spatial effects. It was shown in section 2.5 that binaural unmasking can, in the worst circumstances, have an effect on masking thresholds by as much as 8-12 dB. Some effect would be expected in sound zoning systems which produce some perception of spatial separation, so when comparing between sound zoning methods with different degrees of spatial separation this effect is likely to decrease the accuracy of the current acceptability model. The second validation set to the acceptability model was such a demanding task in part because the scores were based on binaural recordings of a sound zoning method which produced a pronounced spatial separation. Despite this the predictions were still somewhat appropriate, although it cannot be ruled out that spatial characteristics will sometimes be important to acceptability. For this reason, one area of future work would involve the quantification of the effect of spatial characteristics of programmes upon acceptability. The scope of the acceptability model produced in chapter 8 is strictly limited to monaural co-located target and interferer programmes, although the validation demonstrated that approximate predictions can be made for radically different spatial scenarios.

10.7 Novel contributions to the field

Throughout the course of the research a number of novel contributions to the field were made. The first point of original research was carried out after a review of the masking literature, and consisted of an experiment into masking thresholds and thresholds of acceptability. A linear regression between the data sets had a good fit, and thus a relationship was found between audibility and acceptability. Following this an investigation into the computational prediction of masking thresholds was carried out,
Chapter 10: Conclusion

and the CASP model was selected for implementation. Since in its original form the model was used only to test very short, and highly contrived stimuli, some modifications were made to the CASP model to adapt it for use with complex, ecologically valid stimuli. By incorporating an adaptation to the aforementioned linear regression model, the CASP model was adapted to make predictions of acceptability.

The following research focused on speech programmes and, after carrying out further experiments, it was discovered that a very high intelligibility marks the lower boundary of acceptability. In addition it was shown that distraction and acceptability are not the same for scenarios featuring speech targets. Furthermore, for such cases most of the range of acceptability was covered between 0 and 35 dB SNR.

In order to predict the lower boundary of acceptability in practical applications, a comparative analysis of various speech intelligibility models within auditory interference scenarios was conducted and subsequently a model for predicting speech intelligibility was adapted for use within auditory interference scenarios.

To predict acceptability in such scenarios, however, it was necessary to conduct an experiment to gather a large quantity of training data, before training a computational model of prediction. The experiment revealed that listener ratings of acceptability constitute a unimodal distribution, and also that there were no strong effects on acceptability due to musical training or listener age. Additionally, it was determined that most of the variation in acceptability scores occurs between 10 and 30 dB SNR. Subsequently, a feature selection and model training process was conducted and a model for the prediction of the acceptability of speech in auditory interference scenarios was constructed, and shown to be robust to a wide variety of stimuli and listening scenarios.

Finally, the research was drawn together and an algorithm for predicting the acceptability of auditory interference scenarios was proposed, using the previously constructed models.

In summary, the novel contributions to the field are therefore:

1. A linear correlation between masking and acceptability was found, and this relationship can be, using a linear regression model, used to predict acceptability.

2. With minor modifications, the physiologically inspired masking threshold model known as the CASP model, can be used to make predictions of masking thresholds for complex, ecologically valid stimuli.

3. With the additional use of a linear regression model, the CASP model can be used to make predictions of acceptability instead of masking.

4. A high speech intelligibility (around 95%) was found to mark the lower boundary of acceptability.
5. Distraction and acceptability were shown to differ for speech target based auditory interference scenarios.

6. Most acceptability thresholds were found to be in the range 0–35 dB SNR, and most of the variation in acceptability scores occurred between 10 and 30 dB SNR.

7. A comparative analysis of various speech intelligibility models within auditory interference scenarios showed that a time-windowed version of SII gives the best prediction of intelligibility within auditory interference scenarios.

8. Listener ratings of acceptability constitute a unimodal distribution.

9. There were no strong effects on acceptability due to musical training or listener age.

10. It was found that the acceptability of speech in auditory interference scenarios could be predicted for a wide range of stimuli and listening scenarios using a model trained to subjective data.
Acceptability model features

This section outlines the features generated for use building a model of acceptability in chapter 8.

Manual features

As discussed in section 8.3, three manual features were coded according to the presence of speech in the test stimuli. Specifically these were:

1. Speech: coded as a 1 when the interferer contains speech (not including musical vocals), 0 otherwise.
2. Speech Only: coded as a 1 when the interferer contains only speech (e.g. with no background music), 0 otherwise.
3. Instrumental: coded as a 1 when the interferer contains only instrumental music (i.e. does not contain any linguistic content), 0 otherwise.

RMS features

Simplistic features based on the RMS level of the items were also constructed. Since the SNR (and therefore level) of the programmes was considered likely to be important the following features were extracted to estimate such signal properties in a simplistic way.

1. RMS-Target: The RMS level of the target programme.
2. RMS-Interferer: The RMS level of the interferer programme.
3. RMS-SNR: The SNR calculated by dividing the RMS-Target by the RMS-Interferer.
4. RMS-BadFrame50: 15 features calculated by breaking the signals into 50 ms frames and counting the number of frames in which the SNR (in dBs) exceeds a given threshold. The 15 features used thresholds ranging from 0 to 28 dB SNR in steps of 2 dB. The frame length was selected as 50 ms because this is a common frame length in other speech research (based on the average lengths of phonemes).
5. RMS-BadFrame400: A further 15 features calculated by breaking the signals into 400 ms frames and otherwise calculated as RMS-BadFrame50. 400ms was shown to be an good frame length when used in previous work with the CASP model.
6. **RMS-BadFrame10**: A further 15 features calculated by breaking the signals into 10 ms frames and otherwise calculated as RMS-BadFrame50. Since it may be possible to improve the fidelity of this measure by using briefer frames, a 10ms variant is considered.

7. **RMS-BadFrame50-Over50**: 15 features calculated identically to RMS-BadFrame50, however with frames overlapping by 50% (common in other speech models).

8. **RMS-BadFrame100-Over400**: 15 features calculated identically to RMS-BadFrame400, however with frames overlapping by 100ms (used in previous acceptability prediction based on CASP model).

Thus 78 RMS-based features were included in the features pool.

### Zwicker loudness features

Instead of the stimulus levels, the stimulus loudness may offer a useful description of the acceptability of listening scenarios. To that end, the Zwicker loudness model was utilised to generate 10 further features.

1. **ZWICK-Target**: The loudness exceeded during 30 ms of the target programme.
2. **ZWICK-Interferer**: The loudness exceeded during 30 ms of the interferer programme.
3. **ZWICK-Target**: The loudness exceeded during 50% of the target programme.
4. **ZWICK-Interferer**: The loudness exceeded during 50% ms of the interferer programme.
5. **ZWICK-TarMax**: The maximum instantaneous loudness of the target programme.
6. **ZWICK-IntMax**: The maximum instantaneous loudness of the interferer programme.
7. **ZWICK-LoudnessRatio**: The SNR calculated by subtracting ZWICK-Interferer from ZWICK-Target.
8. **ZWICK-LoudnessPeakRatio**: The SNR calculated by subtracting ZWICK-TarMax from ZWICK-IntMax.
9. **ZWICK-TarMax50Ratio**: The SNR calculated by subtracting ZWICK-Interferer from ZWICK-Target.
10. **ZWICK-IntMax50Ratio**: The SNR calculated by subtracting ZWICK-TarMax from ZWICK-IntMax.
Spectral centroid features

Subjects occasionally commented that higher frequency interference was often especially problematic. It was therefore worthwhile to consider spectral centroid features. In total 7 features were included based on the spectral centroid.

1. SPEC-TarMean: The mean spectral centroid of the target programme divided into frames of 2048 samples (approximately 21.5 ms).
2. SPEC-IntMean: The mean spectral centroid of the interferer programme divided into frames of 2048 samples (approximately 21.5 ms).
3. SPEC-TarStd: The standard deviation of the calculated spectral centroids of the target programme.
4. SPEC-IntStd: The standard deviation of the calculated spectral centroids of the interferer programme.
5. SPEC-MeanRatio: SPEC-IntMean subtracted from the SPEC-TarMean (giving an indication of spectral distance).
6. SPEC-StdRatio: SPEC-IntStd subtracted from the SPEC-TarStd (giving an indication of relatively spectral variance).
7. SPEC-LogMeanRatio: the tenth base logarithm of SPEC-IntMean divided by SPEC-TarMean (giving a slightly more perceptually relevant spectral distance metric).

CASP based features

The CASP model was used (excluding the final modulation filterbank stage) to produce internal representations of the target, interferer, and mixed stimuli. From these representations a wide range of features can be derived. The stimuli were divided into 400 ms frames stepping through in 100 ms steps and each frame was processed using the CASP model. Three groups of features were derived from the resulting frames: standard framing (SF), no overlap (NO), and 50 ms no overlap (50MS). SF features are based on the existing framing structure, NO features are based on a signal reconstructed by using only every fourth frame (i.e. as if the stimuli had been processed by CASP in 400 ms non-overlapping frames), and 50MS features are derived using the non-overlapping internal representation signal broken into 50 ms frames.

SF features

A total of 66 SF features were included in the feature pool.

1. SF-TMinLev: The minimum value in the vector produced by summing the energy within frames and across frequencies of the target programme. This describes the 400ms frame with the least energy in the internal representation of the target programme.
2. SF-IMinLev: As above but for the internal representation of the interferer programme.

3. SF-MMinLev: As above but for the internal representation of the mixed stimuli.

4. SF-TMaxLev: Identical to SF-TMinLev, except calculated by taking the maximum value in the vector.

5. SF-IMaxLev: As above but for the internal representation of the interferer programme.

6. SF-MMaxLev: As above but for the internal representation of the mixed stimuli.

7. SF-TStdLev: The standard deviation of the vector produced by summing the energy within frames and across frequencies of the target programme. This describes the variance of energy across 400ms frames of the internal representation of the target programme.

8. SF-IStdLev: As above but for the internal representation of the interferer programme.

9. SF-MStdLev: As above but for the internal representation of the mixed stimuli.

10. SF-TMinSpec: The minimum of the vector produced by summing the energy within frames and across frames of the target programme. This describes the energy of the maximum frequency band.

11. SF-IMinSpec: As above but for the internal representation of the interferer programme.

12. SF-MMinSpec: As above but for the internal representation of the mixed stimuli.

13. SF-TMaxSpec: Identical to SF-TMinSpec except calculated by taking the maximum value in the vector.

14. SF-IMaxSpec: As above but for the internal representation of the interferer programme.

15. SF-MMaxSpec: As above but for the internal representation of the mixed stimuli.

16. SF-TMinF: The number of the frequency channel with the minimum energy. This measure describes a stimulus quality similar to that of the spectral centroid.

17. SF-IMinF: As above but for the internal representation of the interferer programme.

18. SF-MMinF: As above but for the internal representation of the mixed stimuli.

19. SF-TMaxF: Identical to SF-TMinF except calculated by taking the channel number with the maximum energy.

20. SF-IMaxF: As above but for the internal representation of the interferer programme.

21. SF-MMaxF: As above but for the internal representation of the mixed stimuli.
Acceptability model features

23. SF-IRanSpec: As above but for the internal representation of the interferer programme.
24. SF-MRanSpec: As above but for the internal representation of the mixed stimuli.
25. SF-TStdSpec: Calculated by taking the difference between SF-TMaxSpec and SF-TMinSpec.
26. SF-IStdSpec: As above but for the internal representation of the interferer programme.
27. SF-MStdSpec: As above but for the internal representation of the mixed stimuli.
28. SF-TSpec1: The ratio of energy above 1 kHz to that below 1 kHz in the internal representation of the target programme. 1kHz was suggested as a threshold in a paper on brightness referenced in the MirToolBox manual.
29. SF-ISpec1: As above but for the internal representation of the interferer programme.
30. SF-MSpec1: As above but for the internal representation of the mixed stimuli.
31. SF-TSpec3: The ratio of energy above 3 kHz to that below 3 kHz in the internal representation of the target programme. 3kHz was suggested as another useful threshold in a paper on brightness referenced in the MirToolBox manual.
32. SF-ISpec3: As above but for the internal representation of the interferer programme.
33. SF-MSpec3: As above but for the internal representation of the mixed stimuli.
34. SF-SpecFDiff: SF-IMaxF subtracted from SF-TMaxF. This describes the distance between the peak frequencies in the internal representations of the target and interferer.
35. SF-SpecFChange: SF-MMaxF subtracted from SF-TMaxF. This describes the distance between the peak frequencies in the internal representations of the target and the mixture.
36. SF-AbsSpecFDiff: the absolute value of SF-SpecFDiff, thus excluding information describing which programme has the higher frequency (i.e. only the distance between peak frequencies is considered).
37. SF-AbsSpecFChange: the absolute value of SF-SpecFChange, thus excluding information describing which programme has the higher frequency (i.e. only the distance between peak frequencies is considered).
38. SF-MeanXcorr: The mean cross correlation (across frames) is calculated in the same manner as CASP is used to predict masking thresholds (excepting the modulation frequency bands). Specifically, the target and mixture time-frequency units are multiplied (unit by unit) and then summed in both dimensions; this is then divided by the number of elements in the matrix, and the mean of these values (across all frames) constitutes the feature.
39. SF-StdXcorr: Calculated as above except taking the standard deviation.

40. SF-Xcorr5per: Calculated as per SF-MeanXcorr, except instead of taking the mean across frames the value at which only 5% of frames falls below is.

41. SF-Xcorr95per: Calculated above except using a threshold of 95%.

42. SF-MinFrameSNR: the sum of TF units for each frame of the target and interferer internal representations are calculated and the SNR for each frame is then computed.

43. SF-MeanFrameSNR: As above, however the average SNR is used instead.

44. SF-DivFrameMixT: 10 features with various thresholds for the division of the frames by frequency bands internal representation of the target by the equivalent mixture internal representation.

45. SF-DivFrameMixI: 10 features with various thresholds for the division of the frames by frequency bands internal representation of the interferer by the equivalent mixture internal representation.

NO and 50MS features

The same features were calculated for the NO and 50MS conditions where possible, producing close to 200 features.
Spectrograms of stimuli

In this appendix spectrograms are provided showing the frequency content of stimuli used in the experiments discussed in chapter 7.

Figure 10.1 shows the mean and standard deviation of the spectrograms for the target stimuli used in the training data. Since the target stimuli were all speech-based programmes, most of the energy is contained below 4 kHz.

Figure 10.2 shows the mean and standard deviation of the spectrograms for the interferer stimuli used in the training data. Since the interferer stimuli included both music and speech-based programmes there is more energy at higher frequencies.

Figure 10.3 shows the difference between the target and interferer mean spectrograms for the interferer stimuli used in the training data. Since the interferer stimuli included both music and speech-based programmes and the target stimuli included only speech-based programmes the difference score is negative at very high and very low frequencies, but strongly positive between 250 Hz and 4 kHz.
Figure 10.1: Mean and standard deviation magnitude spectra for target programmes (left), and a duplicate showing only frequencies up to 4 kHz (right). Black lines show the mean, with blue and red lines showing upper and lower standard deviations.

Figure 10.2: Mean and standard deviation of the magnitude spectra for interferer programmes (left), and a duplicate showing only frequencies up to 4 kHz (right). Black lines show the mean, with blue and red lines showing upper and lower standard deviations.

Figure 10.3: Difference between the mean magnitude spectra of the target and interferer programmes (left), and a duplicate showing only frequencies up to 4 kHz (right).
Acronyms

ACC  Acoustic Contrast Control
AFC  Alternative Forced Choice
AI   Articulation Index
ANOVA Analysis of Variance
APS  Artefact Perceptual Score
ASR  Automatic Speech Recognition
BMLD Binaural Masking Level Difference
CASP Computational Auditory Signal-processing and Perception
CoRE Component of Relative Entropy
CI   Contralateral Inhibition
CMR  Comodulation Masking Release
CSII Coherence Speech Intelligibility Index
dB   decibels
DRNL Dual Resonance Non Linear
EC   Equalisation Cancellation
EI   Excitation-Inhibition
EM   Energetic Masking
ERB  Equivalent Rectangular Bandwidth
FFT  Fast Fourier Transform
HPF  High Pass Filtered
L1   Increment Level
ILD  Interaural Level Difference
IM   Informational Masking
IPD  Interaural Phase Difference
**IPL** Instantaneous Partial Loudness

**IPS** Interferer Perceptual Score

**IR** Internal Representation

**ITD** Interaural Time Difference

**JND** Just Noticeable Difference

**LPF** Low Pass Filtered

**NBN** Narrow Band of Noise

**OFL** Off Frequency Listening

**OPS** Overall Perceptual Score

**OR** Outlier Ratio

**P_D** Probability of Detection

**PEASS** Perceptual Evaluation methods for Audio Source Separation

**PEASS-OPS** PEASS - Overall Perceptual Score

**P_T** Test Probability

**P_A** Accuracy Probability

**PCA** Principal Component Analysis

**PESQ** Perceptual Evaluation of Sound Quality

**POLQA** Perceptual Objective Listening Quality Assessment

**POSZ** Perceptually Optimised Sound Zone

**PSM** Perceptual Similarity Measure

**QoLE** Quality of Listening Experience

**QoLE_I** the proportion of QoLE attributed to the effect of the interference

**QoLE_T** the proportion of QoLE attributed to the effect of the target

**R** Pearson’s Correlation Coefficient

**RC** Reverse Correlation

**RMS** Root Mean Square

**RMSE** Root Mean Squared Error

**RMSE^*** Epsilon-insensitive Root Mean Squared Error

**SII** Speech Intelligibility Index

**SNR** Signal to Noise Ratio
**Acronyms**

**SPL**  Sound Pressure Level  
**SRT**  Speech Reception Threshold  
**SSN**  Speech Shaped Noise  
**STOI**  Short-Time Objective Intelligibility  
**STPL**  Sound Term Partial Loudness  
**T/FMLM**  Time/Frequency Multi-Look Model  
**TIR**  Target to Interferer Ratio  
**L_T**  Test Level  
**TPS**  Target Perceptual Score  
**VIF**  Variance Inflation Factor
Bibliography

25th Frame Box Office figures (2013).
URL: http://www.25thframe.co.uk/charts/top-grossing-movies-uk.php


Bartus, P. (2013), 'Akustyk'.
URL: http://bartus.org/akustyk/


URL: http://www.bbc.co.uk/radio1/playlist/


Benjamin, E. and Crockett, B. (2005), 'Preferred listening levels in the automotive environment', *119th AES Convention paper 6533* pp. 1-10.


Boersma, P. and Weenik, D. (2013), 'Praat: doing phonetics by computer [Computer program]'.
URL: www.praat.org


Bibliography


Classic FM Ultimate Hall of Fame (2013).
URL: http://ultimatehalloffame.classicfm.co.uk/


URL: http://www.ncbi.nlm.nih.gov/pubmed/8655793


URL: http://www.engr.uky.edu/~donohue/audio/Arrays/MAToolbox.htm


Feldtkeller, E. and Zwicker, R. (1956), *Das Ohr als Nachrichtenempfänger*, Hirzel Stuttgart.


Glass, G., Peckham, P. and Sanders, J. (1972), ‘Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance’, *Review of educational research* **42**(3).

URL: http://link.aip.org/link/?JAS/114/349/1


---

283


Nagel, F. (2013), ‘Spectral centroid matlab code’. URL: [http://www.mathworks.co.uk/matlabcentral/fileexchange/11345-spectralcentroid/content/centroid.m](http://www.mathworks.co.uk/matlabcentral/fileexchange/11345-spectralcentroid/content/centroid.m)


**URL:** [http://www.rajar.co.uk/](http://www.rajar.co.uk/)


URL: http://www.isr.umd.edu/Labs/NSL/Cosyne/masking_counterpoint.htm


URL: http://www.isr.umd.edu/Labs/NSL/Cosyne/Yost.htm


