Towards measuring music mix quality: the factors contributing to the spectral clarity of single sounds

Kirsten Hermes

Submitted for the degree of Doctor of Philosophy

Institute of Sound Recording

Faculty of Arts and Social Sciences

University of Surrey

2017 © Kirsten Hermes
Abstract

Mixing music is the process of combining tracks of recorded audio to an overall piece. This is a complicated process and, hence, automatic mixing or metering tools would be useful. The aim of the current research project was to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes (spectral clarity).

A review of academic and non-academic literature revealed that the high-level parameters that are responsible for determining the perceived quality of a music mix are ‘clarity and separation’, ‘balance’, ‘impact and interest’ and ‘freedom from technical faults’, alongside context-specific parameters. A further in-depth literature review established that clarity and separation—the chosen focus for this research—depend on spectral, spatial and intensity factors, and temporal changes in these factors. Spectral factors play an important role across all areas of literature consulted (namely timbral clarity, clarity in concert halls, masking, loudness, auditory scene analysis and speech intelligibility), and so the impact of mix EQ on spectral clarity was investigated in a series of experiments.

These experiments determined that two important factors contribute to the spectral clarity of single sounds. These are the harmonic centroid (spectral centroid divided by the sound’s average fundamental frequency) and mid-range spectral peakiness (related to sharp peaks in the frequency spectrum). For sounds modified by simple spectral filtering, these two factors are sufficient to model clarity changes with a Spearman correlation ranging from 0.631 (bass and vocal stimuli) to 0.848 (string stimuli). For sounds in a mix, however, other factors become important. Adding a peak audibility measure proved useful. This measure determined whether the audibility of peaks in the spectra of the target sounds was increased or decreased through EQ. Target and overall mix harmonic centroids and mid-range spectral peakiness, combined with peak audibility, correlated positively with target spectral clarity (r=0.568).

Findings could contribute to the development of marketable products such as a piece of software able to judge the overall sound quality of a mix, automatic mixers or sonically improved music production software. Further work will allow a more comprehensive and generalizable model to be developed.
Declaration 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.
# Table of contents

1 Introduction.......................................................................................................................... 1

1.1 Methodology and structure ............................................................................................ 1

1.2 Conclusion........................................................................................................................ 4

2 High-level descriptive quality criteria and lower-level perceptual components of music mixes .................................................................................................................................................. 6

2.1 General comments on the literature.............................................................................. 6

2.1.1 Academic sources on mix quality .............................................................................. 6

2.1.2 Non-academic sources on mix quality ...................................................................... 7

2.2 Mix quality parameters according to the sources consulted ........................................... 7

2.2.1 Clarity and Separation ............................................................................................. 10

2.2.2 Balance .................................................................................................................... 12

2.2.3 Impact/Interest ........................................................................................................ 16

2.2.4 Appropriate context-specific characteristics ......................................................... 18

2.2.5 Freedom from technical faults ................................................................................ 20

2.3 Is there any consensus on what is important in mix quality? ........................................ 21

2.4 Discussion....................................................................................................................... 22

2.5 Conclusion and next steps ............................................................................................. 23

3 Relating psychoacoustic findings to clarity and separation in music mixes .................... 24

3.1 The timbral clarity of single sounds.............................................................................. 24

3.1.1 Defining timbre and clarity ...................................................................................... 25

3.1.2 Acoustic factors affecting timbral clarity .............................................................. 25

3.1.3 Spacing of harmonics ............................................................................................ 28

3.1.4 Conclusion ................................................................................................................ 28

3.2 Relating clarity in concert halls to clarity in mixes ......................................................... 29

3.2.1 What is clarity and what sonic factors affect it? ..................................................... 29

3.2.2 Echoes .................................................................................................................... 30

3.2.3 Noise ....................................................................................................................... 30
3.2.4 Measurement & the ISO standard .......................................................................................... 31
3.2.5 Possible limitations of measurement methods ....................................................................... 32
3.2.6 Are there any perceptual optima of clarity? ........................................................................ 33
3.2.7 Conclusion ............................................................................................................................ 35

3.3 Relating Auditory Scene Analysis to separation in mixes ....................................................... 35

3.3.1 What is auditory scene analysis? ......................................................................................... 36
3.3.2 Primitive Grouping ............................................................................................................... 37
3.3.3 Gestalt principles .................................................................................................................. 37
3.3.4 Sequential grouping ............................................................................................................. 41
3.3.5 Spectral grouping .................................................................................................................. 43
3.3.6 Schema-based grouping ...................................................................................................... 47
3.3.7 The influence of Attention on Auditory scene analysis ....................................................... 48
3.3.8 Auditory stream segregation in music and mixes ............................................................... 49
3.3.9 Conclusion ............................................................................................................................ 50

3.4 Relating masking to clarity in mixes ....................................................................................... 51

3.4.1 Definition of masking .......................................................................................................... 52
3.4.2 Spectrum related factors influencing masking ..................................................................... 52
3.4.3 Temporal factors influencing masking ................................................................................ 56
3.4.4 Spatial factors ....................................................................................................................... 59
3.4.5 Masking in mixes ................................................................................................................ 61
3.4.6 Conclusion ............................................................................................................................ 62

3.5 Relating factors for loudness to clarity in mixes .................................................................... 63

3.5.1 What is loudness? ................................................................................................................. 63
3.5.2 Measuring loudness .............................................................................................................. 64
3.5.3 Factors influencing loudness ............................................................................................... 66
3.5.4 Relation to clarity and separation in mixes ......................................................................... 72
3.5.5 Conclusion ............................................................................................................................ 73

3.6 Relating speech intelligibility to clarity in mixes ................................................................... 73

3.6.1 Acoustic attributes of speech ............................................................................................ 74
4.3 Further aims of the pilot study

4.3.1 Are the chosen stimulus pairs useful and do they have any visible impact on the clarity ratings?

4.3.2 Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

4.4 Conclusion

4.4.1 Does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity?

4.4.2 Are the chosen stimulus pairs useful and do they have any visible impact on the clarity rating?

4.4.3 Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

5 Assessing the contribution of different octave bands to the single sound spectral clarity of piano and guitar stimuli

5.1 The setup

5.1.1 Stimuli

5.1.2 The choice of additional attributes

5.1.3 User interface

5.1.4 The presentation of the stimuli to test subjects

5.1.5 Type of listeners

5.2 Presentation of the results

5.2.1 Prerequisites for using parametric methods

5.2.2 The contribution of octave bands to clarity

5.3 Relation to the current literature and implications for future listening tests

5.4 Conclusion

6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

6.1 The setup for the new listening test

6.1.1 Stimuli

6.1.2 The presentation of the stimuli to test subjects
6.2 Presentation of the results ................................................................. 142
  6.2.1 Prerequisites for using parametric methods .................................. 142
  6.2.2 The impact of octave bands on clarity ........................................ 146
  6.2.3 Spectral centroids ...................................................................... 148
  6.2.4 Harmonics Boosted and cut ....................................................... 155
  6.2.5 Harmonic centroid .................................................................... 157
6.3 Clarity predictors for guitar and piano stimuli .................................... 160
6.4 Relation to the literature and implications for further listening tests ....... 164
6.5 Conclusion ....................................................................................... 165
7 The single sound spectral clarity of vocal and bass stimuli ..................... 167
  7.1 The setup for the listening test ....................................................... 167
    7.1.1 Stimulus generation ................................................................. 169
  7.2 Presentation of the results ............................................................. 170
    7.2.1 Prerequisites for using parametric methods .............................. 170
    7.2.2 Centroid deviations .............................................................. 172
    7.2.3 Harmonics boosted and cut .................................................. 177
  7.3 Discussion ...................................................................................... 179
  7.4 Conclusion ..................................................................................... 180
8 Single sound clarity adjustment task ....................................................... 182
  8.1 The setup for the new listening test ................................................. 182
    8.1.1 EQ controls ............................................................................ 184
    8.1.2 Loudness matching ............................................................... 185
    8.1.3 Rating and text box ............................................................... 185
    8.1.4 Additional attributes ............................................................ 186
    8.1.5 Pilot Study ............................................................................ 186
    8.1.6 Stimuli ................................................................................. 186
    8.1.7 The presentation of the stimuli to test subjects ....................... 191
    8.1.8 Type of listeners ..................................................................... 191
8.2 Results and analysis ................................................................. 191
8.2.1 Requirements for parametric statistics .................................... 191
8.2.2 Ratings data ........................................................................ 192
8.2.3 What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ? ................................................................. 194
8.2.4 Approach to analyzing the qualitative data ................................ 202
8.2.5 How can spectral clarity be defined in the context of EQ adjustments? .... 203
8.2.6 What additional factors are likely to be relevant for spectral clarity? .......... 205
8.3 Discussion ........................................................................... 210
8.4 Conclusion ............................................................................ 212
8.4.1 What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ? ................................................................. 212
8.4.2 What additional factors are likely to be relevant for spectral clarity? .......... 213
8.4.3 How can spectral clarity be defined in the context of EQ adjustments? .......... 213
9 An improved spectral clarity predictor ........................................ 214
9.1 A computational model of mid-range spectral peakiness .................. 214
9.1.1 Devising the model ................................................................ 214
9.1.2 Incorporating mid-range spectral peakiness into the clarity model ............ 221
9.2 Conclusion ............................................................................ 225
10 Single sound spectral clarity in mixes ......................................... 226
10.1 The setup for the listening test ................................................... 226
10.1.1 Stimuli ............................................................................ 228
10.1.2 Type of listeners .................................................................... 230
10.2 Presentation of the results ....................................................... 230
10.2.1 Prerequisites for using parametric methods .............................. 230
10.2.2 Comparing clarity and separation ......................................... 232
10.2.3 Target harmonic centroids and mid-range spectral peakiness ............... 233
10.2.4 Audible target harmonic centroids and mid-range spectral peakiness ....... 237
10.2.5 Overall target-to-backing track ratio ...................................... 242
10.2.6 The relationship between target and backing track spectra ........................................... 244
10.2.7 Peak audibility .................................................................................................................. 249
10.2.8 Overall mix spectral centroids ....................................................................................... 252
10.2.9 Peak audibility and overall harmonic centroid ............................................................. 254
10.2.10 Target and overall mix harmonic centroid and mid-range spectral peakiness, combined with peak audibility ................................................................. 256
10.3 Discussion .......................................................................................................................... 258
10.4 Conclusion .......................................................................................................................... 258
11 Conclusion ........................................................................................................................................... 261
11.1 Relating findings to the literature ...................................................................................... 269
11.2 Implications for further research ...................................................................................... 271
11.2.1 Further investigation of single sound spectral clarity ................................................ 271
11.2.2 Further investigation of spectral clarity in mixes .......................................................... 272
11.2.3 Investigating additional parameters of clarity and separation ..................................... 273
11.2.4 Interrelation between clarity factors .............................................................................. 273
11.2.5 Other parameters of high quality mixes ....................................................................... 273
11.3 Key contributions to knowledge ....................................................................................... 274
11.4 Summary ............................................................................................................................. 275
Appendices ........................................................................................................................................... 276
A1 Listener instructions ............................................................................................................... 276
A2 Factor classification ............................................................................................................... 284
A3 The mid-range spectral peakiness model introduced in chapter 9 ...................................... 285
A3.1 LTAS ........................................................................................................................................ 285
A3.2 Interpolation ...................................................................................................................... 285
A3.3 Curve fitting ..................................................................................................................... 285
A3.4 Calculation of mid-range spectral peakiness .................................................................. 287
A4 Publications .............................................................................................................................. 288
A4.1 Peer reviewed conference papers .................................................................................. 288
A4.2 Research dissemination activities .................................................................................. 288
List of Figures

Fig. 2-1: the important parameters of high quality mixes .................................................................22

Fig. 3-1: the correlation between timbral clarity and energy concentration in octave bands according to Solomon [1959]. Here, a positive correlation indicates that high energy in a given octave band is associated with clarity and a negative correlation indicates the opposite ........................................................................................................................................27

Fig. 3-2: the range of acceptable noise levels in concert halls and opera houses, where NCB-10 is the recommended limit for the noise and NBC-15 represents the limit above which noise levels would be seriously disturbing [Beranek, 1996] ........................................................................................................................................31

Fig. 3-3: fragments organize themselves more strongly if information for occlusion has been added. On the right, this principle is shown in the context of pure tone glides interrupted by noise bursts [Bregman, 1990] ........................................................................................................................................39

Fig. 3-4: a visual example of “belongingness”. The highlighted contour seems to belong to the dark shape [Bregman, 1990] ........................................................................................................................................40

Fig. 3-5: a visual example of duplex perception: One of the parallel lines in the rectangle is also one of the sides of the square inside the circle [Bregman, 1990] ........................................................................................................................................40

Fig. 3-6: an increase in the noise power with the masker’s bandwidth leads to an initial increase in the target’s threshold until a point is reached where the threshold does not increase further [Moore, 2012] ........................................................................................................................................53

Fig. 3-7: the threshold shift of a pure tone signal as a function of its frequency, in the presence of a narrow band of noise centred at 410Hz. For the high frequency functions, the slopes for the curves become shallower at high sound pressure, as measured by Wegel and Lane [Moore, 2012] ........................................................................................................................................55

Fig. 3-8: the ISO standard equal loudness contours for loudness levels from 10 to 100 phons for sounds presented binaurally from the front [ISO 226, 2003] ........................................................................................................................................65

Fig. 3-9: spectrograms of the phrase “Jazz and Swing fans like fast music”, recorded as natural speech (top) and reproduced as sine wave speech (bottom) [Remez, 2008] ........................................................................................................................................80

Fig. 3-10: octave band weighting factors for the two versions of the articulation index and the speech transmission index, as presented by Humes et al. [1986]. The weighting factors have been divided by 30, as is the case in the formulae ........................................................................................................................................87

Fig. 4-1: the familiarization interface ...........................................................................................................103
Fig. 4-2: the clarity only 'C' test interface................................................................. 104
Fig. 4-3: the all attributes 'F' test interface................................................................. 104
Fig. 4-4: notched box plot for CF-C (no exposure to other attributes) and CF-F (simultaneous exposure)........................................................................................................... 108
Fig. 4-5: notched box plots for CF-C (no exposure to other attributes) and FC-F (simultaneous exposure). 'b' = boost, 'c' = cut. ........................................................................................................... 108
Fig. 4-6: notched box plots for CF-C (no exposure to other attributes) and FC-C (previous exposure)........................................................................................................................................... 109
Fig. 4-7: notched box plots for FC-C (previous exposure) and CF-F (simultaneous exposure)................................................................................................................................. 111
Fig. 4-8: notched box plots for FC-C (previous exposure) and FC-F (simultaneous exposure). ................................................................................................................................................ 112
Fig. 4-9: scatter plot comparing CF-C and CF-F.......................................................... 112
Fig. 4-10: scatter plot comparing FC-C and FC-F.......................................................... 113
Fig. 4-11: notched box plots for FC-F (no previous clarity-only rating) and CF-F (previous clarity-only rating). ........................................................................................................ 114
Fig. 5-1: smoothed long-term average spectrum of the unequalised guitar programme item. .............................................................................................................................................. 120
Fig. 5-2: smoothed long-term average spectrum of the unequalised piano programme item. ................................................................................................................................................ 120
Fig. 5-3: waveform of the unequalised guitar programme item........................................ 121
Fig. 5-4: waveform of the unequalised piano programme item. ......................................... 121
Fig. 5-5: creating boosts and cuts in Logic 9..................................................................... 122
Fig. 5-6: an example of the test page ............................................................................. 123
Fig. 5-7: the familiarization interface............................................................................. 124
Fig. 5-8: two examples of the histograms used to assess normality of the data. The clarity ratings are approximately normally distributed for stimulus pair 11 but slightly less for stimulus pair 12.................................................................................................................. 126
Fig. 5-9: distribution of differences between the two ratings for all stimulus pairs, plotted individually for each listener.................................................................................................................. 127
Fig. 5-1: Distribution of clarity ratings for all boosts (white) and cuts (grey) in comparison to the unequalised reference stimuli. Here, both piano and guitar ratings are included in each distribution.

Fig. 5-11: Distribution of clarity ratings for all boosts (white) and cuts (grey) in comparison to the unequalised reference stimuli, excluding the three least consistent listeners. Here, both piano and guitar ratings are included in each distribution.

Fig. 5-12: Smoothed long-term average spectra of the unequalised guitar and piano programme items.

Fig. 5-13: Distribution of clarity ratings for all boosts in comparison to the unequalised reference stimuli, plotted separately for piano (grey) and guitar (white) stimuli.

Fig. 5-14: Distribution of clarity ratings for all cuts in comparison to the unequalised reference stimuli, plotted separately for piano (grey) and guitar (white) stimuli.

Fig. 6-1: An example of the test page.

Fig. 6-2: The familiarization interface.

Fig. 6-3: Long-term average spectra for the violin programme items.

Fig. 6-4: Long-term average spectra for the cello programme items.

Fig. 6-5: Two examples of the histograms used to assess normality of the data. The clarity ratings are approximately normally distributed for stimulus pair 7 but slightly less for stimulus pair 25.

Fig. 6-6: Distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.

Fig. 6-7: Distribution of clarity ratings for all boosts in comparison to the unequalised reference stimuli.

Fig. 6-8: Distribution of clarity ratings for all cuts in comparison to the unequalised reference stimuli.

Fig. 6-9: Differences in spectral centroids, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair, each circle is the median.

Fig. 6-10: Differences in spectral centroids, calculated in Cents and plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.
Fig. 6-11: spectral centroids, calculated in Mels, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-12: spectral centroids, calculated in ERBs, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-13: differences in spectral centroids, calculated in Cents, and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-14: differences in spectral centroids on clarity, calculated in Mels and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-15: differences in spectral centroids on clarity, calculated in ERBs and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-16: harmonic number boosted, plotted against clarity ratings. Harmonics are shown on a logarithmic scale for an even spread of x-axis values. This reflects the fact that pitches are arranged logarithmically on the basilar membrane.

Fig. 6-17: harmonic number cut, plotted against clarity ratings. Harmonics are shown on a logarithmic scale for an even spread of x-axis values. This reflects the fact that pitches are arranged logarithmically on the basilar membrane.

Fig. 6-18: changes in harmonic centroid, plotted against clarity ratings.

Fig. 6-19: changes in harmonic centroid, plotted against clarity ratings (boosts only).

Fig. 6-20: changes in harmonic centroid, plotted against clarity ratings (cuts only).

Fig. 6-21: differences in spectral centroids, plotted against clarity ratings.

Fig. 6-22: differences in Mel centroids, plotted against clarity ratings.

Fig. 6-23 differences in ERB centroids, plotted against clarity ratings.

Fig. 6-24: differences in Cent centroids, plotted against clarity ratings.

Fig. 6-25: boosts on harmonic numbers, plotted against clarity ratings.

Fig. 6-26: cuts on harmonic numbers, plotted against clarity ratings.

Fig. 6-27: differences in harmonic centroids, plotted against clarity ratings.

Fig. 7-1: an example of the test page.
Fig. 7-2: the familiarization interface

Fig. 7-3: two examples of the histograms used to assess normality of the clarity data. The clarity ratings are approximately normally distributed for stimulus pair 4 but slightly less for stimulus pair 2.

Fig. 7-4: distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.

Fig. 7-5: differences in spectral centroids, plotted against clarity ratings.

Fig. 7-6: differences in Mel centroids, plotted against clarity ratings.

Fig. 7-7: differences in ERB centroids, plotted against clarity ratings.

Fig. 7-8: differences in Cent centroids, plotted against clarity ratings.

Fig. 7-9: differences in harmonic centroids, plotted against clarity ratings.

Fig. 7-10: the influence of cutting harmonics on clarity.

Fig. 7-11: the influence of boosting harmonics on clarity.

Fig. 8-1: an example of a test page.

Fig. 8-2: recorded fader use on a test page. The participant experiments with different combinations of parameters.

Fig. 8-3: histogram showing the ratings given by listeners.

Fig. 8-4: scatter plot of linear spectral centroid shifts, plotted against ratings for one listener. This listener uses the top part of the slider only (grey points correspond to stimuli with unnaturally high initial centroids).

Fig. 8-5: scatter plot of linear centroid shifts, plotted against ratings for another listener. This listener uses the bottom part of the slider only (grey points correspond to stimuli with unnaturally high initial centroids).

Fig. 8-6: scatter plot of linear centroid shifts, plotted against ratings for a third listener. This listener rates all adjustments nearly equally (grey points correspond to stimuli with unnaturally high initial centroids).

Fig. 8-7: histogram showing the Q values used. The smallest Q values (widest bandwidth) are used most often.

Fig. 8-8: histogram showing the smaller Q values used only.

Fig. 8-9: histogram showing typical gain settings. Boosts are preferred.

Fig. 8-10: histogram showing the frequencies most often targeted.
Fig. 8-11: preferred gain settings for each programme item, plotted separately for boosts and cuts.......................................................... 198

Fig. 8-12: preferred Q settings, plotted separately for each programme item. Some outliers are not shown as they around the biggest possible Q setting and would hence cause the boxes to be compressed vertically. One outlier each exists for 'sit bright' and 'sit rez'........ 198

Fig. 8-13: preferred boosted and cut frequencies, plotted separately for each programme item. .................................................................................................................................................................................. 199

Fig. 8-14: preferred harmonics boosted and cut, plotted separately for each programme item. .................................................................................................................................................................................. 200

Fig. 8-15: no correlation exists between preferred gain and frequency settings. .............. 201

Fig. 8-16: no correlation exists between preferred gain and Q settings........................................ 201

Fig. 8-17: no correlation exists between preferred Q and frequency settings......................... 202

Fig. 9-1: bass, 2000Hz boost. Clarity has decreased significantly despite a raised harmonic centroid (group 1)........................................................................................................................................................................ 215

Fig. 9-2: bass, 500Hz cut. Clarity has decreased significantly despite a raised harmonic centroid (group 1)........................................................................................................................................................................ 215

Fig. 9-3: vocal, 500Hz cut. Clarity has decreased significantly despite a raised harmonic centroid (group 1)........................................................................................................................................................................ 216

Fig. 9-4: bowed cello, 250Hz cut: the harmonic centroid and clarity both increased significantly (group 2)........................................................................................................................................................................ 216

Fig. 9-5: piano, 500Hz cut: the harmonic centroid and clarity both increased significantly (group 2)........................................................................................................................................................................ 217

Fig. 9-6: piano, 8kHz boost: the harmonic centroid and clarity both increased significantly (group 2)........................................................................................................................................................................ 217

Fig. 9-7: the curve for the mid-range spectral peakiness metric is fitted to the vocal 125Hz cut stimulus LTAS (increased peakiness). It fits closely below the fundamental frequency, then floats above the LTAS for frequencies slightly higher than the fundamental frequency, in order to ignore any peaks in this area. In the upper mid-range, the curve lies below all peaks. Therefore, these peaks contribute to the mid-range spectral peakiness metric. At very high frequencies, the curve fits closely to the LTAS again .............................................................. 219

Fig. 9-8: the combined clarity metric, plotted against clarity ratings for strings (chapter 6).... ........................................................................................................................................................................................................................................ 224
Fig. 9-9: the combined clarity metric, plotted against clarity ratings for vocals and basses (chapter 7) .................................................................................................................................................. 224

Fig. 10-1: an example of a test page ............................................................................................................................................................................................................. 227

Fig. 10-2: the familiarization interface for the bass test half, featuring either all bass stimuli or all vocal stimuli .................................................................................................................................................................................................................. 228

Fig. 10-3: two examples of the histograms used to assess normality of the clarity data. The clarity ratings are approximately normally distributed for stimulus pair 4 but slightly less for stimulus pair 19.................................................................................................................................................................................................................. 231

Fig. 10-4: distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.................................................................................................................................................................................................................. 232

Fig. 10-5: scatter plot comparing clarity and separation data: there is no correlation........ 233

Fig. 10-6: harmonic centroid changes plotted against clarity ratings for stimuli where the target was equalized (masking and fusion caused by the backing track are ignored)........ 234

Fig. 10-7: clarity prediction based on harmonic centroid and mid-range spectral peakiness changes plotted against clarity ratings (for stimuli where the target was equalized, masking and fusion caused by the backing track are ignored). Clarity = -4.918 + 1.095•HC, peaky stimuli set to -4.918.................................................................................................................................................................................................................. 235

Fig. 10-8: the influence of bass boosts on clarity, measured across 8 octave bands......... 236

Fig. 10-9: the influence of bass cuts on clarity, measured across 8 octave bands.......... 236

Fig. 10-10: the influence of vocal boosts on clarity, measured across 7 octave bands ...... 237

Fig. 10-11: the influence of vocal cuts on clarity, measured across 7 octave bands ...... 237

Fig. 10-12: the impact of audible harmonic centroid changes on clarity ratings through backing track equalization.................................................................................................................................................................................................................. 241

Fig. 10-13: the impact of audible harmonic centroid changes on clarity ratings through target equalization .................................................................................................................................................................................................................. 241

Fig. 10-14: overall target-to-backing track ratio, plotted against target spectral clarity..... 243

Fig. 10-15: the impact of changes in overall mix loudness on target clarity.................. 244

Fig. 10-16: the influence of bass cuts and backing track boosts on clarity, measured across 8 octave bands .................................................................................................................................................................................................................. 246

Fig. 10-17: the influence of bass boosts and backing track cuts on clarity, measured across 8 octave bands .................................................................................................................................................................................................................. 246
Fig. 10-18: the influence of vocal boosts and backing track cuts on clarity, measured across 8 octave bands ................................................................. 247

Fig. 10-19: the influence of vocal cuts and backing track boosts on clarity, measured across 8 octave bands ................................................................. 247

Fig. 10-20: target-to-backing track ratio between the vocal and backing track in the unequalised reference ................................................................. 248

Fig. 10-21: target-to-backing track ratio between the bass and backing track in the unequalised reference ................................................................. 248

Fig. 10-22: model prediction, plotted against clarity ......................................................... 251

Fig. 10-23: changes in harmonic centroids, plotted against target spectral clarity (resulting from target and backing track equalization) ............................................. 252

Fig. 10-24: changes in Mel centroids, plotted against target spectral clarity ..................... 254

Fig. 10-25: the combined overall harmonic centroid, mid-range spectral peakiness and peak audibility metric, plotted against clarity ......................................................... 256

Fig. 10-26: model prediction, plotted against clarity ......................................................... 257

Fig. A-1: listener instructions for the clarity only test half in the pilot study (chapter 4) .... 277

Fig. A-2: listener instructions for the clarity, brightness, warmth and fullness test half in the pilot study (chapter 4) ................................................................. 278

Fig. A-3: listening test instructions for the guitar and piano clarity rating task (chapter 5) ................................................................. 279

Fig. A-4: listening test instructions for the plucked and bowed violin and cello clarity rating task (chapter 6) ................................................................. 280

Fig. A-5: listening test instructions for the bass and vocal clarity rating task (chapter 7) ... 281

Fig. A-6: listening test instructions for the EQ adjustment task (chapter 8) ..................... 282

Fig. A-7: listening test instructions for the rating task for spectral target clarity and separation in mixes (chapter 10) ................................................................. 283
List of Tables

Table 2-1: frequently mentioned low-level parameters of perceived mix quality and their influence on the high-level parameters, as discussed in this chapter .......................................................... 10

Table 3-1: equations and typical ranges of the C80, C50 D50 and EDT as provided in the British Standards [ISO 3382-3, 2012]..................................................................................................................32

Table 3-2: all parameters that have an impact on clarity and separation can be assigned to the parameters “spectrum”, "spatial factors”, “intensity” and “temporal changes”. ......................92

Table 3-3: previously established low-level parameters of clarity and separation in mixes are related to findings presented in this chapter..................................................................................95

Table 4-1: the stimulus pairs used in the experiment. lmg = loudness-matched guitar, lmp = loudness-matched piano, 250/500/1000/2000/4000/8000 = boost/cut centre frequency (Hz), b = boost and c = cut..................................................................................................................101

Table 4-2: paired t-test results: no exposure to other attributes (CF-C) vs. exposure to other attributes ........................................................................................................................................107

Table 4-3: average interquartile ranges for the ratings. (Full ratings range was 0 to 100). 109

Table 4-4: paired t-test results: previous exposure to other attributes (FC-C) vs. simultaneous exposure ........................................................................................................................................110

Table 4-5: ANOVA test with the stimulus pair and the listeners as factors. R Squared = .811 (Adjusted R Squared = .622).......................................................................................................................................115

Table 6-1: the approximate range of the fundamental frequencies, average spectral centroids and the range of possible fundamental frequencies of 15 musical instruments..138

Table 6-2: the harmonics present towards the middle of each octave band (approximate for higher harmonics). In each cell, the harmonic numbers are given (e.g. “3” denotes “third harmonic”)........................................................................................................................................140

Table 6-3: the distance of the center of each octave band from the spectral centroid of each programme item in Hz........................................................................................................................................140

Table 6-4: correlations and RMSE’s for measured listening test data and predicted data...146

Table 6-5: all correlation coefficients, sorted descending by the Spearman coefficient. All correlate significantly positively, except “harmonic number cut” (significant negative correlation). All p-values are lower than 0.05 ..........................................................................................152
Table 6-6: all correlation coefficients, sorted descending by the Spearman coefficient. All correlate significantly positively, except “harmonic number cut” (significant negative correlation). All p-values are lower than 0.05.

Table 6-7: Spearman correlation coefficients for all clarity predictors for piano and guitar stimuli.

Table 7-1: harmonic centroids of all stimuli.

Table 7-2: Spearman correlation coefficients relating harmonics boosted and cut and centroid deviations to median clarity ratings. Although all EQ settings reduced clarity, all predictors correlate strongly positively with clarity ratings.

Table 7-3: Spearman correlation coefficients relating harmonics boosted and cut to median clarity ratings. Significant correlations are shown in bold. Although all EQ settings reduced clarity, the harmonic number boosted has a strong positive correlation with clarity ratings: the negative contribution to clarity diminishes towards higher frequencies.

Table 8-1: the chosen programme items vary in the factors that may be important for clarity. AS stands for ‘articulation sounds’. The presentation of the stimuli to test subjects.

Table 8-2: elicited clarity factors found in the clarity definitions and the qualitative data collected on each test page. Sums are shown per factor and programme item and overall.

Table 8-3: relation of elicited clarity factors to factors previously found in the literature.

Table 9-1: for all stimuli in experiments 2,3 and 5, changes in clarity, harmonic centroid and the prediction by the mid-range spectral peakiness model are shown in that order. ‘SP’ and ‘SN’ mark significant positive and negative changes, i.e. confidence intervals were completely above or below 0 for clarity ratings, and the harmonic centroid was either lowered or raised by more than 1.5 harmonics. ‘N’ and ‘P’ indicate smaller changes, 0 indicates ‘no change in clarity’. Group 1 is marked in yellow and group 2 is marked in green (significant changes in both HC and clarity).

Table 9-2: linear regression model coefficients for HC/ mid-range spectral peakiness and HC only models.

Table 9-3: correlation coefficients, worst case Monte Carlo Spearman’s Rho and root mean square error for the mid-range spectral peakiness /HC and HC only linear regression models.
Table 10-1: Spearman correlation coefficients, p-values, RMSEs and Monte Carlo simulation minimum Rho values for overall mix harmonic centroid and peak audibility metrics and median clarity ratings (no changes were measured in overall mid-range spectral peakiness).

Table 10-2: Spearman correlation coefficients, p-values, RMSEs and Monte Carlo simulation minimum Rho values for all models tested in the current chapter that exhibited correlation with median clarity ratings.

Table 11-1 clarity factors found in the literature and in the experimental phase of the research project and their overlap.

Table A-1: factor classification.
List of Equations

2.1 Spectral centroid .................................................................15
3.1 \( C_{80} \), Perceived clarity of sound (music) ........................................32
3.2 \( D_{50} \), Perceived definition of sound, often for speech........................32
3.3 \( C_{50} \), Perceived clarity of sound (speech) ........................................32
3.4 \( C_{50} \), Perceived clarity of sound (speech, alternative version) ..........32
3.5 Noise power inside a critical band when a noise just masks a tone..........54
3.6 Equivalent rectangular bandwidth .................................................54
3.7 Loudness as a power function of intensity ........................................65
3.8 Speech transmission index ........................................................85
3.9 Articulation index .......................................................................87
6.1 Mels as a function of frequencies in Hz ...........................................150
6.2 ERBs as a function of frequencies in Hz ...........................................150
6.3 Distance of a frequency value from 27.5Hz, expressed in Cents ..........151
6.4 Calculation of Mel frequency vectors for the Mel spectral centroid calculation ....151
6.5 Calculation of ERB frequency vectors for ERB spectral centroid calculation ..151
6.6 Calculation of ERB frequency vectors for Cent spectral centroid calculation ....151
6.7 Spectral centroid calculated in Mels ...............................................151
6.8 Spectral centroid calculated in Cents ...............................................151
6.9 Spectral centroid calculated in ERBs ...............................................151
9.1 Predicted clarity value for stimuli with increased midrange spectral peakiness ....221
10.1 Time smoothing function for the calculation of target-to-backing-track ratios ....239
10.2 Convolution of the target signal with the time smoothing function ........239
10.3 Convolution of the interferer signal with the time smoothing function ....239
10.4 Calculation of a time-frequency target-to-backing-track ratio mask ........239
10.5 Transformation of the time-frequency target-to-backing-track ratio mask into a binary mask .................................................................239
10.6 Element-wise multiplication of the target output of the auditory model with the binary mask, removing all masked parts ................................................................. 240
10.7 Calculation of a new target vector, containing only unmasked part .................. 240
10.8 Deduction of the reference LTAS amplitude values from the equalized stimulus LTAS amplitude values .................................................................................. 249
10.9 Calculation of a vector of differences in peak audibility between target and backing track, containing values for each LTAS frequency bin ........................................ 249
10.10 Calculation of a weighting function for the vector of differences in peak audibility between target and backing track, first step ................................................. 249
10.11 Calculation of a weighting function for the vector of differences in peak audibility between target and backing track, second step ........................................... 249
10.12 Calculation of the peak audibility metric ................................................................ 250
10.13 Overall linear regression model incorporating target harmonic centroid, mid-range spectral peakiness and overall mix harmonic centroid ........................................ 256
A.1 Calculation of the midrange spectral peakiness curve for a stimulus LTAS ........285
A.2 Calculation of the difference between the current LTAS value and the previous curve value ........................................................................................................ 286
A.3 Function that determines how tightly the curve is fitted to the current LTAS value \( c(n) \) when the LTAS increases in amplitude compared to the previous curve value ......... 286
A.4, A.5 Calculation of a vector of integers used in A.3 .................................................. 286
A.6 Function that determines how tightly the curve is fitted to the current LTAS value \( c(n) \) when the LTAS decreases in amplitude compared to the previous curve value ......... 287
A.7, A.8 Calculation of a vector of integers used in A.6 .................................................. 287
A.9 Weighting function that is is applied to data points that lie further away from the curve before calculating the overall mid-range spectral peakiness value ................................ 287
Acknowledgements

I wish to express my sincere appreciation to those who have contributed to this thesis and supported me in one way or the other during this amazing journey.

First and foremost, I would like to thank my supervisors, Dr Tim Brookes and Dr Chris Hummersone for their invaluable expertise, patience and support above and beyond the call of duty. Without their guidance and constant feedback, this PhD would not have been possible. Heartfelt thanks goes to my principal supervisor Tim for all the support and encouragement he gave me, not only in the context of my PhD, but also in the development of my career. Tim provided me with numerous opportunities to learn and develop professionally. I thank Chris for his tremendous academic support and for sharing his expertise in programming and statistics whenever I needed help. I gratefully acknowledge the funding I have received from the EPSRC, making this project possible and allowing me to travel to New York and Paris to conferences.

I would like to thank the (past and present) members of the Institute of Sound Recording research group for their inspiration, good humor and support: Dr Andy Pearce, Daisuke Koya, Dr Cleo Pike, Dr Jon Francombe, Dr Tommy Ashby, Dr Khan Baykaner, Dr Phil Coleman, Dr Toby Stokes, Dr Marek Olik, and Hanne Stenzel. The group has been a source of friendships, as well as good advice and collaboration. I also learned how to make the perfect cup of coffee.

Thanks to the members of staff of the Institute of Sound Recording and the Centre for Vision, Speech and Signal Processing for giving me technical and subject-related advice and asking me useful questions at IoSR seminar days: Dr Martin Dewhirst, Rupert Flindt, Alan Haigh, Caroline Haigh, Dr Philip Jackson, Dr Russell Mason and Professor Tony Myatt. Thanks also to the Tonmeister undergraduate students who took part in my listening tests.

I thank Dr Simon Lygo-Baker for great discussions over coffee about research and teaching in higher education and for making the Graduate Certificate in Learning and Teaching an inspiring and intellectually stimulating experience. He helped me get closer to my goal of becoming a lecturer and gave me useful advice on the lectures I taught during my PhD. Thanks also to Dr Enrique Perez-Gonzales for giving me the invaluable opportunity to undertake an internship at Solid State Logic and to critically evaluate my academic research in an industry context.
Thanks to my family for their constant unconditional support – both emotionally and financially. Finally, I would like to acknowledge the most important person in my life – my partner Joe. He has been a constant source of strength and inspiration.
1 Introduction

While pieces of music can be recorded straight to stereo, many are recorded as multiple tracks. These tracks are mixed in order to create a single stereo (or surround) track. In this process, the loudness, dynamic envelope, spatial position, spectra and other features of the tracks are adjusted. For a given multi-track recording, there are a potentially infinite number of ways the tracks could be combined into a final mix. In many situations, creating a high-quality mix is not a simple process; difficulties arise from e.g. time constraints and lack of expertise. Certain qualities seem to exist that all successful mixes have in common. Therefore, mix quality meters or automatic mixing tools could be developed.

Such meters and tools might be beneficial to amateur producers intending to assess or improve the quality of their recordings. Around 6000 students graduate from music technology related university courses each year [Graeme, 2012]. Automatic mixing measurements could help A-level/university music technology students improve their mixing skills, as they could directly assess their own progress. Furthermore, live sound engineers often work under tight time constraints [Biederman and Pattison, 2014]. Hence, a tool able to automatically measure aspects of mix quality would be useful. Amateur live sound engineers could also benefit from this (e.g. live sound in churches). Finally, object-based audio has recently become the focus of interest for large corporations like Dolby, DTS and the BBC [2015]. Renderers in object-based audio reproduction systems could monitor their output for mix quality and fine-tune their rendering parameters to optimise it.

The aim of the current research project is to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes. The findings could contribute to the development of marketable products such as a piece of software able to judge the overall sound quality of a mix, automatic mixers or sonically improved music production software.

1.1 Methodology and structure

Scott [2014] points out that many disciplines, including signal processing, music cognition, machine learning and human computer interaction are involved in pursuing automatic multi-track analysis and mixing. There are different approaches to solving this complex problem, the most common of which is knowledge engineering [De Man and Reiss, 2013a]. Here, informally known rules for creating high quality mixes are implemented in technology. Research in this
Introduction

field usually draws upon mixing rules laid out in non-academic mixing guides. The second approach is grounded theory, where basic knowledge is acquired first and subsequently transferred to the intelligent system. Psychoacoustic studies are undertaken to define mix attributes, and perceptual audio evaluation is employed to determine listener preference for mix approaches. De Man and Reiss [2013a] argue that the grounded theory approach is very resource intensive and therefore too limited to constitute a sufficient knowledge base for the implementation of an overall system. However, knowledge engineering is a less formalized approach [Scott, 2014] and many commonly accepted rules in mixing, such as the notion that most elements should be high pass filtered above their fundamental frequency, do not hold true in formalized studies [Pestana and Reiss, 2014a]. Since the current project will ultimately focus on a single perceptual attribute, rather than intending to develop an overall mix measurement system, a grounded theory approach is adopted. In the following, the methodology and structure of the research project is presented.

Chapter 2 — High-level descriptive quality criteria and lower-level perceptual components of music mixes

Goal: compile a list of the high-level descriptive quality criteria and lower-level perceptual components that, according to published literature, contribute to the perceived quality of mixed music.

In an initial literature review, the parameters of high quality mixes are established. Both academic and non-academic literature is consulted. One important parameter (clarity and separation) is selected as the focus for subsequent research.

Chapter 3 — Relating psychoacoustic findings to clarity and separation in music mixes

Goal: establish the acoustic and mix parameters that clarity and separation are likely to depend on, by consulting literature on acoustics and psychoacoustics.

Relevant acoustic and psychoacoustic factors are established through a review of academic literature. Six areas of scientific research (outside the context of music mixes, where clarity has not been investigated sufficiently) are studied. The factors are subsequently grouped into four categories. One category of factors (spectrum) appears to be particularly important across all areas of research and is therefore chosen as the basis for further research.
Chapter 4 — The influence of dumping bias on spectral clarity ratings (pilot)

Goal: evaluate a suitable experimental setup for the assessment of the relationship between spectral factors and clarity, and test for dumping bias.

To investigate the impact of spectrum on clarity, a rating task is chosen for initial experimentation. It is decided that spectral clarity should be tested for isolated sounds first (putting to one side the 'separation' aspect), in order to focus on specific spectral factors, before moving on to sounds in mixes. The research project focuses on the effect of single band EQ on single sound spectral clarity.

The first experiment is a pilot study with the purpose of testing a potential setup. Furthermore, the possibility of dumping bias is considered, whereby the perception of attributes not included in a listening test can influence the ratings of the attribute in question.

Chapter 5 — Assessing the contribution of different octave bands to the single sound spectral clarity of piano and guitar stimuli

Goal: establish how boosts and cuts in different frequency regions influence single sound spectral clarity.

The relationship between spectral equalisation and single sound spectral clarity is assessed for two programme items (guitar and piano) in a listening test.

Chapter 6 — Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Goal: establish how changes in the spectral clarity of equalized programme items with differing fundamentals and original spectral centroids can be predicted.

The relationship between single sound spectral clarity and those spectral factors that seem particularly important (fundamental frequency and spectral centroid) is investigated further on a set of four programme items. One particularly useful predictor for single sound spectral clarity is established.

Chapter 7 — The single sound spectral clarity of vocal and bass stimuli

Goal: test the predictor of single sound spectral clarity on a new set of stimuli.
1 Introduction

The predictor found in the preceding chapter is tested on two further stimuli. It is established that the predictor only works with limited accuracy and the possibility of including further predictors is considered.

Chapter 8 — Single sound clarity adjustment task

Goal: establish what additional factors are likely to be important for spectral clarity, what the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ are and how spectral clarity can be defined in the context of EQ adjustments.

In order to investigate how spectral clarity can be predicted more accurately, a combined verbal elicitation and adjustment task is carried out, leading to a set of additional factors and a more refined definition of spectral clarity. One particularly important factor is selected that has thus far not been related to clarity in the literature.

Chapter 9 — An improved spectral clarity predictor

Goal: establish if the previous predictor of spectral clarity (for single sounds) can be improved by including a metric for unpleasant peaks in the spectrum.

A computational metric is devised for the important factor established in the last chapter. Through a combination of the original clarity predictor and the new metric, the spectral clarity of single sounds can be predicted well.

Chapter 10 — Single sound spectral clarity in mixes

Goal: test and improve the previous clarity metric for sounds in mixes, as affected by spectral equalization.

The clarity of sounds in mixes is investigated. A mixture of the previously introduced metrics and a new, additional metric prove useful in measuring clarity in mixes, as affected by spectral equalisation.

1.2 Conclusion

Mixing music is the process of combining tracks of recorded audio to an overall piece. This is a complicated process and, hence, automatic mixing or metering tools would be useful. The aim of the current research project is to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes.
1 Introduction

A grounded theory approach is employed, as follows: first, the relevant high-level descriptive mix quality criteria and the lower-level perceptual attributes that relate to them will be established through a literature search. One particularly important parameter (clarity and separation) is chosen to act as the focus for the remainder of the project. By drawing on academic literature, the acoustic factors that clarity and separation may depend on are established. These factors are used as a guideline for experimentation, which ultimately leads to the development of a metric of this important aspect of mix quality. Findings could contribute to the development of marketable products such as a piece of software able to judge the overall sound quality of a mix, automatic mixers or sonically improved music production software.
2 High-level descriptive quality criteria and lower-level perceptual components of music mixes

As set out in the introduction, the aim of the current research project is to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high quality mixes. A grounded theory approach will be taken, whereby the relevant high-level descriptive mix quality criteria and the lower-level perceptual attributes that relate to them will be established first.

The current chapter aims to compile a list of the high-level descriptive quality criteria and lower-level perceptual components that, according to published literature, contribute to the perceived quality of mixed music. In section 2.1, the current state of the art and the representation of mix quality parameters therein is briefly discussed. Next, parameters as found in the literature are summarized and structured (section 2.2). Here, it is shown that despite the inconsistent and often incomplete presentation of mix quality parameters in the literature, a consensus exists on what makes a good mix (section 2.3). The next step for the overall research project is discussed in section 2.4. Section 2.5 concludes.

2.1 General comments on the literature

In the following, the sources consulted are introduced, evaluating their relevance in the context of this project. Both academic and non-academic sources were included in the search: non-academic sources tend to mention a greater number of mix parameters; academic sources are supported by more thorough scientific research.

2.1.1 Academic sources on mix quality

Due to the interdisciplinary nature of the research topic, academic publications stemming from a variety of scientific disciplines were incorporated in the search. Although some researchers did attempt to compile a list of important mix parameters, e.g. through verbal elicitation [Wilson and Fazenda, 2016], there does not appear to be a commonly accepted consensus of which parameters constitute mix quality, nor a piece of academic literature presenting all parameters of a good mix as mentioned collectively in all sources. Occasionally, academics comment on the lack of parameters necessary to measure the quality of reproduced sound, e.g. Rumsey [2002] points out the “need for reliable, preferably unidimensional, spatial attributes” for the spatial
quality evaluation of reproduced sound, deeming existing descriptors to be insufficient. As mentioned in the introduction, many researchers favour a knowledge engineering approach where a specific subset of mixing rules, as presented in non-academic mix literature (e.g. EQ rules for masking reduction), are directly implemented in technology without first establishing a formalized ontology of all important quality parameters [e.g. De Man and Reiss, 2013b, Reiss, 2011 and Fejzo and Maher, 2014].

2.1.2 Non-academic sources on mix quality

Due to the small amount of academic literature on mix quality, non-academic journals and books on mixing, recording and music production were examined. Here, the authors (usually mixing engineers or producers) derive quality criteria from their personal experiences. For example, mixing engineer Bill Gibson [2002] strives to create mixes that are appropriate for the style, song and people involved whilst meeting his own taste or the taste of a mass audience. He bases his approach to mixing on the feedback of customers, the mixes of highly grossing popular songs and the current fashion. Often, iconic mixing engineers such as Dave Pensado or Matt Wallace are quoted in order to establish values of a good mix [e.g. Owsinski, 2006 and Clark, 2011]. Similarly, producer and remixer Rick Snoman [2009] stresses that it is the specific style of the mixing engineer that makes the mix special.

Occasionally, authors draw upon scientific findings. For example, Roey Izhaki [Izhaki, 2013] relates delay settings to the Haas window: in his “Haas trick” a sound and its delayed duplicate (up to 35 ms) are panned left and right to achieve a “wide, open, spacious sound”. Usually, non-academic literature about mixing focuses on tools, techniques and ways to alter various parameters but the overall aesthetic goal is not always presented.

2.2 Mix quality parameters according to the sources consulted

A set of mix quality parameters, derived from a review of the relevant literature, is presented. It was attempted to incorporate all parameters mentioned both explicitly and implicitly in the sources in the following overview. As mentioned by Wilson and Fazenda [2016], quality attributes relate to timbre, space, defects, and other concepts. Mix quality parameters are grouped and discussed under the headings “clarity and separation”, “balance”, “impact and interest”, “appropriate context-specific characteristics” and “freedom from technical faults”, as these descriptors are frequently mentioned in the literature and can be thought of as a set of high-level criteria for a high-quality mix (i.e. such a mix should have all of these properties).
All lower-level parameters found can be related to one or more of the above categories. There is a degree of overlap between these categories and some lower-level parameters can have an impact on several high-level parameters at the same time: tonal qualities for instance can be altered to create clarity and frequency balance, as well as interest. The various relationships between high and low-level parameters are summarized in Table 2-1. In the following subsections, all of these parameters are discussed in more detail.

<table>
<thead>
<tr>
<th>Low-level parameters</th>
<th>Clarity and separation</th>
<th>Balance (Spatial balance (Stereo balance and depth))</th>
<th>Timbral balance</th>
<th>Impact and interest</th>
<th>Parameters specific to the context of a piece</th>
<th>Correcting “imperfect” musical performances or recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analogue sound</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appealing acoustical space</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artificial staging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artistic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Artist’s vision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atmosphere</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audibility of dynamic movement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authenticity, naturalness (“sounds possible”)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brightness/bright</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boxy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean/dirty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clean sound</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear transients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compressed sound</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context of the song</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crunchy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Darkness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Definition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distinctiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamics/dynamic contour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edgy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elements conveying instrument timbre and rhythm</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enhancing “magical elements” of a performance</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every element of the soundscape has its place</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Every element serves its sonic purpose</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exaggeration/ larger than life</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exciting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8
## 2 High-level descriptive quality criteria and lower-level perceptual components of music mixes

<table>
<thead>
<tr>
<th>Low-level parameters</th>
<th>Clarity and separation</th>
<th>Balance Spatial balance (Stereo balance and depth)</th>
<th>Timbral balance</th>
<th>Impact and interest</th>
<th>Parameters specific to the context of a piece</th>
<th>Correcting &quot;imperfect&quot; musical performances or recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fashion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fat/thin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feel/emotion/mood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foreground, middle ground, background</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency content/interplay/range</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fullness/full</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good blend</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gutsy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harmonic excitement</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harsh</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High fidelity</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imaginative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Individual style, art</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Intelligibility</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intimacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of masking</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Larger than life</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localization</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical dependence of sound sources</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Loudness</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meaning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Muddy</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No artefacts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>No bad tuning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>No clicks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No harshness</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No hum</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No pops</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No sibilance</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No unwanted noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Blend</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No unwanted resonances</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Not &quot;samey and uninspiring&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Not cluttered</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not ghostly, unfocused</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not jarring</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not muddy</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other aspects of tonality/timbre</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panorama</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perceived loudness</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Pitch density</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2 High-level descriptive quality criteria and lower-level perceptual components of music mixes

<table>
<thead>
<tr>
<th>Low-level parameters</th>
<th>Clarity and separation</th>
<th>Balance Spatial balance (Stereo balance and depth)</th>
<th>Timbral balance</th>
<th>Impact and interest</th>
<th>Parameters specific to the context of a piece</th>
<th>Correcting “imperfect” musical performances or recordings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polished</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Power (fat, powerful lows)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Powerful</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precedent Cues, non-precedent Cues and environmental cues</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Presence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present, shape and enhance musical materials</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Realism (or illusion thereof)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rhythmic subtlety easier to perceive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sizzly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoothness/smooth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sonic detailing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sound right</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spaciousness/spacious</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparkly, tinkly highs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subvert, exaggerate, and/or parody conventions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surprise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetry/not lopsided</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taste</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vision/reference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Warmth/warm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight, fatness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wet/dry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wide</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-1: frequently mentioned low-level parameters of perceived mix quality and their influence on the high-level parameters, as discussed in this chapter.

2.2.1 Clarity and Separation

Clarity has been defined as the audibility of instruments within complex orchestrations [Toole, 1985], as the opposite of “muffled” [Lorho, 2005], the impression of declining transparency with increasing reverberance and the ability to distinguish elements in an auditory scene and detect their properties [Lindau et al., 2014]. However, no commonly agreed definition exists for clarity.

“Clarity and separation” is defined here as the extent to which individual components can be heard in a mix.

In a study conducted by Wilson and Fazenda [2016], loudness, as altered by dynamic range compression and fader positions, played an important role in mix quality. Loudness appears to
be directly linked to clarity, e.g. lead vocals, which should be particularly clear, are often set to the same loudness as the rest of the mix according to Pestana and Reiss [2014a]. The importance of the clarity of individual sound sources in a mix is widely agreed on, especially in the context of lead vocals or other lead sounds [Ronen, 2015]. Clarity plays an important role in audio quality perception in popular music [Wilson and Fazenda, 2016]. Fenton and Wakefield [2012] explain clarity as follows: "[...] you can hear a clear vocal that doesn’t suffer heavily from masking, clear dynamics are evident, clear drum hits/transients, bass notes, a point whereby dynamic movement is clearly audible." They explain that listeners do not always perceive clarity according to the same criteria which can depend on the genre: "for example, an ensemble recording may be judged on clarity by considering the tonality, spaciousness and localisation of each individual instrument whilst a contemporary heavy rock recording could be judged on the clarity of vocal and/or drum sources and overall bass & guitar frequency interplay.”

The authors point out that generally, enough elements conveying instrument timbre and rhythm need to be clearly audible in the mix. Similarly, according to Pachet and Delerue [2000], several sound sources may be logically dependent, such as the bass, drum and guitar tracks composing the rhythm section in some pop songs. This relationship needs to be clearly audible. Bob Clearmountain, as quoted by Clark [2011], stresses that the drums need to blend well with the bass in the mix.

Further authors comment on clarity and separation as follows. Izhaki [2008] introduces “definition” as a parameter of a good mix which fits into the context of clarity, as something that is “defined” is also distinguishable and clear. Bazil [2008] mentions “clarity” and “separation” and Jeff Strong [2009] states that a good vocal mix is characterized by fullness, clarity, brightness and presence, without “muddiness” or “sibilance”.

Similarly, Rick Snoman [2009] mentions the parameters of clarity and contrast. Owsinski writes "clarity is what you aim for" [Owsinski, 2006]. Later on, he quotes a number of iconic mixing engineers, most of which also comment on the importance of clarity: Ken Kessie strives to create a “clear, bright vocal” [Owsinski, 2006] and Kevin Shirley aims for a “present and clear” vocal sound [Clark, 2011]. The importance of vocal presence and clarity is also (implicitly) mentioned in the Computer music magazine [Musicradar, 2012]: “of course, you want some sounds to stand out—the hooks, lead vocals and so on”.

Clarity and separation can be impaired by masking, as is often stressed. Perez Gonzalez and Reiss [2010] attempt to “minimize spectral masking” in their Real-Time Semi Autonomous Audio Panning System for Music Mixing for that reason. In non-academic literature, factors counteracting clarity are often described as “mud” [Strong, 2009], “clutter” or “jarring” [Music
Radar, 2012], etc. However, too much separation can take away from perceived quality: various authors comment on the importance of “glue” in a mix [e.g. Owsinski, 2006] where instruments are perceived as belonging together and don’t sound too separated. This indicates that there might be an optimum value for clarity and separation.

Outside the context of music mixing, clarity and intelligibility are introduced as important qualities of both musical and non-musical sound. Breshears [2001] introduces “intelligibility and clarity” as a “general primary sound system design criterion”. Letowski’s “Mural” [1989] features “clarity” as a parameter contributing to the “distinctiveness” of an auditory image.

Importantly, maximum clarity is not always desirable: Pestana and Reiss mention that blending tracks can also often be effective (informally often described as “glue”), e.g. by using bus and mix compression. Similarly, Moylan stresses that tracks can be fused/blended or given clarity [2007].

All in all, the important characteristics (e.g. timbre, groove, melody etc.) of individual components, especially lead and vocal sounds, need to be clearly audible (and sometimes exaggerated) for a mix to be perceived as clear and separated while other tracks may benefit more from being blended together.

### 2.2.2 Balance

“Balance” in general describes an even distribution of energy in the spatial and frequency domains and is depicted as an important parameter of music mixes in the literature. It is almost always quoted in non-academic literature about mixing [e.g. Izhaki, 2008 and Owsinski, 2006], often referring to level, tonal and spatial balance [Moylan, 2007].

Børja says that in a “horrible mix”, “overall balance may not be what you wanted” [Børja, 1976]. In the same way, Fenton et al. [2011] introduce “poor balance” in general as a characteristic of a badly engineered and produced recording. Similarly, Owsinski [2006] introduces “balance” as one of “the six elements of a mix”. Balance is also mentioned by some of the iconic engineers quoted, such as Dave Pensado and Bill Schnee [Owsinski, 2006].

Mostly, differing attributes of mix quality are referred to as “balance”. De Man and Reiss relate balance to appropriate relative loudnesses of instruments [2013b]. Stereo balance, depth and tonal balance are often referred to as aspects of the creation of an imaginary space: Bazil [2008] stresses the importance of balance of both the frequency content and stereo field. Moylan [2007] talks about musical balance and timbral balance. In non-academic literature, mixes are often described as having three dimensions, height, width, depth [e.g. Savage, 2014, and Gibson,
2 High-level descriptive quality criteria and lower-level perceptual components of music mixes

[2005], relating to tonal or level balance (height), stereo balance (width) and spatial depth. Owsinski writes: “most great mixers think in 3 dimensions. They think ‘tall, deep and wide’” [Owsinski, 2006] which is in accordance with the three dimensions mentioned above [e.g. Savage, 2014]. Equivalently, Snoman [2009] comments on the “three dimensions” of “harmonic balance”, “stereo balance” and “depth”. Zagorski-Thomas [2007] describes this phenomenon as “the ‘artificial’ staging of performances in a virtual environment” and stresses the “practical use of staging as a way of highlighting textual elements that are important in defining the functional meaning of the music.” Moylan [2007] establishes the same dimensions, breaking up the overall parameter of “spatial qualities and relationships” further into the “dimensions of the sound stage”, “placement and size of the sources on the sound stage”, “listener to sound stage distance” and “environments of sources and depth of sound stage”. Another example is: “You want to put the listener in an appealing acoustical space […] enhance the overall impact of the final recording, making sure that every element of the soundscape has its place” [Franz, 2004]. In the following, the overall parameter of “balance” will be investigated further by exploring the “three dimensions” of horizontal (stereo) and tonal balance, as well as depth (foreground and background) separately.

Horizontal balance

Horizontal or stereo balance is the extent to which within any given frequency range sound energy is distributed symmetrically and “evenly” between the left and right channels. Here, personal experiences and the analysis of various mixes lead to the assumption that an “even” distribution features a peak of energy is in the centre that gradually diminishes towards the far left and right.

Horizontal balance is usually achieved through panning or stereo (surround sound) effects. It is frequently commented on in the literature. In stereo recordings, the overall frequency content should be similar in both channels while maximizing the dynamic use of the stereo channels and minimizing spectral masking [Perez Gonzalez and Reiss, 2010, Pestana and Reiss, 2014]. According to Pestana and Reiss, low-end frequencies and the main track should be centrally panned, while other tracks are panned away from centre [2014a]. Terrell, Simpson and Sandler [2013] note that Gonzalez and Reiss’s “automatic panner” sets panning controls to spatially separate sounds with similar frequency content, subject to additional subjective constraints. Mixing engineer Dave Pensado describes hard left or hard right as “sacred/noble territory” that should be used sparingly [Owsinski, 2006] but this did not seem the case in a study by Pestana and Reiss [2014]. Børja [1976] says that in a “horrible mix”, “the panorama may be too wide or too narrow”. Similarly, width is important parameter of mix quality perception [Wilson and
According to De Man et al. [2015], side to mid ratio correlates positively with mix quality perception.

According to Izhaki [2008], examples of unbalanced stereo images are the asymmetrical spreading of frequencies, a mix with more instruments panned to the extremes than the middle (V-mix), concentrated energy in the middle (I-mix) or symmetrical gaps around halfway between the middle and both extremes (W-mix). Gerzon [1992] comments on a “sense of realism” in the stereo field. Similarly, Mitchell [1998] explains that stereo sound should provide a good illusion of an infinite number of real sound sources. He quotes Richard Heyser: "stereo is merely an attempt to create the illusion of reality through the willing suspension of disbelief". However, he admits that not all multichannel musical expression attempts to represent the real (or natural) acoustic environment. Strong [2009] writes that stereo panning is important but that the exact rules for this are unclear.

Music Radar [2012] depicts a stereo mix that is “too narrow” as one of the “10 tell-tale signs of an amateur mix”. Cooper [2013] writes about stereo panning: “stereo-imaging plug-ins can help immensely, but use them recklessly and you’ll get a ghostly, unfocused sound. Another strategy is to hard-pan tracks, but a careless approach here will make the mix sound lopsided”. Despite the commonly agreed rules for panning, it appears that some iconic recordings deliberately deviate from these rules, such as some Beatles tracks, as confirmed by informal listening. At the same time, instruments are often arranged in the same way horizontally as they may be heard on stage in a live environment (this seems to be the case in most orchestral recordings, and in drum mixes where e.g. the snare is not centrally panned).

Depth

Depth is defined here as a sense of perspective in a mix, where sound sources can be placed at various distances from the listener and inside a fictional, reverberant space of a certain size and shape. Berg and Rumsey [2006] quote Toole, where “impression of distance or depth” of reproduced sound is described as “... a satisfactory impression of instruments at various distances. An unsatisfactory reproduction would have all of the instruments at one distance (two-dimensional), or some of them too close or too far, and so on.” This can be applied to mixing. “Spacious” and “deep” were also elicited important parameters in pop production audio quality parameters [Wilson and Fazenda, 2016].

Authors often point out that a sense of perspective (foreground and background) must exist in a high quality mix without specifying perceptual optima. Rudolph & Leonard [2001] stress the importance of “front to back depth” when creating soundscapes. Mitchell [1998] quotes Günther Theile: “a multi-dimensional sound field should exhibit three main types of sound
High-level descriptive quality criteria and lower-level perceptual components of music mixes

signals: Precedent Cues, or direct sound; Nonprecedent Cues, or indirect sound; and Environmental Cues, or atmospheric sound. This is artificially recreated when mixing.

Izhaki [2008] describes reverb-free sounds as 'strange and frontal' which is equivalent to a lack of perspective. As stated by Everest, artificial reverberation is necessary for the recreation of "the richness of the room effect" of realistic spaces, such as concert hall [Everest, 2009] which, again, can be related to the parameter of perspective. Similarly, Senior notes that "natural sounding reverb" will contribute to a good mix [2011]. Howard and Angus [2006] point that the "overall balance" is altered in order to place some sounds in the front and others in the back. Further research would need to be undertaken to establish the perceptual optima of depth in music mixes.

Balance across the frequency spectrum (tonal balance)

Tonal balance is the extent to which sound energy is distributed "evenly" across the frequency spectrum: Chapman [1996] analyses the spectral averages of various genres of recorded music. Despite the slight differences between genres, the resulting functions have similar shapes. Similarly, Owinski [2006] points out that mixing engineers often adapt the frequency spectrum to a reference point whilst trying to achieve a "high fidelity" sound. More recent studies confirm that although every song is unique in its spectral/timbral contour, there exists a target equalisation curve that stems from practices in the music industry but also seems to mimic the natural, acoustic spectra of ensembles [Pestana et al., 2013]: Pestana et al. [2013] explain that spectra of professionally produced commercial recordings show consistent trends, which can roughly be described as a linearly decaying distribution of around 5 dB per octave between 100Hz and 4000Hz, becoming gradually steeper with higher frequencies, and a severe low-cut around 60Hz. The preferred position for the overall mix spectral centroid appears to lie around 2900Hz [Wilson and Fazenda, 2015]. The spectral centroid is a weighted mean, indicating the frequency (Hz), at which the centre mass of energy of a spectrum is situated [Grey, 1978]. It is defined as:

$$C = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$  \hspace{1cm} (2.1)

Here, \(x(n)\) is the magnitude of frequency bin number \(n\), and \(f(n)\) is the centre frequency (Hz) of that bin that result from performing a Fourier transform on the audio signal. \(N\) is the overall number of bins. It is likely that tonal balance is influenced by equalization, as well as the loudness as individual tracks with different spectra.
Tonal balance has been described in terms of timbral attributes, in both academic and non-academic literature. Wilson and Fazenda [2016] found the terms “full”, “harsh”, “thin”, “bright”, “smooth”, and “crunchy” in a verbal elicitation experiment. Katz [2014] points out that sound can be described as “boomy”, “boxy”, “warm”, “present”, “nasal”, “muddy”, “cutting”, “edgy”, “airy”, “sizzly” etc., which he relates to the absence or presence of certain frequencies. Similarly, Moylan [2007] defines tonal balance as the overall sound quality timbral balance. Automatic tools for the correction of tonal balance exist, e.g. for mastering [Mimilakis et al., 2013].

As commented on by Izhaki [2008], “achieving frequency balance (also referred to as tonal balance) is a prime challenge in most mixes”. Case [2007] defines a mix as “balanced” when every element serves its sonic purpose. In the same way, Alkin [1996] stresses the importance of an “Even frequency response”. Moylan [2007] describes the same parameter as “Pitch density and timbral balance” and Bazil [2008] introduces the “balance of the frequency content” as contributing to the quality of a good mix. Owsinski [2006] introduces “frequency range” as one of the six elements of a mix. Here, all frequencies need to be “properly represented”, featuring “sparkly, tinkly highs and fat, powerful lows”.

2.2.3 Impact/Interest

Impact/interest is defined here as the extent to which the mix grabs the listener's attention. Frequently, “interest”, “impact” and “punch” and other such mix qualities are mentioned in the literature. Franz [2004] emphasizes that a mix needs to be “exciting, artistic, powerful and imaginative”. Although the exaggeration of the above parameters can create impact, there are further low-level parameters that have not been mentioned yet.

Zagorski-Thomas [2007] comments on the perceived naturalness of a production which can support impact: "If, at a subconscious level, the perception of music involves hypothesizing what it would feel like to produce that sound, it would be useful for both music and musicology to study the grey area between edited performances that are perceived as possible and those that are perceived as impossible or unnatural”. “Unnatural” music mixes that are perceived as “natural” can for instance feature the warmth or fullness of a concert experience but with an added clarity that is impossible to attain in the natural world: Zagorski-Thomas [2007] states that the exaggeration of intimacy is an aesthetic that is popular in modern productions. This effect is often called “larger than life” and fits into the “Interest and Impact” category. Similarly, Wilson and Fazenda [2016] name “punchy”, “loud”, “strong” and “aggressive” as important mix parameters. Another parameter named here is “synthetic” which may be related to “natural”, as mentioned by Zagorski-Thomas.
In another paper, Zagorski-Thomas [2007] relates musical elements to “physical manifestations of emotions, gesture and being in space”, describing music recordings as sonic metaphors for physiologically and culturally determined gestures and morphologies. Similarly, Théberge links adjectives used to describe low-level mix parameters to bodily sensations. It is assumed that the mimicking of such bodily sensations, e.g. through the manipulation of transients to create punch can create impact.

In the following, further approaches to “impact and interest” are presented. Moylan [2007] introduces contrast (“attention”, “meaning”, “surprise” and “dynamic contour”) as a characteristic of a good mix and states that a mix should present, shape and enhance musical materials. Similarly, Owsinski [2006] comments on “dynamics and interest”. He praises the “sonic detailing, impact and space” found in Sting’s music. Pestana and Reiss [2014] note that contrasting volume automations can help the listener focus on different tracks over time. Izhaki [2008] mentions “interest”, Gibson [2002] talks about “power” and “impact”. Huber and Runstein [2005] state that a good, mastered recording should be “clean, punchy, gutsy …” and that it needs to sound “right”. Fenton and Wakefield [2012] also comment on “punch”: “punch is a subjective term often used by engineers to describe a particular moment in a production where there is a degree of change in power in the music”. Fenton et al. [2014] present descriptors of “punch” within complex musical signals, i.e. “thud”, “weight”, “fast attack”, “thump”, “gated feel”, “energy burst”, “hard”, “dense”, “focussed”, “tight”, “narrow” and “defined”. "Punch" depends on the sound’s frequency spectrum and intensity over time (punch correlates with a spectral centroid at 1217.34Hz, a higher intensity ratio, a higher rhythm strength, i.e. sum of the magnitudes of the power spectrum in the onset of the signal, higher spectral energy in the onset of a sample and lower crest factor) [Fenton et al. 2014]. Fenton et al. [2009] define a good music production as “clear’, ‘defined’, ‘punchy’ or ‘highly polished’”, referring to a badly engineered and produced recording as “‘woolly’, ‘distorted’, [...] or ‘muddy’”. An overlap between the overall categories becomes apparent: “muddiness” can influence impact as well as clarity.

The tonal qualities of individual instruments are often altered to create interest. Howard and Angus [2006] indicate that EQ should be used to create “sparkle”, “weight” or “punch” which also fits into the category of tonal balance. In the context of pop and dance music, lows and highs are often boosted to achieve a “powerful” sound. Snoman [2009] introduces the values of “dynamics”, “harmonic excitement” and “loudness” as refined in the mastering process. Owsinski [2006] notes that listeners perceive “louder” as “sounding better”, but warns of hypercompression as resulting in a weaker sound. Loudness can also be related to “impact” because it helps grab the listener’s attention. Bazil [2008] notes that a good mix should match
2 High-level descriptive quality criteria and lower-level perceptual components of music mixes

the genre and should be suited to different playback environments. Zagorski-Thomas [2007] uses dance music as an example for this: here, the mix helps make “the rhythmic subtlety easier to perceive, to help facilitate dancing”. In this way, impact can also be related to how successfully a mix helps fulfil the purpose of a piece of music.

“Warmth” is often connected to the sound of traditional, analogue studio equipment. According to Music Radar [2012], “the elusive warmth of analogue studio gear is a much sought-after sound in modern recording” and Garba [2006] presents analogue warmth and high-end enhancement as contributing to a good mix. Although “warmth” can also be related to tonal balance, it is often used to describe specific tonal qualities that can create “interest”.

2.2.4 Appropriate context-specific characteristics

A mix fits into its intended context when it conforms to current trends, fashions and norms, complements artistic purpose of the recording and supports the musical content. Although qualities such as clarity, balance and interest are sought after in most mixes, other characteristics of good mixes depend on the context and vary between pieces. De Man and Reiss [2013a] point out that the very books that provide mixing tips stress that mixing is highly non-linear, unpredictable, devoid of ‘hard and fast rules’, ‘magic settings’ or one-size-fits-all equaliser presets. Reiss [2011] stresses that mixing has a creative side that cannot be automated. Similarly, Wilson and Fazenda [2016] note that although certain mixing techniques are more likely than others, aspects of preference and liking are distinct from the interpretation of quality and might not be the best descriptors for studies where technical quality is the percept being sought. The authors note that quality in music production is revealed as a perceptual construct distinct from hedonic, musical preference. King et al. [2012] assume that there may be different schools of mixing where e.g. some engineers prefer a hot lead vocal while others tuck the vocal lower into the mix.

As mentioned above, Gibson [2002] points out that the musical style, song, details and people involved, as well as the mixing engineer’s own taste and the taste of a mass audience will have a strong impact on the final product. Bob Clearmountain says in an interview: “everything depends on the context of the song” [Clark, 2011]. Reed [2000] mentions the tonal qualities of “brightness, darkness, and smoothness” “in a context-dependent fashion”. Théberge [1997] notes: “Speech about music is always metaphoric and somewhat vague in nature but I believe it can also be quite precise”. He comments on the great variety of specialized terms and slang used to describe an “unspecifiable but ordered sense of something”, such as “fat/thin”, “wet/dry” or “clean/dirty”. These contrasting attributes can be related to the fact that mix quality is in part context-specific.
What is perceived as a good mix is not always predictable and fashion can change audience perception over time. Fisher [2012] describes the “very loud, quite compressed” sound of modern pop songs as a fashion trend for example. Senior [2011] emphasizes the importance of mixes to sound like comparable commercial tracks and Garba [2006], too, notes that mixes need to be adapted to other mixes “out there”. Furthermore, Gibson [2002] states that “Each style of music developed its own traditions of how it is mixed”. Fisher gives an example for this in an interview [Wells, 2012]: “Gregorian plainchant was designed for big spaces and to be listened to from quite a long way away and you think ‘well, that’s the sort of overall mix that I’m aiming for’”. Owsinski [2006] identifies differences in mixing styles across different regions. He claims that each modern mix can be assigned to either the “London style” or the “LA style” which he relates to the use of the iconic technology available at each location respectively. He observes that styles have become more similar as the same technology has become available worldwide. On the other hand, he comments on the more individual mix styles developed in differently equipped home studios. Fisher [Wells, 2012] comments on the technical restrictions of mixing for vinyl: “you couldn’t have a large out-of-phase signal because it would make the groove depth go to zero and, of course, when the cutter head leaves the surface of the disk who knows where the replay stylus will land!” Analogously, Zagorski-Thomas [2007] notes: “less frequently considered is the way in which the sound of recordings changed as a result of technological innovations. He also describes the exaggeration of intimacy in vocal tracks, as mentioned above, as a fashion trend. Furthermore, he notices a similar phenomenon in dance music fashion: “despite dance music being primarily aimed at an audience listening in group-based contexts such as dubs, the production aesthetic has moved away from a concert-based sound towards proximity and clarity in rhythm-section mixing and with the live aesthetic applied only to vocals and melodic instruments.” He comments on the changes in the perception of technology: “why is it that Queen felt the need to inscribe ‘no synthesisers were used in the making of this album’ on their early records and yet Brian May felt entirely comfortable constructing multiple layered performance ‘patchworks’ of guitar tracks? Why might the use of one type of technological mediation be considered more or less authentic than another?” This can be related to the influence of fashion on the quality perception of music mixes, where certain characteristics can be perceived as sounding more “authentic” than others. It is unclear whether the general preference of analogue over digital sound quality (“warmer”, “more authentic”) is a result of fashion or whether it can be related to low-level perceptual attributes.

Not only fashion determines how music is mixed. The purpose or emotional expression of a piece impacts on mixing techniques: Snoman [2009] states that some sounds need to be prioritized, depending on the piece and its composition. He claims that the individual style of
the mixing engineer makes the final product special. Owsinski [2006] writes that “the artist's vision” needs to “shine through” and that the mix should enhance “magical elements of a performance”. Kevin Shirley mixes for purpose, trying to create “proud”, present clear vocals and using e.g. dynamic interest to enhance the “emotion of the song” [Clark, 2011]. On Music Radar online, [2012], similarly vague statements are made: “Getting the right 'feel’ on a track is probably the single most important consideration when composing and mixing.” It is stated that the “feel” needs to be captured and the groove “enhanced” without specifying further what “feel” means. Izhaki [2008], too, writes: “a mix can, and should enhance the music, its mood, the emotions it entails, and the response it should incite”. Théberge [1997] notes that the mix is always part of the composition. Many more authors comment on this indefinable, artistic, “magical” “something” specific to the piece to be mixed that a good mixing engineer intuitively knows how to create.

Zagorski-Thomas [2007] provides an example of “mixing techniques that subvert, exaggerate, and/or parody conventions that have arisen for functional reasons”: The Flying Lizards’ 1979 single ‘Money’ is deliberately mixed to sound thin and weak to “support the mannered ineptitude and weakness of the recorded performances”. In this case, the mix is completely adapted to the message of the song. Music Radar [2012] claims that today’s music technology can often make tracks sound “samey and uninspiring” and that instead, songs should sound individual.

All in all, a variety of factors such as fashion and taste, as well as characteristics that are specific to individual pieces of music can have an impact on what is perceived as a good mix. This would make it harder to generalize values of a good mix or to even automate the mix process. However, as stated by Fisher [Wells, 2012] in an interview “there's a sort of acceptance window into which, once you have this skill and experience, your output will fall” which indicates that there can be several “correct” approaches to mixing a piece of music.

2.2.5 Freedom from technical faults

In this section, the extent to which individual sound source within the mix are free from technical errors is examined. In addition to the factors mentioned above, a good mixing engineer needs to compensate for “imperfect” musical performances or recordings. As summarized by Case [2007], there should be “no hum, pops, clicks or other stuff”. Numerous authors such as Snoman [2009] and Huber and Runstein [2005] state that there should be no audible noise. Dylan Dresdow says that there should be “no artefacts or horrid tuning” [Clark, 2011]. The aim to avoid “unnatural” sounding pitch and time correction is commented on in a number of sources such as the Computer music magazine [Music Radar, 2012].
engineers often need to “compensate for problems that have arisen during the recording process”, e.g. by removing unwanted resonances [Clark, 2011]. In the same way, Reed [2000] writes: “EQ is used to make sounds sound more natural and to compensate for characteristic sound of recording environment”. Bob Clearmountain, as quoted by Clarke [2011], aims to avoid vocal harshness. “Clean” vs. “distorted sound” also play an important role in pop music production audio quality in a study by Wilson and Fazenda [2016].

2.3 Is there any consensus on what is important in mix quality?

In the above paragraphs, mix quality parameters have been summarized. None of the individual sources consulted encompasses all of the above parameters and findings are usually structured differently between authors. Especially the “interest” and “impact” criterion is described in many different ways, often metaphorically, as mentioned above. These parameters, often coined as “impact”, “interest”, “magic” etc. by mixing engineers are usually underrepresented in academic literature, which tends to focus more on balance and separation. This may have contributed to the lack of standardized terminology in this area. Fenton and Wakefield [2011] claim that terms like “clear”, “defined”, “punchy” or “polished” do not allow for a consistent qualitative measure to be established. However, it has been shown that “interest” and “impact” are always sought after and merely described differently between authors.

In addition to the differing approaches of authors, there also appear to be mix quality parameters that are entirely context dependent. Sometimes it is specifically stated that the aesthetic values of a mix can be arbitrary or have not yet been examined scientifically. Case [2007] talks about the “highly nonlinear, unpredictable creation” of a mix. Reiss [2011] writes about compression: “Admittedly, such choices are often artistic decisions, but there are many technical tasks in the production process for which listening tests have not yet been performed to even establish whether a listener preference exists”. Børja comments on the producer's “vision” or “reference”: “This is by no means a stable, never-changing factor. It is under constant influence of many different impressions, physical and psychological, like mood, spirit and atmosphere, [...] and experience from past mixdowns. Normally, the final sound of a recording is the result of the musical taste of the record-producer, the recording engineer and the musicians who do the mixdown together and throw in ideas, likes and dislikes”.

All in all, no universal, complete set of parameters of a good mix has been established and it is unclear whether this could even be formulated. However, it is possible to derive a number of characteristics of high quality mixes from the literature. None of the sources consulted deemed
balance, clarity or separation unimportant or counterproductive to the creation of a good mix. On the contrary, they are widely discussed as important criteria of high quality mixes. The same is true for the “impact and interest” parameter. Authors seem to approach this similarly, merely using varied terminology. It is in particular possible to derive genre-specific sets of parameters from the literature.

In conclusion, qualities of good mixes seem to be highly complex and partially dependent on time, fashion, genre, taste and other such factors. However, the parameters balance, clarity/separation, and interest/impact are always part of the creation of a high quality mix. Therefore, it would make sense to research these parameters further. In the following section, clarity/separation is established as a useful starting point for further research.

2.4 Discussion

Despite the inconsistent and often incomplete presentation of mix quality attributes in the literature, a consensus exists on what makes a good mix. The parameters ‘clarity/separation’, ‘balance’ and ‘interest/impact’ are generally important, alongside the correction of ‘imperfect’ musical performances or recordings. Additionally, certain context specific parameters can determine mix quality. All high level perceptual attributes are laid out in Fig. 2-1.

![Fig. 2-1: the important parameters of high quality mixes](image)

Research on any of the parameters would help to measure the perceived quality of music mixes. However, they are not all equally suitable for the current research project, as explained in the following.

Context specific parameters are unsuitable for a study on psychoacoustics because they are difficult to generalize. Impact and interest are likely to depend on at least some of the other
High-level descriptive quality criteria and lower-level perceptual components of music mixes

Factors, i.e. the extent to which a mix grabs a listener's attention might depend on clarity/separation and balance. Therefore, it would be useful to investigate this after the other parameters are fully understood. Freedom from technical faults may also, at least to some degree, depend on context-specific parameters (e.g. distortion as stylistic device). The same applies to balance (e.g. 'lopsided' Beatles mixes). Lastly, it appears that several studies on tonal balance already exist (section 2.2.2).

The extent to which instruments can be heard in a mix ('clarity and separation') is likely to depend on acoustic and psychoacoustic factors only (e.g. masking and loudness). The appropriate degree of clarity/separation may still be context dependent. However, predictors for relative changes in clarity/separation may be established. In this way, it may be determined which version of a given mix sounds clearest. This would make it possible to e.g. measure or automatically adjust clarity/separation. Therefore, clarity/separation will be the focus of the remainder of this thesis.

2.5 Conclusion and next steps

This chapter aimed to establish high-level descriptive quality criteria and lower-level perceptual components of music mixes through a search of the current literature. The establishment of such parameters is necessary for the measurement of the perceived quality of mixed music. Despite the inconsistent and often incomplete presentation of mix quality attributes in the literature, a consensus exists on what makes a good mix. The parameters 'clarity/separation', 'balance' and 'interest/impact' are generally important, alongside the correction of 'imperfect' musical performances or recordings. These high level perceptual attributes are influenced by a variety of low-level attributes. Some of the proposed relationships between high-level and low-level perceptual attributes have been introduced above (Table 2-1).

Research on any of the parameters would help to measure the perceived quality of music mixes and therefore, the choice of which to tackle first is somewhat arbitrary. However, most of the parameters are at least in part context-dependent and therefore unsuitable for the current research project (balance, freedom from technical faults and context specific parameters). Impact/interest is likely to depend on clarity/separation and balance; which should be investigated first. The extent to which instruments can be heard in a mix ('clarity and separation') is likely to depend on acoustic and psychoacoustic factors only (e.g. masking and loudness) and will therefore be the focus of the remainder of this thesis.
3 Relating psychoacoustic findings to clarity and separation in music mixes

In the previous chapter, clarity and separation was found to be an important parameter of high quality music mixes. This was defined as the extent to which individual components can be heard in the mix. Further context-specific definitions will be given in each of the following sections. In order to measure and model the perceived clarity and separation of music mixes, the primary aim of this chapter is to establish the acoustic and mix parameters that clarity and separation are likely to depend on, by consulting literature on acoustics and psychoacoustics. Additionally, the following questions will be answered: is it likely that the factors influencing clarity and separation can be manipulated during the mix process? How do the lower-level parameters of clarity and separation presented in chapter 2 relate to the factors established in this chapter? Are there any characteristic parameters of clarity and separation that are presented across several areas of research? If so, this may offer a useful starting point for designing subsequent experiments.

Various starting points relating the clarity and separation of music mixes to lower-level perceptual and acoustic parameters can be found in the literature. In the current chapter, timbral clarity (section 3.1), concert hall clarity (section 3.2), auditory scene analysis (section 3.3), masking (section 3.4), loudness (section 3.5) and speech intelligibility (section 3.6) are investigated. Each section explores clarity and/or separation in the context of one of these research areas in order to establish its meaning in that particular area and the acoustic parameters which affect it. Section 3.7 is a discussion and section 3.8 concludes.

3.1 The timbral clarity of single sounds

The timbral clarity of single sounds has been investigated in previous studies. This is likely to be relevant for music mixes. Firstly, music mixes exist that feature only a single sound. The overall clarity of music mixes is also likely related to the timbral clarity factors presented here, since a mix could be considered as a single complex sound. Lastly, it is likely that the timbral clarity of individual sounds in a sound mixture can be adjusted with mix parameters. In this section, research questions will be addressed in the following order. Is there a definition for timbral clarity? What measurable acoustic parameters affect timbral clarity? Is timbral clarity related to brightness or other timbral characteristics?
3.1.1 Defining timbre and clarity

In order to give a more precise definition of timbral clarity, it will be attempted to define timbre and clarity in the following subsections.

Timbre

Currently, no all-encompassing definition exists for timbre. According to Letowski [1989], “Timbre is that attribute of auditory image in terms of which the listener judges the spectral character of sound. Timbre enables the listener to judge that two sounds which have, but do not have to have, the same spaciousness, loudness, pitch, and duration are dissimilar”. Plomp [1989] states that the ear is very sensitive to small shifts in timbre, although its exact dimensions are unknown. Grey [1977] describes timbre as multidimensional as opposed to the unidimensional parameters of pitch and loudness. He presents a three-dimensional scaling solution for timbre characterized by the variables of: spectral energy distribution; presence of synchronicity in the transients of the higher harmonics, along with the closely related amount of spectral fluctuation within the tone through time; and the presence of low-amplitude, high-frequency energy in the initial attack segment. Additionally, Grey finds a sensitivity of the ear to phase differences and Plomp [1976] notes that even pitch can influence timbre perception (lower simple tones sound duller than their higher counterparts which can sound sharper). Onset time, onset noise and steady state effects like modulation can also influence timbre.

Clarity

Similarly to the definition of timbre, the descriptive terms used to describe the timbres of instruments, such as bright, clear or dull are “blurry” [Plomp, 1976]. In this way, the meaning of clarity or clear may not be commonly agreed on and e.g. differ across languages. Kvist et al. [2004] define clarity as “natural timbre” and a “lack of distortion”. Lorho [2005] defines clarity as the opposite of “muffled”. No commonly agreed definition of timbral clarity was found in the literature. However, Disley and Howard [2004] found a common understanding of the terms “bright” and “clear” while attempting to establish a number of descriptive words used to describe the timbres of pipe organs. “Bright” and “clear” were also the most popular words. The experiment is summarized in more detail in section 3.1.2. Plomp [1989] used “clarity” as one of the timbral parameters for judging sung vowels, as the opposite of “dull”.

3.1.2 Acoustic factors affecting timbral clarity

In the following, the acoustic factors affecting timbral clarity will be presented. Contrary to common belief, there seems to be no specific lower mid frequency area that can be attenuated to increase clarity [Pestana and Reiss, 2014] but it appears that the balance between high and
Relating psychoacoustic findings to clarity and separation in music mixes

low frequencies plays an important role. However, researchers seem to disagree as to the direction of this balance.

More LF, less HF

In the context of sound synthesis, Ethington & Punch [1994] found an increase in timbral clarity in sounds with a decreased number of harmonics and a decreased spectral centroid. They state that pure tones with just a single fundamental frequency can sound particularly clear. No other literature was found that mentioned this relationship. Jeans [1937] suggests that clarity is related to the strength of the lower harmonics in a sound, in particular the even harmonics. He observes that exciting a string at a point a third along its length leads to a missing third harmonic while the second and fourth harmonic will be “fairly strong”, producing a clear and brilliant tone. In particular the second harmonic adds clearness and brilliance according to Jeans. Again, no other sources seem to confirm this, and it appears that Jeans has not undertaken any formal studies to confirm his claim. In the context of string clarity, Dünnwald [1991] found that clarity was associated with a lower level between 4200Hz and 6400Hz.

More HF, less LF

Disley, Howard & Hunt [2006; 2004] found “clear” to exhibit a significant positive correlation with the spectral centroid. This was the case in the context of pipe organs [Disley and Howard, 2004]. In this study, ninety-nine English-speaking subjects were asked to describe recordings of pipe organs with timbral adjectives. Out of the elicited terms, seven useful, unambiguous and frequently used words were chosen, including “clear”. Subjects then rated the pipe organ recordings in terms of these adjectives. Subsequently, the following spectral correlates of words were examined: the average spectral centroid, its fall-off and consistency; the average harmonic strength, its fall-off and consistency; average spectral smoothness and its standard deviation, the average spectral slope and its standard deviation, and the average inter-quartile spectral slope and its standard deviation. The Pearson correlation coefficients between these spectral analyses and the timbral adjectives were calculated. The Pearson correlation coefficient for pipe organ clarity and spectral centroid was 0.955 [Disley & Howard, 2004]. As opposed to Dünnwald, Fritz et al. [2012] found violin clarity to correlate with an increase in level between 1520Hz and 6080Hz.

Solomon [1959] assesses the relationship between energy concentration in eight octave bands and the rating of clarity as given by test subjects. Similarly to Disley et al., his findings indicate that a higher spectral centroid may lead to higher timbral clarity. The result of his experiments can be seen in Fig. 3-1. According to this graph, it could be concluded that by boosting the frequency region above 2.4kHz or by reducing the 150-600Hz area of a sound through
equalization its clarity may be increased. Izhaki [2008] points out that frequencies above 2kHz stand for clarity, definition and crispness and that an excess in lower frequencies can lead to “muddiness” which seems to confirm this assumption. Similarly, clarity is achieved by HF boosts according to Savage [2014]. For overall mixes, there seems to be a preferred region for the spectral centroid, centred at around 2.9kHz [Wilson and Fazenda, 2015]. It seems clear that the spectral centroid has an influence on the perceived clarity of sounds, even if the exact relationship is disputed.

Relation to brightness

Occasionally, authors comment on the relationship between brightness and clarity, e.g. Disley and Howard [2003]. Pipe organ sounds that were perceived as bright were usually also perceived as clear. Zacharakis et al. [2012] assessed the semantic description of musical timbre in a verbal attribute magnitude estimation test. They attempted to find the underlying structure and dimensions of timbral descriptions through cluster and factor analysis techniques. In the resulting diagram, descriptors like “brilliant”, “bright” and “distinct” are close to “clear”, whereas “rough”, “dull”, “warm” or “full” are quite far away. Disley et al. [2006] came to similar conclusions in a similar experiment.

Like clarity, brightness is affected by spectral centroid. For example, the Pearson correlation coefficient was 0.999 for pipe organs [Disley and Howard, 2004]. Terasawa et al. [2012] have observed a strong correlation between the spectral centroid and the perceived brightness of a sound and Plomp [1976] describes brightness as a function of location of centre of energy distribution on the frequency continuum.

Conversely, as mentioned above, Disley et al. [2004] observed the opposite in another paper, where bright and clear showed significant negative correlation. Ethington and Punch [1994]
also find differences between brightness and clarity. Although both correlate with a reduced harmonic density, the overall harmonic content and slope of relative weight need to be increased for a brighter timbre but decreased for a clearer timbre. All in all, clarity and brightness are similar in some ways but not identical. This may partially be because of the blurred meaning of “clarity” and “brightness”, as well as uncertain definitions of some of the acoustic parameters like the spectral slope, as mentioned above.

Overall, it appears that in more studies, clarity is linked to a larger amount of HF energy and a lower amount of LF energy than vice versa. The studies supporting the latter may deviate from the former because they were focusing on very specific contexts (sound synthesis, violin clarity). Another reason may be differences in the authors’ interpretations of clarity or some of the acoustic parameters. Both were not always provided.

### 3.1.3 Spacing of harmonics

Plomp [1976] found the spacing of harmonics to be responsible for the timbral dissimilarity of sounds with different pitch but similar spectral envelopes. Ethington and Punch [1994] note that the harmonic spacing should be expanded for a clearer sound. Kvist et al. [2004] found a positive correlation between the unevenness of the treble frequency response (in addition to the decay time of the midrange impulse response), which might be related to this. However, this was only tested on one listener. The spacing of harmonics in a mix could only be increased by removing harmonic content as changing the frequency of overtones would completely change the timbre and in most cases lead to inharmonic sounds.

### 3.1.4 Conclusion

The present section summarized findings on timbral clarity. The following research questions can now be answered: is there a definition for timbral clarity? What measurable acoustic parameters affect timbral clarity? Is timbral clarity related to brightness or other timbral characteristics?

Although the definitions of *timbre* and *clarity* are “blurry”, there seems to be a common agreement as to which sounds are timbrally clear. Several acoustic parameters seem to influence timbral clarity that nearly all relate to the balance of HF and LF energy: the number of harmonics, spectral centroid, relative level of low-order even harmonics, spacing of harmonics, and spectral slope.

Most sources support the hypothesis that clarity is linked to a greater amount of HF energy and a smaller amount of LF energy. The studies that were not in accordance with this may have been
based on different definitions of clarity and the acoustic parameters tested. Both were not always provided. At the same time, clarity was always assessed in specific contexts (string clarity, sound synthesis). Clarity was also found to be related to brightness, having some acoustic parameters in common.

All parameters found are related to the overtone spectrum of sounds. Mix tools able to manipulate the spectrum of sounds are equalizers and filters, harmonic exciters, denoisers, distortion plugins and experimental plugins such as formant shifters. It is possible that some of these are able to influence the perceived clarity of single sounds or overall mixes. An experiment where single sounds are manipulated with the above mixing tools could be useful to learn more about the exact relationships.

3.2 Relating clarity in concert halls to clarity in mixes

The following section gives an overview of findings on clarity and separation in concert halls. The aims are: to determine what clarity is in the context of concert halls, what factors can affect it, how it can be measured and what the limitations of current measurements might be. Furthermore, it will be established whether perceptual optima exist for clarity and whether maximum clarity is desirable.

3.2.1 What is clarity and what sonic factors affect it?

In the context of concert halls, clarity describes the audibility of elements in musical performances and their spatial and temporal separation. Beranek [1996] distinguishes horizontal definition (the degree to which sounds played in succession stay apart) and vertical definition (the degree to which simultaneous sounds stay apart). Clarity can be impaired by reverberation, echoes, noise, tonal distortion, sympathetic ringing tones and non-uniformities of listening conditions [Beranek, 1996]. Non-uniformities are the uneven distribution of sound throughout a concert hall. Beranek states that these effects can only detract from the beauty of the music and are therefore undesirable in concert halls. Although some of these effects can impair clarity, reverberation, echoes, noise and distortion are often added artificially to mixes to create interest. It is therefore useful to discuss them in this chapter; it will be established to what extent echoes and noise can impair clarity by reviewing the relevant literature. The impact of tonal distortion and resonances in concert halls on clarity can be compared to EQ in mixes and relates to the timbral clarity of sound sources. This was covered in detail in the section about timbral clarity (section 3.1).
3 Relating psychoacoustic findings to clarity and separation in music mixes

3.2.2 Echoes

Although discernable, single echoes can be undesirable in concert halls but they are often added to sounds in a mix. The main difference is that echoes and delay effects in mixes can be adjusted precisely to the tempo of the piece. The timbre, directions and loudness of echoes can be adapted flexibly to the context. Echoes can, however, impair clarity under certain circumstances which should be avoided in mixes. Haas [1972] studied the influence of a single echo on the intelligibility of speech which is a relevant source in this context. His findings are briefly summarized in the following.

When an echo is delayed from the original sound by up to 30ms, an increase in loudness and a “pleasant broadening” of the original sound source is perceived and clarity is not affected. The echo is not heard as a separate signal unless it is at least 10dB louder than the original. (The latter case would usually not be applied in mixes as it may disturb the perception of rhythm.) Greater delay times cause speech to be disturbed and the “critical delay difference” value is inversely proportional to the speed of speech in the range of 3.5 – 7.4 syllables per second. The intensity of the echo plays an important role in the critical delay difference: an attenuation of the echo by 5 dB doubles the critical delay difference. Echo intensities more than 10dB below the original sound source do not impair the intelligibility of speech at all.

Interestingly, Haas found that the high frequencies of the echoes determine the amount of subjective disturbances, which may correspond to Solomon’s findings above. An attenuation of higher frequencies raises the critical delay difference. Haas found that the quantity of the echo disturbance does not depend on the loudness in the range belonging to speech. Lastly, the direction of incidence of the echo does not affect the critical difference as long as the direct sound comes from the front. This may indicate that panning of echoes in a delay effect has little effect on the clarity of the original sound. Lastly, if the room has a longer reverberation time, a greater critical delay difference can be observed, showing an interaction between reverberation and delays.

3.2.3 Noise

Beranek [1996] summarizes the maximum amount of acceptable noise in concert halls and opera houses in the graph below (fig. 3-2). The acceptable noise level is different for each octave band and decreases towards lower frequencies. Curves for the recommended limit for the noise (NCB-10) and the limit above which noise levels would be seriously disturbing (NCB-15) are shown. The influence of noise on clarity will be covered in more detail in the section covering the impact of masking on clarity and separation (section 3.4).
3 Relating psychoacoustic findings to clarity and separation in music mixes

Fig. 3-2: the range of acceptable noise levels in concert halls and opera houses, where NCB-10 is the recommended limit for the noise and NCB-15 represents the limit above which noise levels would be seriously disturbing [Beranek, 1996]

3.2.4 Measurement & the ISO standard

The most commonly used measures of perceived clarity, definition of speech and reverberance in concert halls are summarized in the International Standard [ISO 3382-1, 2009]. These are the $C_{80}$, $C_{50}$ and $D_{50}$ and the EDT (early decay time). The first three are measures of balance between the early and late arriving energy which can be calculated for either a 50ms or an 80ms early time limit. The $C_{80}$ is often used for music, whereas the other two are more useful in the context of speech. The sensitivity of the $C_{80}$ for speech signals can be maximized by averaging the $C_{80}$ of the frequency bands most relevant to speech cues: 500Hz, 1kHz and 2kHz [Beranek, 1996]. The $C_{80}$ correlates highly with the reverberation time and early decay time. Table 3-1 gives the equations and typical ranges of the $C_{80}$, $C_{50}$ $D_{50}$ and EDT as provided in the British Standards.
3 Relating psychoacoustic findings to clarity and separation in music mixes

<table>
<thead>
<tr>
<th>Subjective listener aspect</th>
<th>Acoustic quantity</th>
<th>Equation</th>
<th>Just noticeable difference</th>
<th>Typical range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceived clarity of sound (music)</td>
<td>$C_80$, measured in dB</td>
<td>$C_{80} = 10 \log \frac{\int_0^{80} p^2(t)dt}{\int_0^\infty p^2(t)dt}$ dB (3.1)</td>
<td>1 dB</td>
<td>-5 dB; +5 dB</td>
</tr>
<tr>
<td>Perceived definition of sound, often for speech</td>
<td>$D_{50}$</td>
<td>$D_{50} = \frac{\int_0^{0.05} p^2(t)dt}{\int_0^\infty p^2(t)dt}$ dB (3.2)</td>
<td>0.05</td>
<td>0.3; 0.7</td>
</tr>
<tr>
<td>Perceived clarity of sound (speech)</td>
<td>$C_{50}$, in dB</td>
<td>$C_{50} = 10 \log \frac{\int_0^{50} p^2(t)dt}{\int_0^\infty p^2(t)dt}$ dB (3.3)</td>
<td>1.1 dB</td>
<td>-3.7 dB; 3.7 dB</td>
</tr>
<tr>
<td>Perceived reverberance</td>
<td>EDT (early decay time), in seconds</td>
<td>(the decay time, measured over the first 10 dB of decay)</td>
<td>Rel. 5%</td>
<td>1s; 3s</td>
</tr>
</tbody>
</table>

Table 3-1: equations and typical ranges of the $C_{80}$, $C_{50}$, $D_{50}$ and EDT as provided in the British Standards [ISO 3382-1, 2009].

3.2.5 Possible limitations of measurement methods

Griesinger [2013] claims that most common ISO 3382 measures fail to predict sound clarity precisely. Firstly, he stresses the fact that unclear sound, in particular speech, can often be understood according to the ISO standards, but is more difficult to pay attention to and remember than clear sound. He assumes that this also plays an important role in music, as musical phrases may not only need to be heard clearly, but also interpreted and remembered.

In addition, Griesinger [2013] suggests that the ear might be sensitive to the sharp peaks created by the phase coherence of the upper harmonics in tones with definite pitches, rather than the sound power measured by e.g. the $C_{80}$. He claims that these peaks can be found in the acoustic pressure at the fundamental frequency of the sound but become blurred when reverberation is present, leading to an increase in muddiness and reduced clarity. Griesinger assumes that the ear can only detect these peaks when there are two or more harmonics at the same time within one critical band. As critical bands have approximately a 1/3rd octave width, the peaks are only detectable above about 1500Hz for male voices and 2500Hz for female voices according to Griesinger. This would need to be critically tested through experimentation.

Griesinger [2013] concludes that “the information content of speech and much music, the sense of sonic distance, the perception of clarity, the ability to sharply localize instruments, and the
ability to separate multiple sound streams all depend critically on the harmonics above 1000Hz from sounds with a definite pitch". Further study would be required to verify these claims.

The relationship of higher frequencies with timbral clarity has also been assessed in the section on timbral clarity (section 3.1). According to Solomon [1959], an energy concentration in the frequency region above 2.4kHz can make sounds appear clearer. This may indicate a potential similarity between timbral clarity and clarity in concert halls. Griesinger [2013] notes that binaural localization of low-passed music and speech in reverberant environments is very difficult. He notes: “If we study the threshold of localization of male speech as a function of the direct to reverberant ratio (D/R) in octave bands we find the threshold drops 6dB/octave as frequency rises, up to about 1000Hz, and then holds constant”. In contrast, the auditory system can easily localize sounds in anechoic chambers by utilizing the ITD for low frequencies and the ILD for higher frequencies.

Further studies have shown the limitations of the clarity metrics presented above. Imamura et al. tested the relationship between arrival direction and delay time of the first reflections and perceived clarity of reproduced sound. The variation of the first reflections did significantly influence perceived clarity and spatial impressions, demonstrating the limitations of the C80 and other methods presented earlier [Imamura et al., 2014]. Panton et al. [2015] investigated concert hall acoustics from musicians’ perspective. They found no correlation between clarity and reverberance (r=0.15/0.05). Similarly, Lokki [2014] states that ISO3382-1 cannot explain the details of subjective perception nor preferences of the listeners.

### 3.2.6 Are there any perceptual optima of clarity?

Certain perceptual optima seem to exist for clarity and separation, where too much clarity can have a negative effect which, again, is not covered by e.g. the C80. Griesinger [2013] points out that halls need a "cushion of reverberation". Pleased with the effect of electronically added artificial reverberation to the Wintergarden of the Elgin theatre in Toronto, he explains that adding some extra late energy can make the sound more "beautiful", even if it reduces clarity slightly. He also explains that although listeners may choose clarity over added reverberation in an A/B comparison test, a slight amount of reverberation is usually preferred to a dry space. In the context of music mixes, artificial reverberation is necessary for the recreation of "the richness of the room effect" of realistic spaces, as stated by Everest [2009]. The amount of desirable reverberation is context dependent as discussed in the following paragraphs.

According to Beranek [1996], Clarity in concert halls is the “degree to which the discrete sounds in a musical performance stand apart from one another”. He states that clarity depends critically
Relating psychoacoustic findings to clarity and separation in music mixes

on musical factors and the skill and intention of the performers, as well as the acoustics of the room. He distinguishes two kinds of definition, as mentioned above: horizontal definition (the degree to which sounds played in succession stay apart) and vertical definition (the degree to which simultaneous sounds stay apart). They depend on the composition, room acoustics, performers and the auditory acuteness of the listeners. Both can be relevant to the mix process.

Beranek [1996] describes horizontal definition as opposed to fullness in tone. He points out that while Gregorian chant should be quite blurry and full in tone, other genres may require greater separation. The same is the case in the context of vertical definition, where faster passages can get blurred more easily through reverberation than slower passages. Similarly, louder sounds are blurred less easily in reverberant spaces than quiet sounds [Beranek, 1996]. Griesinger [2013] comes to a similar conclusion, stressing the importance of the interaction between held notes and the space. He compares the acoustics in a church in Cambridge before and after the acoustic plaster is covered with dusky gold paint. Whereas the addition of the paint has a positive effect on the sound of the organ, it has a negative effect on chamber music. This was not only due to the location of the instruments (whilst the organ is located flat against one of the walls of the crossing projecting into the space, chamber music is performed from the centre of the crossing) but also the nature of the music played. Griesinger points out that the loudness of reflections builds up with time and that multiple late reflections dominate in the church due to the new reverberation time of 2.5 seconds at 500Hz. Due to this, the long, held notes played by the French horn in “Serenade for Wind Instruments” by Dvořák were much more audible than shorter notes, as the latter did not excite audible reverberation in the space. Griesinger adds that in halls, it is the onset of sounds that determine clarity, as during a held note the direct component is nearly always inaudible. In this way, if sound decay masks sound onsets both clarity and intelligibility can suffer. These findings can be applied to the mix process, where the effect of reverberation settings on clarity and relevant perceptual optima depend on the genre, sounds and the nature of the piece.

Beranek [1995] quotes Barron, who found that preference varied between listeners in an experiment: some listeners preferred the presence and intimacy of music performed in a dry space, while others preferred the fullness of increased reverb. Analogously, personal taste and fashion also seem to influence the amount of reverberation used in mixes. Informal listening reveals that in the 1980’s for instance, more reverberation was applied to mixes than nowadays. In conclusion, the perceptual optima of clarity are unclear and seem to depend on a number of factors, although it was shown that the maximum amount of clarity possible is usually not desirable.
3.2.7 Conclusion

The aims of this section were to define clarity is in the context of concert halls and to establish what factors can affect it, how it can be measured and what the limitations of current measurements might be. Another aim was to establish whether perceptual optima exist for clarity and whether maximum clarity is desirable.

Clarity can be described as the audibility of individual elements in musical performances and their separation. This depends on the composition, room acoustics, performers and the auditory acuteness of the listeners. Vertical clarity is the extent to which simultaneous sounds stand apart and horizontal clarity is the extent to which sounds played in succession stand apart. Both can be impaired by reverberation, echoes, noise, tonal distortion, sympathetic ringing tones and non-uniformities of listening conditions. According to the findings of several authors, it seems that high frequencies in particular contribute to the perception of clarity, also in the context of reverberation.

Clarity in concert halls is usually measured using the $C_{80}$, $C_{50}$ and $D_{50}$ and the EDT. These methods ignore how incoming sound is processed by human listeners, however. At the same time, perceptually important acoustic factors such as the arrival time and direction of early reflections are not taken into consideration. Although it impairs clarity, reverberation is usually perceived as pleasant and preferred over a clearer, dry sound. Hence, certain perceptual optima exist for clarity in concert halls. The combination of factors determining these is highly complex and not necessarily clearly defined, as the factors depend on the context (genre, fashion, taste, etc). The maximum amount of clarity is not always desirable.

Reverberation and delay effects are commonly added in the mix process and, by increasing the level of late sound, are likely to reduce clarity, both in terms of $C_{80}$ and, for delay, in terms of lyric intelligibility. However, the resulting sound may be perceived as more aesthetically pleasing. The findings in this chapter can be used as initial step towards finding perceptual optima of clarity in mixes.

3.3 Relating Auditory Scene Analysis to separation in mixes

The perception of clarity and separation in mixes (the extent to which individual sounds can be heard) is largely determined by the auditory system’s ability to group or segregate sonic elements into streams. Therefore, according to Woszczyk and Bregman [2005], ASA principles might supply a framework upon which the recording engineer’s craft could be systematized. Hence, by reviewing the current literature, the following questions will be answered in this
section: what is auditory scene analysis? What factors influence the grouping or segregation of sonic elements? Does attention play a role in auditory scene analysis? How does this relate to music and mixing? In section 3.3.1, auditory scene analysis will be defined. Sections 3.3.2—3.3.6 give an overview of the factors influencing how the sound reaching the ears is analyzed, structured into processes of primitive and schema-based grouping. In section 3.3.7, the influence of attention is discussed and in section 3.3.8, auditory scene analysis is related to music and mixing. Section 3.3.9 concludes.

3.3.1 What is auditory scene analysis?

Auditory scene analysis (ASA) is the process of forming mental representations of individual sound sources from the summed waveforms that reach the ears. This process consists of the following two conceptual stages [Bregman, 2008, 2007, Wrigley & Brown, 2001]. First, the auditory system divides the input into its constituent atomic units, i.e. packages of acoustic evidence (segmentation). Following segmentation, any packages that appear to have arisen from the same source are either grouped (to form a stream for a given source) or segregated (to form separate streams for different sources). In this way, parts of the neural spectrogram are combined into auditory streams. Elements that fall in the same auditory stream are perceived as stemming from the same sound source.

Bregman [1990] describes the mental processes responsible for the grouping or segregation of elements as complementary with the structure of the surrounding world, basing his findings on concepts from Gestalt psychology, computer modeling, syntactic theory and physiological explanations, using heuristics to establish functional (rather than physiological) explanations for auditory source separation. The author explains that stream segregation is context-dependent, involving the competition of alternative organizations.

Bregman [2008] distinguishes sequential and spectral grouping, the first being the grouping of consecutive elements and the latter being the grouping of simultaneous elements. Furthermore, he distinguishes primitive and schema-based grouping. Primitive grouping follows innate constraints, whereas schema-based grouping follows learned ones. All four categories will be discussed in this section. As stated by Bregman [1990], the exact properties that determine grouping are not known and the findings summarized in this section cannot be easily generalized. Different people resolve auditory illusions differently, and most findings stem from simple experiments in controlled environments, using tones with little complexity, such as pure tones. Broad areas of ASA have not been investigated yet. However, current literature offers useful starting points for investigating clarity and separation in mixes, e.g. [Bregman and Woszczyk, 2005].
3.3.2 Primitive Grouping

Bregman [1990] defines primitive grouping as an innate processes of ASA that are not influenced by learned constraints. Primitive grouping processes do not attempt to predict the future position of a sound in frequency or time and they do not make use of the fact that a sequence may be governed by rules, such as the trajectory of a rising or falling pure tone. Instead, they use the factors outlined in the following sections for ASA.

As the strategies for stream segregation can be unreliable, the different cues both collaborate and compete to control the grouping and different cues have different strengths as explained by Bregman [2008]. When cues compete, it can be difficult to predict which sounds will group as some of the factors are hard to quantify; apparent spatial separation, for example, depends on many parameters such as the timbre, loudness, pitch, number of harmonics and can even be a result of stream segregation itself, when some of a sound’s energy becomes grouped with other, spatially separate sounds. Bregman states that usually, failure in inclusion is a better way to predict stream segregation than success in exclusion. When a sonic element features some of the factors promoting segregation, exclusion is more likely to occur than grouping in the opposite case [Bregman, 1990].

Although sequential and spectral grouping are separate phenomena, they can also collaborate. Before both will be discussed in more detail, the Gestalt principles will be introduced, as they are relevant for both.

3.3.3 Gestalt principles

Gestalt psychology explores the perceptual ability to acquire and maintain stable percepts in a noisy environment, analyzing complex interactions among various stimuli. Bregman [1990] relates some of the Gestalt principles originally used to describe visual perception to the perceptual organization of sound. These will be introduced in the following.

Gestalt principle of similarity

Moore [2012] states that sonic elements that are similar in timbre, pitch, loudness or subjective location are likely to be grouped.

Gestalt principle of common fate

The Gestalt principle of common fate occurs when the different frequency components arising from a single sound source vary in a coherent way, as pointed out by Moore [2012]. They start, finish and change in intensity and frequency together. Partials are grouped if they feature synchronised amplitude changes or pitch modulation patterns. Bregman [1990] distinguishes
three types of modulation: the small pitch and amplitude fluctuations present in all natural sounds, musical vibrato and the slow pitch changes present in e.g. portamento. At least 1% of vibrato is always present in natural instruments and singing. Even if parts of a vibrating sound are panned or separated in any other way, they can still group. In mixing this can be an advantage, as the stereo image of one sound can be widened by panning some of the frequency regions without causing these to fall into separate streams. As mentioned above, opera singers use vibrato to be heard over the orchestra and various types of modulation can be used in sound synthesis to create separation between sounds. This is similar in the context of masking: when a masked tone is grouped to another not masked tone, this can counteract the masking (comodulation masking release).

Bregman and Woszczyk [2005] provide insight into the relevance of these findings to the mix process: ”several sources or the entire mix can be compressed in amplitude by a compressor that imposes common dynamic changes. The result is always increased blend and interdependence of sounds subjected to the commonality of motion.” They present further techniques of creating “common fate” between sounds in a mix, such as pitch modulation and Doppler modulation, an effect that can be achieved by reproducing sources via a rotating Leslie loudspeaker or its digital emulation. Here, “The common spectral side-bands created by modulation are derived from the individual spectra and thus bind the individual sound together” [Bregman and Woszczyk, 2005].

**Gestalt principle of proximity**

Bregman [1990] relates the Gestalt principle of grouping by proximity, where visual elements that are spatially close together are more likely to be grouped, to ASA mechanisms. Proximity can be in the dimension of frequency or in time.

**Gestalt principle of closure**

According to the Gestalt principle of closure, a continuous sound can form an auditory stream even if it is interrupted by another sound. This is particularly relevant for sequential grouping, where subsequent elements of the spectrum are grouped into one stream. According to Gestalt psychologists, the fragments organise themselves more strongly if information for occlusion has been added, as shown in Fig. 3-3 [Bregman, 1990]. Bregman relates this to auditory streams, where the illusion of continuity works best if the gap in the target sound has been filled by a stimulus that could have masked the target sound. The masker does not need to contain all frequencies present in the target but both sounds need to stimulate similar neural activity. Bregman [1990] gives the example of a sound rising and falling in pitch, interrupted by white noise bursts that mask the sound. Here, the sound is still perceived as continuous. In the same
3 Relating psychoacoustic findings to clarity and separation in music mixes

way, a vocal in a mix could still be perceived as continuous auditory stream even if it is occasionally masked by a snare.

Bregman notes that the target sound could even move or change during the interruption and will be perceived as continuous as long as the fragments share certain attributes. If the target sound is attenuated before and after the break or changes drastically in some other way, the continuous sound may fall into separate streams. The Gestalt principle of closure is similar, but not identical, to the continuity illusion explained below.

Fig. 3-3: fragments organize themselves more strongly if information for occlusion has been added. On the right, this principle is shown in the context of pure tone glides interrupted by noise bursts [Bregman, 1990].

Gestalt Principle of exclusive allocation

A contour separating two visual elements is usually assigned to only one of the elements, as illustrated in figure 2. Bregman [1990] relates this description of a visual perceptual phenomenon to auditory scene analysis: in many cases, one sonic element can only belong to one auditory stream. This applies especially to sequential grouping (section 3.3.4), where sonic elements played in succession are allocated to different streams. Bregman [1990] describes this phenomenon as "exclusive allocation" and Moore calls it "Disjoint allocation" [Moore, 2012].

However, visual and auditory stimuli are perceptually processed differently which can make it difficult to adopt principles of visual perception to auditory scene analysis. In visual perception, light emitted or reflected by an object behind another object does not physically reach the eye unless it is reflected again by another, not hidden object. Sonic elements, on the other hand could be described as "transparent", as the sound emitted from an object that is visually hidden can still reach the eardrums. Sonic energy can, however, be occluded by auditory processes such as masking. Due to these differences between the perceptual domains, the Gestalt Principle of Exclusive Allocation cannot always be transferred well to ASA.

In spectral grouping for example (the grouping of simultaneous elements of the spectrum, section 3.3.5), there are exceptions to the principle of exclusive allocation. Although spectral components are usually interpreted as belonging to only one sound (Bregman, 1990, Fig. 3-4), they can sometimes be used to derive properties of several sounds (duplex perception, Fig. 3-5).
3 Relating psychoacoustic findings to clarity and separation in music mixes

When exclusive allocation is violated, it is ambiguous where a sonic element belongs. At the same time, as pointed out by Gestalt Psychologists, the properties of an object are sometimes not just derived from the local attributes of that object but also other objects (Fig. 3-3). Bregman [1990] gives the example of vocal formants that are presented as two separate sounds. Here, the sounds are segregated but the speech sound can still be perceived. In this way, a sonic component can have two identities at the same time.

Arguably, the principle of exclusive allocation and its exceptions cannot always explain the perception of sounds parts of which have been masked. Often, spectral components present in the masker are interpreted as part of the masker but still contribute to the perception of the target sound. Here, the target sound is interpreted as containing the masked spectral components because they exist in the masker.

Fig. 3-4: a visual example of "belongingness". The highlighted contour seems to belong to the dark shape [Bregman, 1990].

Fig. 3-5: a visual example of duplex perception: One of the parallel lines in the rectangle is also one of the sides of the square inside the circle [Bregman, 1990].

Gestalt Principle of Good Continuation

According to Moore [2012], the Gestalt principle of good continuation can be applied to sound perception: smooth changes in frequency, intensity and location do not impair the perception of
3 Relating psychoacoustic findings to clarity and separation in music mixes

A continuous auditory stream, whereas abrupt changes indicate a change in source. A continuous auditory stream can be split into a number of substreams, as demonstrated by Bregman and Crum [2006]. The authors explain that this formation of smaller units in a longer period of a changing sound enables listeners to recognize a change in the timbre of the sound faster than when the sound appears to change continuously. They conclude that the formation of units causes the timbral difference between parts of the overall transition to be more pronounced when those parts are packaged in separate units.

The figure-Ground Phenomenon

Moore [2012] points out that attention is usually directed to one auditory stream at a time and that other streams form a kind of background. The Gestalt psychologists call this separation into attended and unattended streams the “figure-ground phenomenon” [Moore, 2012]. Here, certain parts of the spectrum can be selected for conscious analysis. At the same time, attention may also influence the formation of streams itself. The role of attention in auditory scene analysis will be discussed in more detail in section 3.3.7.

3.3.4 Sequential grouping

In sequential grouping, subsequent elements in the spectrum are perceived as belonging to the same auditory stream. Here, an increased biasing toward forming a distinct stream builds up with longer exposure to sounds in the same frequency region, as pointed out by Bregman [2008]; alternating high and low tones take about 4 seconds before segregating into two separate streams for example. Elements that are similar, such as the successive sounds played by an instrument sharing the same instrument timbre, fall into the same auditory stream. Bregman states that complex sounds, like vowels spoken by different voices can resemble each other in many different ways but it is not completely clear how. At the same time, as introduced in the context of the Gestalt principle of good continuation, streams can undergo slight changes in e.g. timbre over time and still stay coherent.

As explained by Bregman [2008], “the perceptual segregation of different subsets of sounds in a sequence into separate streams depends upon differences in their frequencies, pitches, timbres (spectral envelopes), center frequencies (of noise bands), amplitudes, and locations, and upon the suddenness of the changes of these variables from one sound to the next.” The duration of silence between sounds can also influence segregation. All of these cues can be manipulated in the mix process. In the following, spatial segregation and timbre will be discussed in further detail.
3 Relating psychoacoustic findings to clarity and separation in music mixes

Spatial segregation

When sounds are spatially separate, they are less likely to fall into the same auditory stream. When a sound rapidly changes location, it can be harder to understand, as shown in the context of speech in an experiment by Bregman [1990]. Similarly, alternating clicks seem slower and are harder to count [Bregman, 1990]. According to Izhaki [2008], sounds in a mix can appear clearer and more defined at the extremes than in the middle.

In most cases, spatial separation alone is not a strong cue for segregation [Bregman, 1990]. When two voices alternate between both ears it is difficult to follow the speech at one ear, rather than the speech of one of the voices, as each voice falls into its own auditory stream. This makes it possible to automate pan pot settings in a mix without losing clarity. The perception of spatial separation itself depends on auditory stream segregation, as mentioned earlier. The influence of spatial segregation on spectral grouping will be discussed below.

Timbre

The timbres of sounds influence whether they fall into the same auditory stream. Cusack and Roberts [2000] assessed the effect of timbre on sequential grouping. They found that target sounds can be selected from distractors in the same spectral region more easily when they differ in timbre. Differences in timbre impaired performance, indicating the occurrence of primitive stream segregation. Similarly, Iverson [1995] investigated auditory stream segregation by musical timbre which appeared to be influenced by "gross differences in static spectra and by dynamic attributes, including attack duration and spectral flux". Bregman [1990] states that although scene analysis uses timbre as a cue for source segregation, it is also responsible for the computation of timbre. This is why timbre is not discussed further as a cue for spectral grouping in this section.

It is not known whether there is a limited number of dimensions of timbre and whether there might be any metameric timbres (very similar timbres with no obvious shared physical properties). Bregman [1990] names the dimensions of the brightness of the spectrum, the bite of the attack and the simplicity of behaviour of the harmonics over time. The first two can be manipulated with compressors and equalizers in mixes. Further aspects of timbre could e.g. be the lowest two formants in a sound, like in vowels. It can be difficult to predict when timbres blend and Bregman [1990] assumes that ‘textural’ (fine structural) features like crunchiness may play a role in source separation. In that way, distortion effects may create separation in mixes. According to Bregman [1990], sounds of the same brightness (as measured by their spectral centroid) are often assigned to one stream. The impact of brightness on clarity has also
been investigated in some of the other sections. Noises also tend to be segregated from pure tones.

### 3.3.5 Spectral grouping

Spectral grouping is the grouping of elements of the spectrum that are heard simultaneously. This is particularly relevant to the mix process, where separation needs to exist between sounds that are played at once. Audible simultaneous parts of the spectrum are either grouped or segregated. When they are grouped, they become part of the same auditory stream and lose their individual identities. When they are segregated, they are treated as separate entities. In masking, a target sound or parts of a target sound become inaudible when another, usually louder sound is played at the same time. Both fusion and masking can compromise the audibility and therefore clarity of a target sound, as they prevent the target sound from being an alone standing, separated unit. Mixing techniques are often applied to avoid both for that reason.

According to Bregman [2008], “the perceptual fusion of simultaneous components to form a single perceived sound depends on their onset and offset synchrony, frequency separation, regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation, and parallel gliding of components”. He explains that composers create separation between soloists and other instruments intuitively by using differing pitches and different frequency bands. Soloists employ vibrato and singers can create a vocal formant making them stand out of the mix. As indicated above, the onsets of sounds can also be used to create separation by playing in a rubato style. Bregman and Woszczyk [2005] state that a group of similar instruments, such as electric guitars, can be blended into an ensemble “when their individual envelopes are trimmed into synchrony using gates or keyed (synchronous) expanders.” Bregman [1990] adds that a soloist sounds particularly separated from the accompanying instruments if the latter are as different in the above variables from the soloist as possible. The relevant cues for spectral grouping will be discussed further in the following.

**Common fate**

The Gestalt principle of common fate has been introduced in section 3.3.3. Elements of a sound scene are grouped if they feature synchronised amplitude changes or pitch modulation patterns.

**Old plus new heuristic**

Bregman [2008] introduces the “old plus new heuristic” as an important organisational principle in primitive grouping processes. Here, the auditory system tries to interpret a spectrum that has become more complex or some of whose parts have become more intense as
Relating psychoacoustic findings to clarity and separation in music mixes

continuing old spectrum with new, added components. When the necessary cues to support this interpretation are present, the auditory system can subtract the "old" components from the mixture and assess the newly added partials separately. It is assumed that as a result, the segregation experience in mixes at a given time, and the resulting clarity and separation are not only dependent on what is currently heard, but also what was heard just before.

Onsets

Bregman [2006] further relates the "old plus new heuristic" to the fact that the onset of a new sound plays an important role in stream segregation. Phillips et al. [2002] researched the effect of temporal asymmetries on auditory perception and found evidence that onset times can impact on auditory segregation, although this was not the case for the offsets.

Sudden rises in intensity usually indicate the beginning of a new event. Furthermore, the auditory system can determine which partials started at the same time and use this relation to group them [Bregman, 2008]. Roads [1996] notes that more neurons fire at the onsets of sounds than if a sound is constant which may support this finding. It may be possible that the creation of sharper attacks in rhythm instruments through compression may aid separation in mixes. On the other hand, Bregman and Woszczyk [2005] demonstrate that "when onsets of instrumental sounds are edited out, leaving only the sustained sounds, perception and classification of musical instruments is confused". As mentioned earlier, instruments can be blended by trimming their envelopes into synchrony using e.g. gates [Bregman and Woszczyk, 2005].

Bregman [2009] summarizes that the onsets of sounds seem to be the most important because they tell the listener most about timing: in nature, sounds with soft on and offsets usually play the roles of ambient, atmospheric sounds where timing is not as important. Sudden changes in loudness attract much more attention as they may indicate the presence of a predator.

Spectral cues

Bregman [1990] states that the auditory system looks for correlations or correspondences among parts of the spectral content that would be unlikely to have occurred by chance. He introduces the harmonicity principle, where partials that seem to form a harmonic series are perceived as part of the same auditory stream. The auditory system makes use of the fact that the partials in harmonic sounds (sounds that are perceived to have a definite pitch) are integer multiples of a common fundamental. If played in succession, the overtones of all harmonic sounds follow the same pattern of intervals. The pitch of quasi-harmonic sounds is ambiguous, as the spectrum is inharmonic but the ear can segregate the partials into several perceived pitches. An example is a bell sound. Inharmonic sounds do not feature this characteristic spread of partials. Noises can however be grouped with other sounds if they feature similar peaks.
The harmonicity principle also makes it possible to hear a masked or missing fundamental. This allows for the application of drastic filtering to sounds in a mix and to layer sounds that mask each other partially. Bregman [2008] also found that the phases of partials can influence stream segregation, as sounds with the same frequency content but differing phases can sound different. Laitinen, Disch and Pulkki [2013] confirm that humans can perceive differences in the phase spectrum of otherwise identical sounds and that the phase spectrum affects the perceived timbre, especially in sounds with lower fundamental frequencies.

Although timbre is usually not a cue for but a result of spectral grouping, certain timbral features can be altered to create separation. Bregman gives the example of a flute sound used to brighten the timbre of a violin in order to make the solo stand out of the orchestra [Bregman, 1990]. Singers can appear louder by enlarging their pharynx cavity, leading to a boost in the high-mid frequency area. The loudness of the frequency area that the auditory system is particularly sensitive to is increased. That same frequency area is usually not occupied by a lot of other instruments, as explained by Bregman [1990]. This sensitivity is due to the uneven sensitivity to different frequencies, as measured in the equal loudness contours [Moore, 2012]. Frequencies between 2.5kHz and 3.5kHz appear to be louder due to the anatomy of the ear. Bregman [1990] points out that it is easy to follow the "bulge" in that frequency area over time, making the soloist stand out. It is assumed that presence peaks in mixes (peaks in this frequency region) can create separation by increasing perceived loudness as well as by creating contrasting brightness between sounds.

Bregman and Woszczyk [2005] state that segregating sounds by artificially induced differences in brightness through filtering can “clean up a muddy stretch of sound”. At the same time, “bandpass filtering of two sounds with the same filter settings will increase their tendency to blend.” They add that filtering can also bring out the pitch of a particular instrument: “since only a few harmonics, particularly the low ones, are needed to define a pitch, if there is a region in the spectrum where the lower harmonics of A are not mixed with those of other instruments, boosting the intensity of this spectral region will strengthen the pitch of A”.

Spatial cues

In the same way as sequential grouping, spectral grouping can be affected by spatial separation of parts of the spectrum. Generally, masking becomes a lot weaker if masker and target are in different locations. In this way, it is often useful to pan instruments occupying similar frequency regions differently in mixes. Therefore, stereo recordings are usually perceived as clearer in stereo than mono. Bregman and Woszczyk [2005] point out that “Common spatial panning of several instruments segregates them out of the mixture and groups them in the unity of
motion”. They add that spatial segregation alone may not be sufficient to separate sounds but that other factors are usually present to aid the segregation. Sounds almost always differ from each other in their amplitude envelopes, the timing of their attacks and their pitches.

In order to establish the spatial location of a sound, the ear uses different spatial cues to arrive at the perceived location of the event. Interaural time differences (ITDs) and phase differences (IPDs) are strong cues for localization in nature, especially for low frequencies. However, level differences between the right and left channel are most often used in stereo mixes. Due to the interchannel crosstalk of the loudspeakers, time panning would cause comb filtering and a lack of robustness of the centre line. Head shadowing creates an ILD for high frequencies and low frequencies diffract around the head of the listener, creating interaural phase shifts [Lipshitz, 1985]. Lipshitz [1985] points out that due to the spacing of our ears, the arrival time difference never exceeds 630µs which corresponds to a path-length that is a half wavelength at a frequency of around 800Hz. Hence, the phase relationship at frequencies below about 800Hz is unambiguous between the two ear signals. Lipshitz [1985] claims that at low frequencies, time delays between the loudspeakers produce only level and polarity differences between the two ear signals and that it is even possible that the ear on the side of the later loudspeaker receives the louder signal. It appears that therefore, time panning will not work for low-frequency signals. Contrary to this, in a study by Lee and Rumsey [2013], time panning was effective for both high and low frequency source stimuli.

Edmonds and Culling [2005] assess the effect of interaural time delays in different frequency bands on the spatial unmasking of speech. They establish that the process responsible for binaural unmasking apparently exploits interaural time differences independently within each frequency channel and that hence, binaural unmasking is indifferent to the perceived direction of sounds [Edmonds and Culling, 2005]. Full binaural advantage could be achieved even when the high- and low-frequency bands were presented with ITDs of equal but opposite magnitude.

Sometimes, phase effects can be used to alter the perceived location of sounds in a mix. Izhaki [2008] introduces the ‘out of the speaker trick’ where, by inverting the phase of a duplicated, panned signal, a sound can appear to jump out of the speakers. The influence of panning on masking is discussed in section 3.4.4 and may work similarly to fusion.

Reverberation can also affect perceptual organization [Bregman and Woszczyk, 2005]. An individual sound with reverberation will stand out from a mixture and other potentially masking sounds due to the lengthening effect of reverberation and its differentiating characteristics [Bregman and Woszczyk, 2005]. When reverberation is added to a masker, a
reverberation-free target sound may stand out more, as its onsets become more audible. Sounds can be blended and grouped by employing the same reverberator.

**Continuity illusion**

Continuity illusion is the perceived continuation of a sound through another sound [Bregman, 1990]. A pure tone and a complex tone containing the pure tone alternate. Rather than being perceived as two alternating sounds, the pure tone stands out of the complex tone and is perceived as continuing and separate. This is similar to the Gestalt Principle of Closure, where a target sound is also perceived as continuous, as introduced in the context of sequential grouping, but here, the target sound becomes inaudible during the interruptions. During the continuity illusion, the target sound is still physically heard and spectral segregation takes place. The complex tone is segregated into two streams, one containing the pure tone and one containing the remaining spectral components of the complex tone. In this way, the identity of the complex tone is changed. In the Gestalt principle of closure, however, the original sound is masked during interruptions and the masker keeps its original identity. Although the auditory complex perceives the target as continuing, it is not physically heard as a separate stream.

The continuity illusion also works when two versions of the same tone but of different amplitudes alternate. For the continuity illusion to work, no discontinuity (such as silent gaps) should be present in the target sound and there should be sufficient evidence that the target sound is continuing. During the higher amplitude tone, some subset of the neural activity in the auditory system should be indistinguishable from activity that would have occurred if the lower amplitude version had continued. At the same time, the two sounds should not be perceived as a single stream that transforms over time.

**Contralateral Induction**

In the same way as in the continuity illusion, a sound is also heard as part of another sound in contralateral induction. A soft sound is played in one ear and a loud, inducing sound in the other, causing the perceived location to be pulled toward the centre. Again, the distractor does not need to contain all frequencies present in the target but both sounds only need to stimulate similar neural activity.

**3.3.6 Schema-based grouping**

Learning and remembering a sound can affect audition, as Bregman [1990] shows in an experiment. Six pure tones of different pitches that are slightly panned to different positions group into a complex timbre. However, when the stereo position of one tone is altered, it
suddenly stands out. The auditory system therefore must have remembered the position of every single tone despite fusing them.

Bregman [1990] assumes that the knowledge of particular classes of signals is acquired through learning. In schema-based grouping, these classes of signals are recognized by the auditory system and schemas are activated when all the necessary attributes are detected. Bregman [1990] established that two vowels together are heard as such even when none of the cues for segregation presented in the above sections are present. Processes of schema-based grouping can thus be used to select evidence out of a mixture that has not been subdivided by processes of primitive scene analysis yet and put together evidence that had been portioned by primitive processes [Bregman, 1990]. This phenomenon, called duplex perception shows again that exclusive allocation can be violated. Bregman concludes that the descriptions of schemas and the components of sensory evidence may be two different sorts of entities and subject to different kinds of rules, which might explain duplex perception.

While primitive processes of stream segregation partition the spectrum into streams, schema-based grouping selects evidence for only a single stream. Schema-based processes also seem to be able to look at relations over a longer time than primitive ones can [Bregman, 1990]. Schema-based grouping is relevant to the mix process, as it may be possible to add more sounds in later choruses of a song, as the auditory system may recognise and segregate elements from earlier parts of the song.

3.3.7 The influence of Attention on Auditory scene analysis

Wrigley and Brown [2004] state that the term “attention” could refer to both the selectivity of the auditory system and its capacity limitation. Bregman [2007] notes that source separation and the interpretation of the cues outlined in this section may depend on attention. He quotes Näätänen et al. [1992, 2001] who found evidence for pre-attentive processing of auditory inputs into streams in the brain. When a regular pattern of tones changes, auditory event related potentials feature a mismatch negativity component, possibly indicating pre-attentive source segregation mechanisms in the brain. When focussing ones attention to one stream, however, other streams can become less distinct.

Wrigley and Brown [2001] state that attention is required for stream formation and not only for stream selection. They show that auditory streaming of a target does not occur when listeners attend to an alternative stimulus. When listeners direct their attention towards the target, however, stream segregation functions as outlined in this section. The authors tested this phenomenon using simple alternating sequences of high and low tones, as well as a complex
3 Relating psychoacoustic findings to clarity and separation in music mixes

tone containing a mistuned harmonic. Further research would need to be undertaken to assess the influence of auditory attention in the context of more complex sound spectra, such as music mixes.

Wrigley and Brown [2001] suggest that attention could be divided into the two levels of low-level exogenous attention, where acoustic elements are grouped to form streams, and higher-level endogenous attention, where stream selection takes place. Exogenous attention can overrule conscious (endogenous) selection, for example in response to a sudden loud bang [Wrigley and Brown, 2001]. They add that schema information can aid the grouping of the exogenous processing outputs as well as being used to detect salient information to reorient conscious attention. The authors [2001] relate their findings to the oscillatory correlation theory, where neural oscillators corresponding to grouped auditory elements are synchronised and desynchronised from oscillators encoding other groups [Wrigley and Brown, 2001].

Wrigley and Brown [2004] state that attention could either be described as being evenly distributed across a discrete range of frequencies or that the attentional focus may be a gradient in which “the density of the attentional resources is greatest at the cued frequency and declines gradually with frequency separation from the focal point of attention”. The authors [2004] also suggest that a finite amount of time is required before attention is fully allocated to a particular frequency region.

3.3.8 Auditory stream segregation in music and mixes

In the context of mixing, the fusion or segregation of individual elements results from the way they are processed. This needs to be appropriate for the piece to be mixed, where e.g. lead sounds need to be clearly audible and distinguishable from the other sounds present. Bregman [1990] explains that on the one hand, music must defeat stream segregation tendencies for clarity and on the other hand work with them when instruments need to blend to make new timbres. He points out the hierarchical form of music, where parts exist within larger parts. Groups of sounds can exist within larger groups and shifts in timbre can delineate musical units to create phrasing. Unpleasant sensations like psychoacoustic dissonance can be diminished when the dissonant notes fall into separate streams. Hence, dissonant effects can often be added to music mixes without sounding unpleasant. A change in perceptual grouping can also alter the perception of rhythms, melodic patterns, and overlap of sounds [Bregman, 2008]. As indicated by Bregman and Woszczyk [2005], ASA in mixes works in the same way as in real life. Clarity and separation between sounds can be increased by modifying the mix parameters that can manipulate the factors presented in this section (e.g. panning for spatial separation or chorus effects for modulation).
3 Relating psychoacoustic findings to clarity and separation in music mixes

Composers can create separation between soloists and other instruments by using differing pitches. As mentioned above, soloists can employ vibrato and singers can create a vocal formant making them stand out of the mix. As indicated above, the onsets of sounds can be used to create separation by playing in a rubato style.

In mixes, notes and phrases played by the same instrument often need to be perceived as one consecutive stream, making the perception of melody and harmonic progression easier. As explained by Bregman [1990], sequential patterns involving elements of the same stream will be more easily perceived than patterns with elements from several streams. When sounds fall into separate auditory streams, however, it is harder to understand their rhythmic attributes. Bregman [1990] noticed in an experiment that it is hard to tell whether there are gaps of silence between fast altering high and low tones, as they fall into separate streams.

Bregman and Woszczyk [2005] state that a group of similar instruments, such as electric guitars, can be blended into an ensemble “when their individual envelopes are trimmed into synchrony using gates or keyed (synchronous) expanders.” All in all, ASA can be used to explain clarity and separation in mixes.

3.3.9 Conclusion

In the present section, the literature dealing with auditory scene analysis (ASA) that is likely to be relevant for the mix process was reviewed. The questions presented in the introduction can now be answered. These were as follows: what is auditory scene analysis? What factors influence the grouping or segregation of sonic elements? Does attention play a role in auditory scene analysis? How does this relate to music and mixing? The answers to these questions will be summarized in the following paragraphs.

What is auditory scene analysis? As established in this section, “Auditory scene analysis” is the process of forming mental representations of individual sounds from the summed waveform that reaches the ears. First, the auditory system divides the input into packages of acoustic evidence (segmentation). Packages that appear to have arisen from the same source are then recombined (grouped). Packages can be grouped into a single stream or segregated into different streams by employing mechanisms of primitive grouping (following innate constraints) and schema-based grouping (learned constraints). In sequential grouping, subsequent sonic elements are grouped, and in spectral grouping, simultaneous sonic elements are grouped.

What factors influence the grouping or segregation of sonic elements? Cues for sequential segregation are the frequencies, pitches, amplitudes, locations and timbres of sonic elements,
the center frequencies of noise bands, the suddenness of the changes of these variables from one sound to the next and the duration of silence between sounds. Spectral grouping depends on the onset and offset synchrony, frequency separation, regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation and parallel gliding of spectral components. The Gestalt principles of closure, grouping, exclusive allocation and good continuation, the principle of common fate and the old plus new heuristic can explain some phenomena found in ASA. Does attention play a role in auditory scene analysis? Relevant literature suggests that attention does play a role in ASA, where the formation of streams can depend on this.

How does this relate to music and mixing? In mixes, separation needs to exist between some sounds while others need to blend. This can be achieved in the composition and production process by taking the above factors into consideration. All in all, auditory scene analysis appears to play an important role for the study of clarity and separation in mixes.

### 3.4 Relating masking to clarity in mixes

In the context of clarity and separation (the audibility of individual instruments) in mixes, masking is likely to play an important role. Generally, mixing engineers attempt to minimize masking of lead instruments, vocals and other important instruments in order to increase their presence and clarity. The present section provides an overview of masking, its definition, the factors influencing it and the psychophysical models used to explain masking. The following research questions will be answered. What is masking? What factors (spectral, temporal, spatial) affect the degree to which masking is likely to occur? Can this be explained through psychophysical models such as the auditory filter or the power spectrum model of masking? And how does it relate to mixing?

In section 3.4.1, masking in defined. In section 3.4.2, the spectral factors influencing masking are established by introducing the auditory system’s frequency selectivity, the auditory filter model, the power spectrum model of masking and critical bandwidths and by comparing tonal and noise maskers. In section 3.4.3, the role of temporal attributes in target and masker will be discussed, introducing comodulation masking release, dip listening, profile analysis, the overshoot effect, as well as forward and backward masking. In section 3.4.4, the influence of spatial cues on masking will be assessed and lastly, in section 3.4.5, masking will be discussed in the context of music mixes.
3.4.1 Definition of masking

According to the American National Standards Institute [ANSI/ASA, 2013], masking is the process by which the threshold of audibility for one sound is raised by the presence of another, masking sound. The amount by which the threshold of audibility is raised is expressed in dB [ANSI, 1994].

Masking works best if the targets and masker lie in similar frequency regions, as shown in section 3.4.2. According to Moore [Moore, 2012], masking reflects the limits of the auditory system’s frequency selectivity which will be explained in further detail in the next section.

Masking does not only occur when masker and target are presented simultaneously, but also when the masker is played before (forward masking) or after the target (backward masking). Certain spectral, temporal and spatial factors can influence the threshold of audibility of the target as established in the following sections.

3.4.2 Spectrum related factors influencing masking

As mentioned above, masking works best if the masker and target lie in similar frequency regions. This can be explained with the auditory system’s frequency selectivity, the resulting auditory filter model and the power spectrum model of masking. From this, the concept of critical bandwidths or equivalent rectangular bandwidths can be derived. These phenomena will be presented in more detail in the following sections.

Frequency selectivity and the auditory filter

The peripheral auditory system can be described as a bank of band pass filters with overlapping pass bands, the auditory filters [Fletcher, 1940; Helmholtz, 1863]. Fletcher [1940] relates the auditory filters to the anatomy of the basilar membrane, although neural coding also plays a role [Moore 2012]. Moore [2012] points out that according to the current literature, each location on the basilar membrane responds to a limited range of frequencies, so different points correspond to filters with different centre frequencies. The auditory filters are discrete, as the number of hair cells is finite. However, their spacing is so close that they could be described as continuous. The shape of the auditory filters can be described as a weighting function with a rounded top and sloping edges.

Two sound sources with different pitches (e.g. two different tuning forks) can be separated unless they are very close together in frequency (this ability is also called “frequency resolution” or “frequency analysis” [Moore 2012]. In the latter case, the pitches can merge into one sound,
3 Relating psychoacoustic findings to clarity and separation in music mixes

in which case roughness and beating can occur [Helmholtz, 1863]. The two frequencies lie within one “critical band”.

Critical bandwidths and the power spectrum model of masking

The phenomenon of critical bandwidths introduced above can be demonstrated and measured by studying masking. Fletcher [1940] measured the threshold for detecting a sinusoidal signal as a function of the bandwidth of a band-pass-filtered noise masker. The noise was centred at the frequency of the target sound with a constant noise power density. The noise power was then increased with the masker’s bandwidth, leading to an initial increase in the target’s threshold. Eventually, a point was reached where the threshold did not increase further but instead stayed constant, although the masker became louder. This is illustrated in Fig. 3-6. Fletcher [1940] concluded that only the components in the masker that pass through the auditory filter containing the target sound can have a masking effect. The maximum bandwidth up to which the noise bandwidth can increase before masking is constant is the critical bandwidth [Fletcher, 1940]. Fletcher [1940] proposed specifying the bandwidth of a rectangular filter passing equivalent energy, for simplification. He thus established the concept of an “equivalent rectangular bandwidth”.

![Graph showing the relationship between signal threshold and masking noise bandwidth.](image)

Fig. 3-6: an increase in the noise power with the masker’s bandwidth leads to an initial increase in the target’s threshold until a point is reached where the threshold does not increase further [Moore, 2012]

Patterson and Moore [1986] derive the power spectrum model of masking from the findings above, where the threshold of a signal is determined by the amount of masker energy passing through an auditory filter centered on the signal’s frequency. This also explains why a low-complexity sound is more likely to mask a high-complexity sound than vice versa. The authors
assume that the threshold of a signal corresponds to a certain signal-to-noise ratio. Stimuli are represented by their long term power spectra, ignoring the relative phases of components and short term fluctuations in the masker [Patterson and Moore, 1986]. The power spectrum model of masking cannot explain all phenomena associated with masking, such as masking occurring for a masker and target of significantly different pitches, but it proves useful in many situations.

In the following paragraph, again, all findings are summarized by Moore [2012], unless otherwise specified. When a noise just masks a tone, the power of the tone $P$ divided by the power of the noise inside the critical band is a constant $K$. Moore describes the constant $K$ as a measure of the efficiency of the detection process following the auditory filter. He adds that the value varies between people. Noise power is specified in terms of power in frequency bands 1 Hz wide, resulting in the noise power density measure $N_0$. As the power-per-Hz of white noise is independent of frequency, the total white noise power falling in a critical band that is $W$ Hz wide is $W \cdot N_0$. Based on this, Fletcher [1940] presents the equation

$$K = \frac{P}{W \cdot N_0} \tag{3.5}$$

Moore [2012] points out that $N_0$ is known, $P$ can be measured, $K$ can be estimated, and hence $W$ can be estimated. According to Fletcher [1940], $K$ is roughly equal to 1, leading to a “critical ratio” of $P/N_0$. According to Moore [2012], $K$ has been found to approximate 0.4 in recent studies, but the resulting value usually depends on the method used. Moore [2012] observes that the auditory filter becomes broader with increasing level, especially on its low frequency side. He states that the equivalent rectangular bandwidth for a tone $ERB_N$ in healthy listeners can be measured as

$$ERB_N = 24.7(4.37F + 1) \tag{3.6}$$

where $ERB_N$ is measured in Hz but $F$ (the centre frequency) is measured in kHz. Moore [2012] describes the power spectrum model of masking and equivalent rectangular bandwidths as a useful approximation. However, although the critical bandwidth is a good measure of the effective bandwidth, the auditory filters are not rectangular. Instead, as shown by Fletcher [1940], they are rounded at the top, featuring sloping edges. No distinct break point exists for the critical bandwidth, i.e. the width of the auditory filter depends on the output level. Furthermore, the auditory filter is roughly symmetric on a linear frequency scale at moderate sound levels. At high sound levels, however, the low frequency side of the filter becomes shallower than the high frequency side [Moore, 2012]. Moore [2012] also introduces the phenomenon of the upward spread of masking which is also not included in the model. As illustrated in Fig. 3-7, Wegel and Lane [1924] kept a masker constant and varied the stimulus...
Relating psychoacoustic findings to clarity and separation in music mixes

frequency. They assessed how the threshold of a pure tone depends on its frequency for a narrow band of noise centred at 410Hz. For the high frequency functions, the slopes for the curves become shallower at high sound pressure levels [Moore, 2012]. Moore [2012] explains that if the level of a low-frequency masker is increased by a certain amount, the masked threshold of a high-frequency signal is elevated by a larger amount; hence the amount of masking grows nonlinearly in the high-frequency side. As a result, low-frequency maskers can mask high-frequency signals more easily than vice-versa.

![Graph showing threshold shift as a function of frequency](image)

**Fig. 3-7:** the threshold shift of a pure tone signal as a function of its frequency, in the presence of a narrow band of noise centred at 410Hz. For the high frequency functions, the slopes for the curves become shallower at high sound pressure, as measured by Wegel and Lane [Moore, 2012].

Spectrum-related masking effects not described by the power spectrum model

Most masking phenomena can be explained with the assumption that the listener monitors only a single auditory filter with the highest signal to masker ratio and it is often possible to predict if a complex sound will be detected in a given background noise by calculating the detection thresholds of its most prominent frequency components [Moore, 2012]. However, there are exceptions to this assumption that will be presented in the following.

Listeners can compare the outputs of different auditory filters [Moore, 2012]. This is called “off frequency listening” or “off place listening”. Here, the listener can make use of a lower-frequency filter when the masker is higher in frequency than the stimulus to attenuate the masker more than the stimulus. Moore [2012] quotes Spiegel [1981] who suggests that the ear is capable of integration over bandwidths much greater than the auditory filter bandwidth. This is discussed below, where comodulation masking release is discussed. At the same time, masking can still occur when a target and masker differ strongly in pitch.
Relating psychoacoustic findings to clarity and separation in music mixes

The spectral content of masker and target has a strong influence on the target’s threshold. Taghipour et al. [2013] point out that the masking effect of narrowband noise is stronger than that of a tone of the same power placed in the center frequency of the noise. Hence, the authors stress that some kind of tonality estimation needs to be included in masking models. Moore [2012] explains that a sinusoidal masker and target can cause roughness and beating when both are close together in pitch. The auditory system features a varying sensitivity to beats, where slower beats are easier to detect, which can help the auditory system to notice the target. The spacing of complex tone harmonics can also influence a signal’s threshold: the threshold plotted as a function of bandwidth increases monotonically as bandwidth increases beyond 50Hz [Moore, 2012].

3.4.3 Temporal factors influencing masking

In the following, the effect of temporal factors on masking will be established. The following sections present the comodulation masking release, the effects of dip listening and profile analysis, the overshoot effect, as well as forward and backward masking.

Comodulation masking release

Moore [2012] explains that when the masker amplitude is modulated, leading to a correlation in different frequency bands, a reduction in the signal threshold can be observed. The same happens when the target is modulated. This is called “Comodulation Masking Release” (CMR).

Moore [2012] points out that across-filter comparisons can enhance the detection of sinusoidal signals in fluctuating noise maskers. When a correlation across frequencies is observed, comodulation masking release is possible. The signal threshold decreases as the bandwidth increases for noise bandwidths greater than 100Hz which, again, shows the limitations of the power spectrum of masking: a frequency band centred at the stimulus frequency could be masking the latter but as soon as a flanking band further away comodulates with the masking frequency band, masking release can be possible. This phenomenon works even if the flanking band is spatially separated [Moore, 2012].

Schooneveldt and Moore [1987] show that modulation on a masker can produce release from masking even if the masker’s bandwidth is less than the auditory filter bandwidth, concluding that CMR does not always depend on comparisons between outputs of different filters but sometimes on “within-channel cues”.

Moore [2012] adds that CMR is less effective in brief signals and that it does not vary significantly with signal frequency. According to Moore [2012], CMR is greatest when the masker modulates at a low rate and covers a wide frequency area. He assumes that the ear
Relating psychoacoustic findings to clarity and separation in music mixes

possibly compares envelope modulation patterns at the outputs of different auditory filters and detects the presence of an added signal (the stimulus) as a disparity in modulation across filters [Moore, 2012]. Alternatively, the envelope fluctuations at the outputs of the auditory filters tuned away from signal frequency may make the minima in the masking envelope and (hence the optimum times to listen for the signal) more obvious. Moore [2012] adds that the two assumptions are not mutually exclusive.

Dip listening

The latter assumption in the paragraph above can be related to a phenomenon called “dip listening” which may, as pointed out by Moore [2012], be related to the compression that occurs on the basilar membrane. Low amplitude portions of a signal are amplified more than high amplitude portions, causing the signal to be enhanced when dips occur in the masker [Moore, 2012]. Hence, when the masker features a high peak factor (a high ratio of peak to RMS levels), the stimulus can be heard between peaks. As a result, the phase relationships in a complex tone have an impact on the stimulus threshold. As explained by Moore [2012], a waveform with a very high peak factor can e.g. result from all sinusoid components in a harmonic waveform starting in a 90 degree phase. When starting phases are random, the peak factor is lower. This, again, deviates from the power spectrum model of masking where the influence of phases on the target threshold is ignored.

Profile analysis

Another temporal phenomenon that can lower the threshold of a signal is “profile analysis”, where the auditory system compares the outputs of different filters to detect a signal, detecting increment in one component relative to the level of other components [Green 1988]. Even if the magnitude of the output from any single auditory filter is an unreliable cue to the presence of the signal (e.g. by randomizing the overall sound level of each stimulus), subjects can still detect the signal by comparing the outputs of different filters. Green [1988] adds that this is most effective when the background (masker) has large spectral range and many components (not too close to the stimulus frequency), when the signal is at the edge of the background, when the signal is of similar or greater loudness to the background and when background components are also similar in level to each other.

The overshoot effect

Moore [2012] presents the overshoot effect, where the threshold for detecting a brief signal in a noise masker is greater if it is presented near the masker onset or turned on and off simultaneously with an equally brief masker, rather than when it is presented after a long onset delay in a continuous masker. According to Moore [2012], the effect is greatest when the
masker covers a broad frequency range, when the signal is at a high frequency and when the masker is at a moderate level. He adds that it is unclear why this is the case.

**Forward and backward masking**

The temporal separation of a target and masker does not necessarily lead to masking release. The poorly understood phenomena of forward and backward masking are presented in this section. Non-simultaneous masking is often studied using short signals called "probes".

In forward masking, the masker is presented before the signal [Moore, 2012]. This may occur due to the persistence of responses on the basilar membrane (ringing), short time adaptation in the auditory nerve, the persistence of neural activity, or an inhibition evoked by the masker at some level of the auditory system. Alternatively, the efferent system may be activated (the efferent neurons conduct impulses outwards from the brain, as a reaction to incoming signals carried by afferent neurons), resulting in reduced gain of active mechanisms and reducing the effective level of the signal [Moore, 2012]. Furthermore, a temporal overlap of patterns of vibration on the basilar membrane might be important especially for small delay times between signal and masker [Moore, 2012]. In the following, the main characteristics of forward masking, as established by Moore [2012], are summarized.

Firstly, forward masking is greatest when the masker is presented spatially near the signal. The rate of recovery from forward masking is greater for louder maskers but it always decays to 0 after 100 to 200 milliseconds. While the threshold in simultaneous masking usually corresponds to a fixed signal-to-noise ratio, this is not the case in forward masking. When the level of the masker is incremented, the masking effect does not increase by the same amount. Moore [2012] assumes that this could be due to the nonlinear input-output function of the basilar membrane (compression effects). The phenomenon may also depend on the frequency relationship between masker and target [Moore, 2012].

The psychophysical tuning curves introduced are sharper in forward than simultaneous masking [Moore, 2012]. Moore quotes Houtgast who suggests that this could be because the internal representation of the masker might be sharpened by a suppression process. Alternatively, in simultaneous masking, the signal may be suppressed by the masker, increasing the effectiveness of the masker for masker frequencies well above and below the signal frequency [Moore, 2012]. In non-simultaneous masking, however, the masker does not suppress the signal, making it less effective.

Moore [2012] states that backward masking, where the stimulus is presented before the masker, is poorly understood and depends strongly on the amount of practice test subjects have received; some subjects show no backward masking. Moore [2012] suggests that the larger
3 Relating psychoacoustic findings to clarity and separation in music mixes

masking effects found for unpractised subjects may reflect some sort of “confusion” with the masker [Moore, 2012].

3.4.4 Spatial factors

In the following, the influence of spatial cues (interaural intensity and phase differences, spectral cues) on masking will be discussed before introducing masking caused by reverberation.

The influence of spatial cues on masking

Differences between the phases, intensities and spectral shapes reaching the ears can reduce masking, as found by a number of authors. The interaction between the interchannel time differences and the temporal characteristics of sound sources plays an important role. In this context, Moore [2012] introduces the "equalization and cancellation" model. According to this, the auditory system tries to eliminate the masking components of a binaural-masking stimulus by transforming the total signal in one ear relative to the total signal in the other ear until the masking components are exactly the same in both ears. Hence, the masked threshold of a signal can be lower when listening with both ears, rather than one. The improvement in the detectability of a signal under binaural listening conditions is measured with the binaural masking level difference [Moore, 2012]. This is the difference between the threshold of the signal when the signal and masker have the same level and phase relationships and the threshold when the phase and/or level relationships or the signal and masker are different.

Experiments using speakers and headphones can also produce different results, presumably due to the channel crosstalk in speakers, leading to changes in the binaural cues listeners are perceiving. The spacing of sounds in the vertical plane can also decrease thresholds. However, spatial cues for unmasking do not always correlate with sound localization. A masker and target appearing to arrive from spatially separate locations are not necessarily less likely to cause masking phenomena and spatial cues can aid unmasking even when localization is blurred.

Bronkhorst and Plomp [1988] found that binaural unmasking through interaural time and level differences contributed to the spatial unmasking of speech and speech-shaped noise presented over headphones. For noise azimuths between 30 and 150 degrees, the gain due to ITD was between 3.9dB and 5.1dB and the gain due to ILDs was between 3.5 and 7.8dB.

Saberi et al. [1991] assume that the release from masking is related to a discrimination between spectral shapes under monaural and vertical plane conditions, and to binaural phase or intensity cues horizontally, and not necessarily to a difference in the perceived location of the target and masker.
Culling et al. [2004] investigated the roles of interaural time and level differences in spatial unmasking in multi-source environments. They observed that in the context of interaural level differences, speech reception thresholds were only lowered when the interfering sources were also in one hemifield, giving the contralateral ear an advantageous signal-to-noise ratio [Bronkhorst and Plomp, 1988]. This seems to contradict the theory that sounds that appear to be spatially separate are always less likely to mask each other. Instead, the authors state that best ear listening and binaural unmasking, but not sound localization, can explain the effect of ILDs and ITDs on masking. This is in accord with Saberi et al.’s [1991] assumption.

Lee [2011] studied test subjects’ masked and localised thresholds when presented with stimuli and maskers through two vertically placed channels. He defined the masked threshold as the “level of delayed height channel signal at which any subjective effect of delayed signal became completely inaudible” [Lee, 2011]. Nine different delay times ranging from 0ms to 50ms were tested using cello and bongo samples. Lee [2011] states that although no interchannel time or level difference relationship was present in the vertical domain, the average level reductions of the delayed signals required for a masking effect were significant (between 9 and 10dB). Following the findings of his experiments, Lee [2011] also notes that the interaction between the interchannel time differences and the temporal characteristics of sound sources determines the masked thresholds of vertically distributed signals and maskers, especially for interchannel time differences larger than 10 ms.

Ahonen and Pulkki [2008] investigated the influence of spatially wide noise maskers on a frontal signal. The noise sources were presented over symmetrically positioned loudspeakers in the frontal horizontal plane and in anechoic conditions. The authors found no relationship between the detection threshold of the signal and the masker width, concluding that frontal unmasking does not exist in loudspeaker listening [Ahonen and Pulkki, 2008]. The authors note that their results differ from results in previous headphone listening experiments with corresponding coherence values. They point out that this may somehow be related to the lack of crosstalk in headphone listening. A separation of the stimulus and masker horizontally or a difference by interaural time or level differences led to a degree of masking release, however [Ahonen and Pulkki, 2008]. Aichinger et al. [2011] suggest that adding reverberation to a masker could also minimize the masking effect.

Masking caused by reverberation

Spatial cues cannot only minimize but also cause masking. Zarouchas and Mourjopoulos [2011] used a computational auditory masking model to evaluate masking and smearing generated in the time-frequency domain due to reverberation decay in stereophonic sound reproduction.
Time-frequency maps were used to show monaural and binaural attributes affected by reverberant cues in reverberant, compared to dry signals. The authors establish a reverberation masking index (RMI) to model and quantify some of the masking effect of reverberation in the same way as masking noise. For simplicity, interaural cues were not taken into consideration. Zarouchas and Mourjopoulos [2011] found a correlation between the RMI and well-established, signal-independent statistical acoustical properties of the room, as measured by the RT or D/R values. Longer reverberation times as found in larger rooms or at more distant receiver positions lead to higher RMIs. In an earlier, similar study, Zarouchas & Mourjopoulos [2009] also established that monaural masking due to reverberant decay is proportional to reverberation interference and to room reverberation time. However, the exact nature of this relationship can differ between audio signals.

3.4.5 Masking in mixes

Although the stimuli used in most masking experiments are simple tone and noise samples, some of the findings can be related to the mix process, where instruments occupying similar frequency regions are likely to mask each other. Signal processing such as equalisation and panning can be used to reduce masking.

Partial masking can reduce the loudness of the target in the mix [Ma et al., 2014]. This is likely to lead to a reduction in clarity. Related to this, Pestana and Reiss [2014] point out that in music mixes, EQ should be applied to ensure that no element masks any of the frequency content of the vocals. They also note that reverberation can be carefully substituted for delays to avoid masking. Related to this, they state that masking caused by delays and reverberation is more likely for low frequencies and transients. This may be due to the fact that masking is particularly critical here. Several tools have been developed to tackle masking in mixes, such as MixViz [Ford et al., 2015]. Further research on automatic masking reduction in mixes was undertaken by Perez Gonzalez et al. [2008], Hafezi and Reiss [2015] and [2014]. However, none of the studies explicitly link masking to clarity.

Aichinger et al. [2011] attempted to predict the transparency of mixdowns by assessing masking effects. The masked-to-unmasked-ratio is used to relate the original loudness of an instrument to its loudness and presence in a mix, taking the frequencies of instruments that compete against each other into consideration. Although spectral, binaural, reverberant and temporal effects influence masking, the paper focuses on spectral effects for simplification [Aichinger et al., 2011]. The authors assume that instruments could be separated in mixes through panning or different reverberation settings and they point out that the acoustic
properties of the transient onset of an instrument is likely play a significant role, too, although the features that are likely to be important are not investigated further.

Aichinger et al. [2011] find a significant correlation of identification probability (IP, the extent to which instruments in mixes can be identified) and the masked-to-unmasked-ratio in masked-to-unmasked-ratios of 10% or more. Below this ratio, the IP is at the baseline level that would be achieved by guessing [Aichinger et al., 2011]. The authors suggest that therefore the masked-to-unmasked-ratio of each instrument in a mix should be at least 10%. In general, it is assumed that the masked-to-unmasked-ratio of each instrument should be as high as possible although this also depends on the context. The authors point out that instruments in a polyphonic melody or voices in a choir may sound better as a merging sound while solo instruments may be best presented as disunited sound.

3.4.6 Conclusion

The present section summarizes the current literature about masking and relates it to the mix process. The following research questions have been answered. What is masking? What factors (spectral, temporal, spatial) affect the degree to which masking is likely to occur? Can this be explained through psychophysical models such as the auditory filter or the power spectrum model of masking? And how does it relate to mixing? To summarize the findings presented in this section, the questions will be answered in short in the next paragraphs.

Masking is the process by which the threshold of audibility for one sound is raised by the presence of another, masking sound. The amount by which the threshold of audibility is raised is expressed in dB.

Generally, sounds are likely to be masked when other, louder sounds are present. In the spectral domain, a masker is most effective when it has energy close to the target’s frequency. This relationship can be modelled by the auditory filter model and the power spectrum model of masking. It is assumed that the threshold of a signal is determined by the amount of masker energy passing through an auditory filter centered on the signal’s frequency. Hence, the components of the masker that fall in auditory filters other than those detecting the target will have no effect on the degree of masking. Due to the upward spread of masking, sounds that are lower in frequency are more likely to mask sounds that are higher in frequency than vice versa.

Three frequency-related phenomena are not considered in the power spectrum model of masking: off-frequency listening, the ear’s capability of integration over bandwidths much greater than the auditory filter bandwidth and the fact that masking can still occur when a target and masker differ strongly in pitch. Another phenomenon that is not modelled by the
power spectrum model is the fact that narrowband noise can have a stronger effect than harmonic tones.

Temporal factors like comodulation masking release, dip listening, profile analysis and the overshoot effect can have an influence on masking. Hence, modulating either target or masker can reduce the masking effect. Masking can also occur when the masker occurs before (forward masking) or after the target (backward masking). Spatial cues like interaural time and level differences, spectral differences caused by differences in placement in the median place and reverberation, in combination with temporal characteristics of the sounds and the use of headphones or speakers, influence masking. The perceived distance between sounds is not a reliable cue, however. Reverberation can also cause masking.

In mixes, the masked-to-unmasked ratio could possibly be used to indicate clarity, presence and loudness of instruments. Generally, masking of important instruments should be kept to a minimum with a high masked-to-unmasked ratio, in order to ensure that they can be identified, although this is context dependent.

### 3.5 Relating factors for loudness to clarity in mixes

Clarity and separation in mixes is defined as the extent to which individual sounds can be heard. In general, sounds can be heard well when they are loud relative to other sounds which makes literature on loudness perception relevant for the study of clarity and separation in music mixes. In this section, the following research questions will be addressed. What is loudness (section 3.5.1)? How can it be measured (section 3.5.2)? What factors does loudness depend on, apart from physical sound pressure? Can a sound be made louder without increasing the SPL (section 3.5.3)? In order to answer these questions, literature on loudness perception will be reviewed. In section 3.5.4, loudness will be related to clarity and separation in music mixes.

#### 3.5.1 What is loudness?

Moore [2012] provides the following definition of loudness: “Loudness is that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud”. Loudness is a subjective quantity and different from sound pressure (a measure of the local deviation from the ambient atmospheric pressure) or sound intensity (the product of sound pressure and acoustic particle velocity). The loudest sound we can hear without immediately damaging our ears (approximately 120 dB SPL) is about $10^{12}$ times greater in intensity than the quietest sound we can perceive [Moore, 2012]. At the same time, the auditory system is able to detect very small changes in loudness. Moore [2012] argues that loudness may
3 Relating psychoacoustic findings to clarity and separation in music mixes

depend on the total neural activity evoked by a sound, or possibly a summation of neural activity across critical bands. It is also possible that the timing of neural spikes in the Auditory nerve could provide cues for the perceived intensity of a sound and phase locking may play a role in coding the relative levels of components [Moore, 2012]. Further research is necessary to establish how exactly loudness is processed.

3.5.2 Measuring loudness

Moore [2012] points out the difficulty in measuring loudness, quoting Helmholtz: "We are exceedingly well trained in finding out by our sensations the objective nature of the objects around us but we are completely unskilled in observing these sensations per se and the practice of associating them with things outside of us actually prevents us from being distinctly conscious of our pure sensations." Hence, test subjects really judge the apparent distance, the context and the nature of a sound when they are trying to judge loudness, making it difficult to establish an objective measure for loudness [Moore, 2012]. Due to this, it may be possible to make a sound in a mix appear loud by giving it certain sonic attributes that usually only loud sounds have (e.g. the frequency response that only a loud sound can have at a certain distance). Further study would be required to assess this assumption.

In order to develop a scale for loudness, a magnitude estimation test can be executed, where subjects rate the loudness of a stimulus on numerical scale. Another approach is to ask subjects to match the loudness of a stimulus to that of a given 1kHz tone. Alternatively, test subjects can adjust the loudness of the 1kHz tone itself to match that of the comparison tone. In this way, the quantity "phon" was established as a measure for loudness [Moore, 2012]. The loudness of a 1kHz tone is defined to be equal to its sound pressure level measured in dBs. The SPL of the 1kHz tone that matches the loudness of a test stimulus is then defined as the loudness of that stimulus, measured in phons. Fletcher and Munson [1933] executed the experiment using sine tones, deriving the well-known equal loudness contours. Nowadays, various versions of the equal loudness contours exist. Fig. 3-8 shows the ISO standard equal loudness contours.
3 Relating psychoacoustic findings to clarity and separation in music mixes

Magnitude estimation tests can also be performed for complex tones, either with or without a comparison stimulus. Alternatively, subjects can adjust a stimulus in a magnitude production test to an absolute level or relative to a comparison stimulus.

An alternative measure to the phon is the “sone”, as established by Stevens [1972]. One sone corresponds to the loudness of a 1kHz tone at 40dB, presented binaurally from the front in a free field [Moore, 2012]. At 50dB, the tone appears to be twice as loud, resulting in a measure of 2 sones. At low levels, loudness changes more rapidly. As a result, loudness can be described as a power function of physical intensity. A twofold change in loudness results from a tenfold change in intensity. This is expressed in the equation

$$L = kI^{0.3}$$  \hspace{1cm} (3.7)

where $k$ is a constant depending on the subject and units used [Moore, 2012] and $L$ stands for “loudness” in sones. It is not specified which quantity the intensity $I$ is measured in here but Moore points out that the equation shows that a twofold change in loudness is produced by a tenfold change in intensity, corresponding to a 10dB change in level [Moore, 2012]. Phons measure loudness level, i.e. an increase in loudness corresponds to an increased phon value, but the increase is not proportional. Sones, however, measure loudness more directly, i.e. as a doubling in Sons corresponds to a doubling in loudness.
Relating psychoacoustic findings to clarity and separation in music mixes

Often, SPL meters weight the intensity of complex tones at each frequency according to the equal loudness contours before performing a summation across frequencies [Moore, 2012]. In this way, the measurement becomes more representative of perceived loudness. A-weighting is used to reduce the contribution of low frequencies to the meter reading at low levels. Similarly, the C-network can be used for loud levels and the B-network for intermediate levels (the differing effect of overall levels on the relative loudness of different frequency areas can be seen in Fig. 3-8). In this way, loudness perception can be approximated.

3.5.3 Factors influencing loudness

In the following sections, the factors determining the loudness of a sound are summarized.

Fundamental Frequency

As shown in Figure 1, the perceived loudness of pure tones depends on their frequency. Humans are most sensitive to frequencies between 3kHz and 5kHz, where pure tones appear loudest. The threshold for hearing quiet tones is also much lower in this area. In complex tones, the predominant frequency and signal bandwidth play an important role in loudness perception [Cabrera and Miranda, 2011].

Intensity

When a sound has a greater physical intensity than another, otherwise identical sound, it will appear louder, as long as the difference in intensity is equal to or larger than the JND (just noticeable difference). Measurements for the relationship between loudness and sound intensity have been introduced in section 3.5.2. According to Pavel [1976], the loudness of short signals with rectangular envelopes is proportional to the product of their power and their duration (i.e., their energy).

The uneven sensitivity to different frequencies mentioned above is flattened at high intensities. The rate of growth of loudness level with increasing intensity is greater at low frequencies (and to a lesser extent at very high frequencies) than at middle ones [Moore, 2012]. Moore adds that in reproduced complex sounds, the relative loudnesses of different frequency components changes as a function of overall level, altering the perceived tonal balance of the sound.

Temporal changes in intensity

According to Johnston [2013] overall loudness models for extended time periods are still in development. However, several authors have found that the duration and amplitude fluctuations of a sound need to be taken into consideration when assessing loudness. These factors will be discussed in the following.
Duration

The duration of a sound can influence its loudness and hence its absolute threshold. As mentioned above, the loudness of short signals with rectangular envelopes is proportional to the product of their power and their duration [Pavel, 1976]. Glasberg and Moore [2002] note that for sinusoids of fixed peak level and up to 100ms in duration, the loudness level increases by roughly 10 phons for each tenfold increase in duration. This is equivalent to a 3-phon increase per doubling of duration [Glasberg and Moore, 2002]. The authors add that both short-term and long-term loudness take time to build up and also to decay after the input has stopped. They conclude that the long-term loudness (the latter is the overall loudness of longer segments) of sounds may correspond to a memory for the loudness of an event and that this could possibly be “reset” by a new sound event.

Sound envelopes

Several authors note that the envelope of a sound can influence its loudness. Glasberg and Moore [2002] examine the instantaneous loudness, the short-term loudness and the long-term loudness over time for a 200ms 4kHz tone burst, gated on and off abruptly. They note local maxima at times corresponding to the onset and offset of the sound in the instantaneous loudness. They also found a slight rise in the short-term loudness at the end of the signal. The authors point out that these phenomena are caused by spectral spreading related to the abrupt gating which shows the influence of onset times on loudness.

Stecker & Hafer [2000] assessed the loudness of sounds with temporally asymmetric amplitude envelopes. They noticed that for sinusoidal (330–6kHz) and broadband noise carriers, stimuli with slow onsets and abrupt offsets were perceived as louder than stimuli of equal energy featuring fast onsets and slow offsets. When playing the latter version first, the biggest difference was noted between stimuli. Similarly, Stecker & Hafer [1996] gated pure-tone (330Hz) signals with asymmetric temporal envelopes whose rise and fall times were unequal, presenting them both forward and reversed to test subjects. In this way, peak amplitudes appeared either early or late in the signals. Late-peaking signals were perceived as louder than early-peaking signals. Raimond and Watkins [2008] also state that stimuli with a slow attacks and fast decays are judged to be louder than stimuli of equal energy with fast attacks and slow decays.

Stecker and Hafer [2000] relate these findings to the parsing of auditory input into direct and reverberant sound. Stecker and Hafer [1996] propose that this may indicate a form of perceptual constancy where the slow decay of a stimulus is treated as reverberation and not
3 Relating psychoacoustic findings to clarity and separation in music mixes

included in the judgment of loudness. Furthermore, forward masking of the stimulus tail by the onset may have occurred in sounds with fast onsets and slow offsets [Stecker and Hafter, 1996].

Stecker & Hafter [2000] note that these results are not compatible with automatic gain control theories, where the ear compresses the intensity range of sounds before coding them in the discharge pattern of auditory nerve fibres [e.g. Moller, 2006]. According to this, signals with more rapid rise times should be less affected by automatic gain control and thus appear louder. In the same way, Roads [1996] notes that more neurons fire at the onsets of sounds than when a sound is constant, and Moore [2012] assumes that the amount of neural activity may be an indicator of loudness. Power-spectrum models of loudness, or predictions obtained using the auditory image model as presented by Patterson and Allerhand [1995] are also not in accord with Stecker & Hafter [2000 and 1996].

Pavel [1976] also assesses the effect of nonrectangular envelope shapes on loudness judgments for short stimuli. A 1kHz pure tone’s amplitude was modulated by various slowly varying functions, including combinations of rectangles, decaying exponentials and growing exponentials with durations ranging from 25ms to 2000ms. The test stimuli were compared to rectangular standard tones of the same frequency and duration. Pavel [1976] found no correlation between the shape of the envelopes and the loudness of short stimuli (25 ms). Results for longer duration signals could not be found.

Modulation

Glasberg and Moore [2002] review relevant literature on the loudness of time-varying sounds and compare conclusions of other authors to their own test results. Both indicate that for carriers amplitude modulated at low rates, perceived loudness corresponds to a level between the RMS and peak level. For modulation rates up to 10Hz, listeners found it difficult to describe overall loudness rather than loudness fluctuations.

Sounds that are modulated at intermediate rates are judged to be slightly quieter than the loudness of the RMS level [Glasberg and Moore, 2002]. Sounds that are modulated at higher rates (e.g. sinusoidal carriers modulated between 30 and 100Hz), are judged to be louder than those modulated at intermediate rates. For high modulation rates loudness decreases again, as the spectral sidebands are resolved and a loudness summation effect across frequency occurs. The modulation rate at which this first occurs increases with increasing centre frequency according to Glasberg and Moore [2002].
3 Relating psychoacoustic findings to clarity and separation in music mixes

Spectrum

As mentioned above, in complex tones, the predominant frequency and bandwidth play an important role in loudness perception [Cabrera and Miranda, 2011]. At the same time, as stated by Zwicker, et al. [1957] the frequency bandwidth of a complex tone has an impact on its loudness. The authors show that when the spacing between a group of pure tones is increased, the loudness remains constant until reaching a critical point, after which it increases. This also applies to bands of noise of constant sound pressure and differing bandwidths. Zwicker, Flottorp and Stevens [1957] found the smallest bandwidth at which loudness summation depends on the energy spread to be approximately the same as the critical bandwidth determined previously by methods involving thresholds, masking, and phase. They conclude that the concept of critical bandwidths is applicable to loudness summation. Johnston [2013] notes that inside of a critical bandwidth, loudness grows as the $1/3.5$ power of the power present in that band. When energy is spread over critical bandwidths, however, the loudness adds across frequencies outside of critical bandwidths, but is compressed inside each [Johnston, 2013]. Similarly, Moore [2012] argues that the spread of excitation may play a role in intensity discrimination but he notes that when the edges of the excitation pattern are masked by noise, loudness perception can remain unchanged.

The loudness of different frequencies within a sound can also influence its overall loudness, due to the uneven sensitivity to different frequency regions mentioned above. As mentioned in the section on auditory scene analysis, singers can enlarge their pharynx cavity and produce a singing formant in that area, making their voices louder and more separated from other instruments.

Spatial factors

The spatial location and spread of a sound can also influence its loudness. In the binaural loudness summation model [Fletcher and Munson, 1933, Stevens, 1972] it is assumed that the loudness of a diotic stimulus is the same as that of an otherwise identical monaural stimulus $10\text{dB}$ greater in sound pressure level, as explained by Cabrera and Miranda [2011]. Moore and Glasberg [2007], however, confirm in a study that loudness summation across ears is not an accurate model, as a diotic sound is less than twice as loud as the same sound presented monaurally. Epstein and Florentine [2012] also criticize the older literature on binaural loudness summation where it is often assumed that a tone presented binaurally is twice as loud as the same tone presented monaurally. They show that binaural loudness summation in the loudspeaker conditions is significantly less than binaural loudness summation in typical
laboratory test conditions using earphones. Roey Izhaki [2008] claims that reproduced sounds appear approximately 3dB louder when panned to the extremes in speakers.

Cabrera and Miranda [2011] found that although diffusivity does affect binaural loudness summation, the loudness effect of direction is greater. They state that in modeling the binaural loudness summation of spatially diffuse stimuli, a binaural gain constant can be observed that is approximately 1 or 2 dB greater than that of the non-diffuse stimuli. The authors derive this fact from a loudness matching experiment involving various filtered pink noise stimuli diffused to simulate eight sound sources evenly distributed in a circle around the listener. The stimuli featured four degrees of diffusivity and were presented over headphones. Furthermore, they state that a larger binaural gain constant exists for stimuli that have a low interaural cross-correlation coefficient, which is associated with a spatially diffuse sound field.

Hirvonen and Pulkki [2008] investigated the effects of interaural time and level differences on loudness. Interaural time differences of up to 1.5 ms did not influence loudness significantly. For very small interaural time differences, the authors assume that this was due to the effect being below the accuracy of the test system and the general loudness discrimination threshold of humans. Interaural level differences caused mainly directional loudness variations due to different sound incidence angles in the horizontal plane. Lastly, Hirvonen and Pulkki [2008] detected a binaural loudness advantage of the diotic compared to the monaural presentation of about 3dB.

The relationship between reverberation and loudness has also been discussed in the literature. Raimond and Watkins [2008] assume that the fact that stimuli with fast attacks and slow decays are judged to be quieter than stimuli with slow attacks and fast decays of equal energy may be because the latter stimuli might be perceptually attributed to room reverberation. Hence, it could be concluded that the loudness of a stimulus is calculated independently from its reverberation tail. Konstantinos and Shen [2006], however, show that the loudness of narrow band noise signal depends only on the sound pressure level at the receiver point and is independent of reverberation time. Therefore, according to the authors, if the same source excites rooms with different reverberation times, loudness is higher in the room with longer reverberation time because it produces higher SPL at the receiver's position. For impulse train stimuli and realistic reverberation time values (larger than 0.1s), loudness depended on the RMS SPL at the receiver's position [Shen & Angelakis, 2006].

Phases

The phases of tones or frequencies in complex tones could also have an impact on their loudness. Griesinger [2013] suggests that the ear is not only sensitive to sound power but also
the sharp peaks present in sounds where all overtones are in phase. He claims that e.g. reverberant sounds, where frequencies are out of phase, appear further away. As mentioned earlier, Helmholz points out that humans are better at establishing the objective nature of objects than observing the sensory stimuli themselves. This may suggest that a sound that is perceived as further away may also appear as quieter.

Loudness adaptation, fatigue

Moore [2012] explains that ear fatigue results from the application of a stimulus in excess of that required to sustain the normal physiological response of the receptor. It can be measured after the stimulus has been removed which is called “post stimulatory auditory fatigue”. Auditory adaptation is the decline of the response of a receptor to a constant stimulus until a steady value is reached [Moore, 2012]. Here, a sound may appear to be getting quieter even though its intensity is constant. Both loudness adaptation and ear fatigue can change loudness perception.

Visual Stimuli

Visual cues can influence loudness perception. Epstein and Florentine [2012] note that binaural loudness summation is significantly less for speech presented through a loudspeaker with visual cues than for stimuli with any other combination of test parameters, such as speech presented via earphones or a loudspeaker without visual cues or speech presented via earphones with visual cues. The authors also observe a subjective effect resulting from expectations about loudness of a familiar, visually present talker, which they call “binaural loudness constancy” [Epstein and Florentine, 2012]. Here, the amount of binaural loudness summation is less for speech from a visually present talker than for recorded speech or tones.

Differences between listeners

Loudness perception also differs between listeners. Hirvonen and Pulkki [2008] found significant differences between individuals in loudness matching when investigating the effects of interaural time and level differences on loudness. The variance between subjects changed depending on the interaural cue utilized where the samples to be matched contained interaural differences. In the monaural case, the deviations between individuals were up to 10 dB according to Hirvonen and Pulkki [2008]. The authors assume that this was due to different strategies adopted by the subjects when comparing the loudness of spatially differing signals.

Loudness perception can also change due to hearing loss. Hearing loss can also change the equal loudness contours of the listener. Furthermore, subjects suffering from hearing loss also often
experience loudness recruitment where the increase in loudness of a sound increasing in volume is perceived to be more drastic than in healthy subjects [Zwicker et al., 1957].

3.5.4 Relation to clarity and separation in mixes

Sounds that appear louder can also be heard better which leads to the assumption that by manipulating the factors for loudness presented in this section, sounds could be made to stand out of the mix. It is possible to increase the loudness of sounds by creating a presence peak though EQ in the mid frequency area. Furthermore, “loudness EQ” can be used on the entire mix to simulate the auditory system's increased sensitivity to very high and low frequencies relative to mid frequencies in louder sounds and the resulting change in tonal balance. In this way, it is sometimes attempted to make a piece of music appear louder.

Terrell, Simpson and Sandler [2013] establish a perceptual mixer based on some of the above loudness findings. They argue that currently no direct mapping between gain and loudness exists in mixing equipment, as it does not take into consideration that loudness is signal and listening level dependent. The track faders also provide no means of controlling the interaction between sound streams, for example in the context of masking [Terrell et al., 2013].

To solve this problem, the authors present an interface that operates within the perceptual domain and takes the listening conditions into consideration. The controls consist of loudness faders for each track, from which loudness ratios are evaluated to determine the loudness balance [Terrell et al., 2013]. The authors define loudness balance as a description of the loudness relationships between the component sounds. They state that it is independent of listening level, and it accounts for the masking interactions between sounds. A master loudness fader is also present, allowing the overall loudness of the mix to be set directly. Findings such as this could contribute to the design of an automatic mixer, where the technological implementation of sound perception by the auditory complex plays a key role.

Ma et al. [2014] assess the effect of partial masking on the loudness of sounds in mixes. They note that when mixing instrument stems together, the perceptual loudness of individual tracks is reduced, depending on individual masking effects. Moreover, the effect of partial masking on the perception of the overall loudness is significant. Interestingly, Ma et al. also found small consistent bias effects related to whether the track in the mix or the solo track was varied, such that the differences at the point of equal loudness obtained in the case of varying the solo track were slightly higher. As stated above, it is likely that both loudness and masking influence clarity, and the latter shows that they also influence each other.
3 Relating psychoacoustic findings to clarity and separation in music mixes

3.5.5 Conclusion

The research questions listed in the introduction can now be answered. These were as follows.
What is loudness? How can it be measured? What factors does loudness depend on, apart from physical sound pressure? Can a sound be made louder without increasing the SPL?

In this section, loudness was defined as an attribute of auditory sensation describing a sound's position on a scale ranging from quiet to loud. Loudness is different from sound pressure or sound intensity and can be measured in e.g. phons or sones.

In phons, the loudness of a 1000Hz tone is defined to be equal to its sound pressure level measured in dBs. The SPL of the 1kHz tone that matches the loudness of a test stimulus is then defined as the loudness of that stimulus, as measured in phons. One sone corresponds to the loudness of a 1000Hz tone at 40 dB, presented binaurally from the front in a free field. At 50 dB, the tone appears to be twice as loud, resulting in a measure of 2 sones. At low levels, loudness changes more rapidly. As a result, loudness as measured in sones can be described as a power function of physical intensity.

Data from experiments such as loudness matching or magnitude estimation tests can be used to develop models of the loudness of simple and complex tones. The factors influencing perceived loudness were summarized above. These are the fundamental frequency of a sound, its intensity, temporal changes in its intensity (its duration, envelope and amplitude modulation), its spectrum, spatial factors, the phases of its partials, loudness adaptation, differences between listeners such as fatigue and hearing loss and the presence of visual stimuli. Hence, loudness can be increased without changing the SPL of a sound.

Clarity and separation was previously defined as the extent to which individual sounds in a mix can be heard. It seems likely that louder sounds in mixes may be perceived as clearer. The loudness of sounds in mixes can be altered by manipulating the established factors and by reducing masking effects. It seems likely, therefore, that these factors might also have an impact on clarity and separation.

3.6 Relating speech intelligibility to clarity in mixes

Music mixes often feature singing and vocal lyrics. Hence, findings on speech intelligibility are likely to be relevant for the measurement of clarity and separation in mixes. Some of the findings may also be transferable to non-speech elements in mixes, such as findings about duplex perception and the ability of the auditory system to recognize very fast sequences of sounds. Therefore, current literature on speech intelligibility will be reviewed, answering the
Relating psychoacoustic findings to clarity and separation in music mixes

following research questions. What factors influence speech intelligibility? Can speech intelligibility be measured or modelled? What are the acoustic attributes of speech and which are most important for speech intelligibility? In section 3.6.1, the acoustic attributes of speech are presented. In section 3.6.2, the factors influencing speech intelligibility are established. In section 3.6.3, cues that the auditory system uses to understand speech are presented. In section 3.6.4, the redundancy of cues in speech is discussed. In section 3.6.5, duplex perception, as introduced in section 3.3.6, is related to speech. Next, in section 3.6.7, models of speech intelligibility are presented. Lastly, in section 3.6.8, several methods for measuring speech intelligibility are presented and discussed. Section 3.6.9 concludes and relates speech intelligibility to clarity in mixes.

3.6.1 Acoustic attributes of speech

In this section, the acoustic attributes of speech will be presented. The findings presented here have been summarized by Moore [2012], unless otherwise specified. Speech consists of complex acoustical patterns and there is no clear and universally accepted way to break these down into atomic perceptual units, as the acoustic characteristic of perceptually identical speech sounds can depend both on the linguistic meaning and the speech sounds surrounding them.

The speech apparatus

Speech consists of harmonic tones and noises. When humans produce speech sounds, they first create an airflow using their lungs and trachea (windpipe). The larynx, where the vocal folds are situated, can be used to create complex tones. When the vocal folds vibrate, the sound produced is a harmonic tone and the resulting speech sound is called "voiced". When the vocal folds do not vibrate, a noisy ("unvoiced") sound, as found in whispers, is produced instead. Turbulent airflows can be created though a constriction or the release of a blockage further up the vocal tract.

The voice sound resonates in the nasal cavities and mouth. It can be shaped into speech sounds in the vocal tract. Movements of the tongue, lips and jaw work together as a complex filter introducing resonances (formants) into the voice sound. The frequency regions in which the formants occur can be altered to produce different vowels. To model this process, formants are usually numbered. The frequencies of the first two formants have the biggest impact on the perceived vowel. The frequency positions of formants are fairly stable over time and do not vary greatly between speakers, hence they can easily be measured and modelled. Formants also exist in whispers.
Consonants are produced by narrowing the vocal tract at certain points along its length and can be classified by the degree and nature of that constriction and/or place of the constriction. This will be shown in more detail after introducing phonemes.

Phonemes

In the literature, words and their syllables are usually broken down into individual speech sounds called phonemes. They are the perceptual building blocks of speech and, hence, the way in which these are distinguished is relevant for the current research project. English, for example, has 40 phonemes that are represented by a set of symbols specified by the international phonetic association [International phonetic association, 1999].

Phonemes are usually categorized according to their perception rather than their acoustic patterns, as the latter can vary according to the context. This phenomenon called "coarticulation" is presented below. Hence, phonemes become abstract, subjective entities. Many phonemes are not pronounceable in isolation and have no meaning unless they are presented in the context of other phonemes. Therefore, according to Moore [2012], several authors argue that phonemes have no perceptual reality and hence should not be classed as basic perceptual units. Phonemes whose acoustic patterns have been altered according to their context are called "encoded", otherwise they are called "unencoded". Unfortunately, no clear dichotomy exists between encoded and unencoded phonemes, making it more difficult to model the acoustic patterns of speech sounds. This is explained in more detail below after introducing categories of phonemes. As mentioned above, phonemes are represented by a set of symbols as presented by the International Phonetic Association [1999]. These describe both vowels and consonants. The nature of vowels has been introduced above.

Consonants are produced by narrowing the vocal tract at certain points along its length. Consonants are categorized into affricates (e.g. “ch” or “j” in English), nasals, (e.g. [m], [n]) approximants (e.g. [l] or [j], as found in “less” or “rest” in English), laterals (e.g. [l]), semi vowels (e.g. [w] in English), fricatives, (e.g. [s] or [z] in English) and stops (e.g. [t], [p] or [g]).

Coarticulation

As touched upon above, the relationship between physical acoustic patterns and phonemes is complex and acoustic patterns vary according to the sounds preceding and following them. This phenomenon is called "coarticulation". The acoustic patterns at the beginning of sounds can differ between words (encoded phonemes) and yet be perceived as the same sound. The
Relating psychoacoustic findings to clarity and separation in music mixes

The phoneme [d] in /di/ and /du/ is perceived as the same speech sound, although in the first case the second formant rises from 2200Hz to 2600Hz, whereas in the second case it falls from 1200Hz to 700Hz. Few invariant cues exist for consonants and vowels often have surrounding consonants merged into them [Liberman et al., 1967].

As pointed out by Moore [2012], it is hard to explain which alterations in coarticulation are acceptable and which are not. Often, coarticulation can improve speech intelligibility in a way that is difficult to measure or model. The patterns of speech sounds in "recognize speech" and "wreck a nice beach" for example, are so similar that machines cannot easily tell them apart [Moore, 2012]. The latter was tested informally on Apple’s voice recognition, which interpreted both phrases as “recognize speech”. Human listeners, however, are very sensitive to the fine differences between them.

According to Moore [2012], the categories and boundaries between encoded and unencoded phonemes in speech perception have evolved in a way that exploits the natural sensitivities of the auditory system. He explains that the boundaries that separate speech sounds lie at those points of the acoustic continuum where discrimination is optimal. He argues that this must be based on certain perceptually, rather than acoustically, equal steps. Smooth transitions are usually perceived as two states separated by a quantal jump at some point between them.

3.6.2 Factors influencing speech intelligibility

According to French and Steinberg [1947], the speech and noise received by a listener depend on the basic characteristics of speech and hearing, the electrical and acoustical characteristics of the instruments and circuits intervening between talker and listener, the conditions under which communication takes place, the behaviour of the talker and listener as modified by the characteristics of the communication system, the conditions under which it is used and the acuity of hearing of the listeners. When speech is presented through any transmission channel, its level and frequency content can be altered. Non-linear distortions, echoes or reverberation can be introduced by e.g. recording or reproduction equipment or by mixing effects. Masking can occur when other sounds or noise are present at the same time as the speech or sung lyrics in a mix. Vasiliauskas et al. [2010] also note that speech intelligibility for binaural reproductions is considerably lower than speech intelligibility in real life. This is due to the fact that the spatial attributes important for reverberation suppression and segregation of sources are impaired in binaural reproductions (i.e. due to the speaker crosstalk, binaural unmasking mechanisms may not work as in real life).
In order to predict the intelligibility of reproduced speech or vocal lyrics, it is useful to establish the cues the auditory system uses to understand speech. Then the audibility of these cues can be assessed to predict speech intelligibility. It is also useful to establish to what extent speech intelligibility differs from the perception of non-speech sounds and to consult current models of speech intelligibility. This is done in the following sections. As above, findings are summarized by Moore [2012] unless otherwise specified.

3.6.3 Cues that the auditory system uses to understand speech

When interpreting speech, the auditory system analyses cues in the dimensions of intensity, frequency and time. A familiar linguistic context and visual cues can also aid speech intelligibility.

According to Fry [1979], the first process in understanding speech is the identification of phonemes. The cues used in this process depend on the native language. A selection of important cues as found in e.g. the English language will be shown below. Fry adds that a great deal of guessing and prediction plays a role in speech perception and that only long practice renders the brain capable of the continuous work of audition prediction and revision that enables it to transform the infinitely complex and variable sound waves of speech into a sequence of words, phrases and sentences.

The acoustic cues presented below are used to decide which phonemes have been spoken. They do not depend on absolute values but on relations between physical quantities [Fry, 1979]. This is illustrated by, for example, the cues for vowel identification where a child's formant frequencies are all higher than a man's. Hence, listeners create a reference against which vowels’ qualities can be differentiated. In natural speech, several cues exist for each phoneme although often only one cue would be enough to understand the phonemes. All important speech cues lie in the frequency region between approximately 120Hz and 8kHz. The fundamental of the human voice lies between approximately 120Hz (male) and 265Hz (child); the most important speech formants range from about 200Hz to 3kHz and noise energy for high frequency plosives and fricatives ranges from about 6kHz to 8kHz [Fry, 1979].

Categories of speech cues

As presented above, speech sounds can be similar to either tones or noise-bursts. Resonances in the vocal tract lead to these sounds having lumps in their frequency spectra and the positions and movements of these lumps (formants) determines the vowels and consonants that result. Cues in the dimensions of intensity, frequency and time help the auditory system identify phonemes. The weight given to the different cues is context dependent but the temporal fine
structure information seems to be particularly important and usable over a wide range of frequencies. The fluctuations in the overall loudness envelope as well as temporal fine structures in each auditory filter are relevant. In the following sections, a selection of categories of speech cues as presented by Fry [1979] will be summarized.

Cues for the manner of phoneme production

Vowels, plosives, affricates, fricatives, nasals, laterals and semi-vowels are each produced differently. Cues for distinguishing these sounds are time and intensity cues, as well as the presence or absence of harmonic tones or noises. Affricates and plosives for example feature a noise component, whereas nasals, laterals and semivowels consist of a continuous tone instead. The extent of the acoustic changes over time, principally changes in the formant frequency and in the noise filtering, play an important role. Fry points out that the second formant transition cue is particularly liable to be used as a cue for the place of articulation. The very rapid acoustic change in plosives for example is reflected in quick changes in that formant frequency.

Time cues

Time cues, such as the duration of silence or noise in speech units can be used to identify phonemes. The interruption in the stream of sound in plosives results in a segment of silence or near silence lasting between 40 and 120 ms in running speech. A small section of noise follows. In affricates, the duration of silence is similar but the noise component is of longer duration than in plosives (70 to 140 ms).

Intensity cues

The intensity variations in speech rarely exceed 7 dB but, although the primary cue to open/close vowel differentiation are formant positions, this can be enough to help the listener distinguish open vowels from closed ones e.g. when a lot of masking noise is present.

Voicing cues

Speech contains cues that indicate if phonemes are voiced or voiceless. The presence or absence of low-frequency energy in the range occupied by larynx vibration is an important cue, especially for consonants in intervocalic positions. When consonants are initial or final in a group, the voice onset time of voiced sounds plays an important role. This is longer after voiceless plosives like ‘p’ in ‘pop’ than e.g. after the voiced ‘b’ in ‘bob’. The noise in consonant production also has a greater intensity in voiceless sounds.
Frequency cues

As explained above, vowels can be identified by their formants, although these vary between speakers. The first two formants play the most important role. As they usually lie between 200Hz and 3kHz, this frequency area is most useful for phoneme identification. As touched upon earlier, the frequency of the second formant is quite important in a lot of cases. Frequency cues are also used for consonant identification; nasals, laterals and semivowels for example do not feature a noise component but instead a continuous tone. Nasals, for instance, are cued by a low frequency resonance and the absence of energy between this frequency band and a band beginning in the region of 2kHz. Laterals and semi-vowels are characterized by vowel-like formats with relatively slow variations in formant frequency, but featuring quicker transitions than semi-vowels [Fry, 1979].

Noise filtering cues

The filtering of noise tells the listener about the location of the noise generator and hence the phoneme produced. For instance, plosives are produced with both lips and have a peak in the low frequency area. Consonants produced with the tip of the tongue at the roof of the mouth (alveolus) feature a high-frequency peak. Sounds pronounced with the back of the tongue near the soft palate feature a mid-frequency peak.

Rhythm and intonation of spoken message

Fry suggests that, in addition to the cues listed above, the rhythm and intonation of the spoken message may also play an important role.

All in all, the temporal fine structure information, namely the fluctuations in the overall loudness envelope as well as temporal fine structures in each auditory filter, especially those containing the first two formants are relevant. Vocoder use this idea of representing speech as a series of band pass filtered signals. Speech is split into several frequency bands and the envelope information in each band is extracted. This information can then be used to modulate e.g. noise bands or sinusoids centred at the respective frequency bands. The result is highly intelligible speech. The idea of using frequency and envelope cues only has been taken to the extreme in the context of sine wave speech. A synthetic speech signal is produced by combining just three sinusoids whose frequencies and amplitudes oscillate in the same way as the first three formants in real speech [Remez, et al., 1981]. The resulting speech signal is intelligible, as long as the listeners have been told that it is supposed to be speech.
3.6.4 The redundancy of cues in speech

In the following, the intelligibility of speech with missing cues is discussed. When some of the sonic elements present in natural speech are missing, speech can often still be understood. This leads to the conclusion that some cues are redundant: no particular components are essential and the information carried by speech is not confined to any particular frequency range. Hence, no single aspect of the speech wave is essential for speech perception. The degree of redundancy in speech can be assessed by eliminating or distorting some of its features, as presented in the following.

The intelligibility of speech with missing cues can be demonstrated in the context of sine wave speech as introduced above. Here, the harmonic structure of speech and the pulsing structure associated with voicing are missing. However, listeners can identify the spoken text as long as they have been told that the signal is supposed to be a speech signal [Remez et al., 1981]. Fig. 3-9 illustrates this phenomenon. Another example is severely distorted speech that can also be understood well, as well as speech containing masked or missing elements.

![Spectrogram of speech](image)

**Fig. 3-9:** Spectrograms of the phrase “Jazz and Swing fans like fast music”, recorded as natural speech (top) and reproduced as sine wave speech (bottom) [Remez, 2008].
3 Relating psychoacoustic findings to clarity and separation in music mixes

The intelligibility of speech with masked elements

Speech is often intelligible even when parts of it are masked or missing. This is the case because listeners usually know what to expect but also because speech signals contain more information than is necessary to understand what is being said [Fry, 1979]. In continuous speech, the auditory system can replace missing words with highly probable words even when there are no acoustic cues indicating what these may be. This demonstrates the importance of the linguistic content [Bagley, 1900]. However, nonsense syllables can be understood too, showing that other cues also play an important role. The intelligibility of masked words is best when the masker contains similar frequencies to the ones present in the masked target. This phenomenon relates to the continuity effect presented in the section on auditory scene analysis (section 3.3).

Experiments involving non-continuous noise as summarized by Moore [2012] show the relationship between the timing and masking effect of such noise. At high interruption rates (above 200 per sec) the masking effect is similar to the effect of continuous noise. According to Kryter [2005], speech intelligibility is accurate as long as the average speech level exceeds that of noise by about 6dB, although the linguistic context and spatial separation of target and noise can increase the intelligibility. When the interruption rates of non-continuous noise lie between 1 and 200 interruptions per second, intelligibility of speech increases as it becomes easier to connect the audible elements to a meaningful signal. At slow interruption rates, this becomes more difficult and the intelligibility drops again.

Moore [2012] distinguishes between energetic and informational masking. In the first case, the neural activity of a signal is similar to that evoked by the masker. In the second case, a signal might not actually be masked energetically but the listener might be confused because two people are speaking at once. This effect is worsened when the two speakers have the same voice characteristics.

As outlined in the section in masking (section 3.4), the masking effect can be reduced when masker and target are in separate locations. According to Freyman et al. [1999 and 2001] this works particularly well when two similar voices mask one another. The authors observed an improvement of 13-30% in that case. For noise maskers, this effect was at just 5-10%.

Cue trading

When contradicting cues are present in a speech signal, a phenomenon called “cue trading” or “phonetic trading” can occur. A change in one cue can be offset by the opposed change in another cue. Moore [2012] points out that this phenomenon can happen for non-speech sounds as well. It is difficult to establish which cues are most important for speech perception as the relationship between contradicting or missing cues and speech intelligibility is complex.
Contradicting cues can also occur when visual and sonic elements are presented simultaneously. In audiovisual integration, both cues are combined to a perceived speech sound. This can differ from both cues: when the word “mama” is played whilst showing a video of a speaker saying “tata”, listeners hear the word “nana” [Mcgurk and Macdonald, 1976].

3.6.5 Duplex perception

As presented in the section on auditory scene analysis (section 3.3), certain cues can influence speech intelligibility even when they have been grouped with the sonic elements belonging to other, non-speech auditory streams. This phenomenon is called “duplex perception”. For example, two formants can belong to two different streams with different fundamental frequencies, leading to the perception of two tones but a single speech sound.

3.6.6 Explaining speech perception

Many aspects of speech perception can be explained in the same way as the perception of non-speech signals, such as phenomena related to masking and auditory scene analysis. Hence, the study of speech intelligibility may also be useful to understand the clarity of non-speech sounds. However, some significant differences can be observed between speech and non-speech perception. Moore [2012] argues that different regions in the brain might be important for speech and non-speech perception. The left hemisphere seems to be most important in speech perception.

Phenomena such as duplex perception occur mainly in the context of speech perception. Furthermore, up to 30 phonemes per second can occur in fast speech [Liberman, 1967] which exceeds the resolution otherwise found in the auditory system. It is possible, however, that this is because phonemes are not actually the basic units of speech perception. At the same time, human listeners can learn to identify sequences of non-speech sounds even when the individual sound segments occur at rates of 100 per second by recognizing the overall sound pattern. Hence, this intelligibility of fast speech could be related to the experience with that language. Samuel [1977] found that training can improve discrimination ability and native speakers are best at understanding their own language. However, as nonsense syllables are also intelligible, learning may not be the only factor. At the same time, infants can already categorize stimuli.

Moore [2012] points out that signals are perceived as either linguistic or non-linguistic and that the perception of a signal as a mixture of both never occurs. He argues that the auditory system switches to “speech mode” when it detects a speech signal, processing the sounds differently from non-speech signals. This was also shown in the context of sine-wave speech [Remez et al., 1981], where some subjects heard a series of hisses and busses, whereas others heard speech.
3 Relating psychoacoustic findings to clarity and separation in music mixes

Once subjects were told that the signal was supposed to be speech, they found it difficult to perceive it differently. Hence, the perception of a signal as speech or non-speech signal may be dichotomous.

3.6.7 Models of speech perception

Moore [2012] points out that speech perception involves processing at many different levels and that separate information at each level may be used to resolve ambiguities or to correct errors that occur at other levels, making it very difficult to model speech perception or intelligibility. He argues that the processing of speech probably does not occur in a hierarchical way but that extensive links exist between levels. Hence, many models of speech perception exist but none of them are generally accepted or complete. In the following paragraphs, examples for such models are presented.

According to Lieberman and Mattingly's [1985] motor theory, listeners perceive the articulatory gestures a speaker is intending to make when speaking. An innate link connects speech production and perception and speech perception depends on the extent to which the listener can understand the intended articulatory gestures of talker. Unfortunately, the authors do not specify how the acoustic signal can be transformed into the perceived gestures.

According to Stevens' [2002] "invariant feature" or "cue-based approach", it should be possible to map acoustic patterns to perceived speech by analysing the acoustic patterns appropriately. Each speech segment is modelled as a set of binary distinctive features. McClelland and Elman's [1986] complex trace model is also based on the idea that the auditory system matches incoming patterns of speech sounds to an existing "database". This happens across several interconnected levels of representation for phonetic features, phonetic segments and words. Higher-level nodes fire when they receive sufficient information from lower-level nodes.

3.6.8 Measuring speech intelligibility

The frequency and loudness variations present in speech over time can be visualized in a spectrogram. As the bandwidths of auditory filters vary with centre frequency, neither wideband, nor narrowband spectrograms can illustrate the perception of voice sounds well. Hence, auditory spectrograms take the representation in the auditory system into consideration, making them a potentially useful aid to predict speech intelligibility. However, they are difficult to interpret, requiring a skilled human to read them. Numerous methods of measuring speech intelligibility exist. In the following, the most relevant current measurements and their limitations will be presented. Subjective measurements are performed by carrying out listening experiments, whilst objective measurements are undertaken without test subjects.
3 Relating psychoacoustic findings to clarity and separation in music mixes

Subjective measurements of speech intelligibility

As indicated above, the subjective measurement of speech intelligibility involves human listeners. This delivers more accurate results than objective measurements but subjective tests are more expensive, time-consuming and work-intensive. Often, logatoms (pseudowords) are used. ITU-T Recommendation P.800 [1996] introduces the absolute category rating method (ACR) for assessing the speech signal in analogue and digital telecommunication channels and speech encoding systems. Simple, short, semantically unrelated sentences spoken by male and female speakers in a room with a noise level below 30dBA are distorted via the communication channel under test. Test subjects rate the listening quality on a scale from 5 (Excellent) to 1 (poor).

The Modified Rhyme Test [House, 1963] provides lists of rhyming or similar sounding monosyllabic English words and listeners pick the word they hear, and this will reveal errors in discrimination between consonant sounds.

As an alternative to subjective quality rating, Huckvale and Hilkhuysen [2012] suggest the use of performance-based testing by measuring the cognitive effort listeners employ to understand speech. Recall accuracy and reaction time can vary with signal quality. The transcript error detection accuracy and processing speed of listeners can hence be a useful method in measuring the intelligibility, especially of high-quality signals. Similarly, Prodi, Visentin and Bellettin [2012] explain that the response time of listeners can be a useful measure for speech intelligibility. Its ratio with word intelligibility is called "listening efficiency" (Prodi, Visentin and Bellettin, 2012). The authors apply the metric to a conference hall under several acoustical conditions, showing that listening efficiency can indeed resolve some ambiguities of e.g. opinion-based rating scales.

Objective measurements of speech intelligibility

As mentioned above, objective measurements of speech intelligibility are more affordable. Many of these techniques assess the impact of the system under test on the temporal structure and the signal-to-noise ratio in processed speech and deliver satisfactory results. Unfortunately, they always simplify speech perception to some extent. Ebem et al. [2011] for example point out that objective speech quality measurements should include the influence of the cultural background of the listener as well as the particular language being spoken. This is not currently the case. The authors show that listeners of Igbo, an African tone language, are more disturbed by additive noise and low listening levels than listeners of American English. Hence, the authors conclude that the low-level parts of the Igbo tone language appear to contain more critically important information than American English, but current objective speech intelligibility
measures do not take this into consideration. In the following, a selection of objective measurements of speech intelligibility will be presented.

The speech transmission index (STI) as introduced first by Houtgast and Steeneken [1971] is standardized by the IEC standard 60268-16 [2011]. It has also been adapted to public address systems (STIPA) and rapid speech (RaSTI). As explained by Humes, et al. [1986], the STI measures the impact of any system (this can be a room, filtering/noise condition, hearing aid or hearing impaired listener) on speech intelligibility by using an artificial test signal consisting of spectrally shaped, random noise with a long term RMS spectrum similar to that of real speech. The signal’s intensity is modulated sinusoidally like a real speech signal (0.63-12.5Hz). The speech to noise ratio as introduced by the system under test is measured across all seven octave bands centered at 125Hz to 8000 Hertz. At the same time, the preservation of modulation in the test signal is measured in each octave band and at each modulation frequency in the output of the system. A modulation transfer function is used to measure the loss and preservation of modulations.

Houtgast & Steeneken [1985] explain that the modulation transfer function (MTF) was first used for the assessment of the performance of optical systems. A lack of sharpness is quantified by an MTF measured with spatially sine-wave modulated light patterns [Houtgast and Steeneken, 1985]. Speech can be described as a flow of sound with a specific distribution pattern of sound intensity over frequency and time. Hence, a comparison between the distribution pattern resulting from the transmission channel under test with that of the original signal can be used to measure the degree of smearing introduced by that system. In rooms, the finer details of the temporal intensity distribution are blurred through reverberation for example.

Houtgast & Steeneken [1985] suggest that the analysis should be performed with a temporally sine-wave modulated test signal and explain that the performance of a sound transmission system as revealed by the MTF can be expressed in the STI. Humes et al., [1986] present the STI as follows:

\[
STI = \frac{\sum_{i=1}^{n} W_i \cdot (SNR_i + 15)}{30}
\]  

(3.8)

\textit{STI} is the speech transmission index, \(n\) is the number of octave bands and \(W_i\) is the specific weighting factor for octave band \(i\). The weighting factors are compared to those for the articulation index after this has been introduced in the following paragraphs (Fig. 3-10). \(SNR_i\) is the signal-to-noise ratio in octave band \(i\), which is based on the MTF for each of the seven octave bands. The result of the calculation is an index that ranges from 0 to 1.
3 Relating psychoacoustic findings to clarity and separation in music mixes

Steeneken and Houtgast [1980] argue that the speech intelligibility index works well for speech recognition in reverberant conditions because it takes temporal parameters into consideration. Other indices, such as the Articulation Index [ANSI S3.5, 1969] explained below, are less useful in that context. However, the authors point out that the weighting factor of the STI is not well suited to all situations. The STI usually performs best when the noise added by the system has similar frequency content to the test signal (e.g. another speech signal or certain types of reverberation). It is, however, less beneficial when the frequency contents differ. Hence, spectrally distorted speech is better assessed by e.g. the Articulation Index (below). In a later paper, Houtgast and Steeneken [1985] suggest a hybrid approach.

Ryan et al. [2013] also state that the STI is well suited for measuring the subjective effects of background noise, reverberation and critical band frequency masking caused by system equalization. They point out, however, that it is unclear whether the STI can account for the effects of discrete echoes or otherwise delayed coherent copies of the original signal. They add that the STI seems to overestimate the degradation to intelligibility caused by multiple arrivals with short delay times (e.g. 5ms). Similarly, Mapp [2013] points out that the Speech Transmission Index (STI) is not a flawless technique and misleading with respect to echoes. Furthermore, he states that the standard speech spectrum assumed by STI often does not replicate the speech spectrum of real announcements. He criticizes RaSTI, stating that it is a poor predictor of STI with respect to sound systems. Different implementations of the simplified STIPA technique led to different results in his experiments. Mapp [2013] adds that while typical amplitude compression that might be applied to an audio signal did not affect the measured STIPA value, sharp limiting reduced the value. Despite its flaws, an STI iPhone application performed surprisingly well given the response of the internal microphone.

The articulation index was coined by Fletcher and Galt [1950] and French and Steinberg [1947]. Based on this, an ANSI standard method was devised [ANSI-S3.5, 1969]. It works similarly to the speech transmission index. The frequency spectrum is divided into twenty bands, where each provides an equal and independent contribution toward the overall speech recognition performance. In each band, the speech to noise ratio between the original and altered test signal determines if that band contributes fully, partially or not at all toward overall speech recognition performance [Humes et al., 1986]. If the signal-to-noise ratio is below 30 dB, the band is not included in the calculation. Above that ratio, the contribution of the band increases linearly for larger signal-to-noise ratios. 12 dB is added to the RMS level of the speech signal in each band in order to represent the peak level of natural speech [Humes, et al., 1986]. The articulation index is calculated with the following equation, where $W_i$ is the weight or
3 Relating psychoacoustic findings to clarity and separation in music mixes

importance of band i and $SNR_i$ represents the difference between the RMS signal level and the RMS noise level:

$$AI = \sum_{i=1}^{n} \frac{W_i (SNR_i + 12)}{30}$$

(3.9)

The similarity between the formulae for the articulation index and the speech transmission index is apparent. The main differences are the differing thresholds for the signal-to-noise ratio (-15dB for the STI and -12dB for the AI) and the differing weighting factors, as shown by Humes et al. [1986] (Fig. 3-10).

Humes et al. [1986] explain that the differences in weighting factors for the indices are due to differences in the databases from which they were developed. While the AI weights were developed from data featuring extreme and abrupt spectral distortion, the STI data base consisted of three similar band pass filtering conditions administered in noise, broadband speech in broadband noise and several forms of temporally distorted broad-band speech administered in broad-band noise [Humes et al., 1986]. The authors point out that the AI is useful for filtered and masked speech but that it needs to be adapted for reverberant speech or otherwise temporally distorted speech.

The speech intelligibility index (SII) has been derived from the AI and it is standardized by ANSI standard S3.05 [ANSI, 1997]. Like the AI, it does not directly include the effect of temporal and non-linear distortions.

Mizumachi [2011] points out that in the context of speech intelligibility, distortion measures should consider the temporal variations of the speech distortion. He investigates the temporal variation of the short-term signal-to-noise ratios of short-term speech distortion based on
Relating psychoacoustic findings to clarity and separation in music mixes

Higher-order statistics, namely variance, skewness, and kurtosis. He notes that the skewness (a measure of the asymmetry of the distribution) of the short-term SNR in particular can explain the discrepancy between subjective methods and current methods of objective evaluation that do not take temporal variations into consideration.

Another example of an objective measure of speech intelligibility is the Percentage Articulation Loss of Consonants (Alcons). It can be computed from measurements of the direct-to-reverberant ratio and the early decay time, making it suitable for the measurement of speech intelligibility in reverberant conditions.

As explained by Brachmanski & Kin [2013], the POLQA (Perceptual Objective Listening Quality Assessment) has been accepted in 2011 as the ITU-T Recommendation P.863 [American National Standards Institute S 3.5, 1997]. It is an objective method that can be used for the measurement of speech intelligibility in telecommunication systems. It is based on an earlier method, PESQ, where the distortion of test signals featuring both female and male voices through a distortion channel is measured. The method takes psychoacoustic findings into consideration, such as the fact that the hearing system features a better frequency discrimination for the lower frequency range in comparison to the higher band and signal-by-noise masking phenomena. Signal levels are measured in sones. The resulting PESQ score can be transformed into a subjective listening quality scale between 1.0 and 4.5.

Gaubitch et al. [2010] explain that the intelligibility of speech in background noise can be measured with a psychometric function (PF) which links the probability of a listener correctly understanding what is being said to the signal-to-noise ratio (SNR). They point out that the PF is often modelled as a sigmoid function that can be parameterized in terms of the SNR corresponding to a defined intelligibility level, Ψ₀, and the slope of the PF at this SNR. In addition, there may be allowance made for guessing and lapses.

3.6.9 Conclusion

In the present section, the findings of speech intelligibility that seem most relevant to clarity in mixes have been summarized. The research questions presented in the introduction can now be answered.

What are the acoustic attributes of speech? Speech consists of harmonic tones featuring resonances (formants) and noises. The characteristic frequencies of formants can easily be mapped to vowels. Speech sounds can be categorized into phonemes, structuring them by their perception rather than their acoustic patterns as the latter can vary according to context (encoded and unencoded phonemes). Phonemes can further be categorized into vowels and
consonants. Consonants are usually subcategorized by the way in which they are physically produced. A complex relationship exists between speech sounds and the acoustic pattern, making it difficult to model speech perception.

What factors influence speech intelligibility? Speech intelligibility depends on the degree to which the linguistic context is known, and on the degree to which the relevant acoustic cues can be heard. The latter can be impaired by the transmission channel, the conditions under which communication takes place, the behaviour of the talker and listener and hearing acuity of the listeners. When speech is presented through any transmission channel, its level and frequency content can be altered. Non-linear distortions, echoes or reverberation can be introduced, masking can occur and noise can be introduced. Speech intelligibility for binaural reproductions is considerably lower than speech intelligibility in real life.

The auditory system uses cues in the dimensions of intensity, frequency and time to understand speech. These include the manner of phoneme production, timing, intensity, voicing, frequency, noise filtering, rhythm and intonation. A familiar linguistic context and visual cues can also aid speech intelligibility. The temporal fine structure information seems to be particularly important and usable over a wide range of frequency areas. Certain cues are redundant, so speech with masked or missing parts can still be understood in many cases.

Can speech intelligibility be measured or modelled? Speech perception differs from non-speech perception but it is unclear exactly how. Many models of speech intelligibility exist but none are generally accepted or complete. Both subjective measurements (based on listening tests) and objective measurements (can be done without listeners) exist for speech intelligibility. In subjective tests, subjects rate speech intelligibility on a quality scale. Alternatively, their reaction time or the accuracy in the transcription of speech is observed. Most objective measurements, such as the STI or the SII analyse the extent to which the temporal fine structure of speech has been skewed and the signal-to-noise ratios in different frequency bands. Current measures for speech intelligibility are not flawless but are reasonably accurate.

It is likely that the above findings on speech intelligibility can be related to clarity and separation in music mixes. Firstly, the intelligibility of sung lyrics may be measured in the same way as the intelligibility of spoken words. Secondly, some of the factors determining the intelligibility of speech could be related to the audibility/intelligibility of sound in general.
3.7 Discussion

In the following discussion, the acoustic parameters influencing clarity and separation, as established above, will be summarized briefly, before answering the following research questions presented in the introduction. Are there any characteristic parameters of clarity and separation that are presented across several areas of research? Is it likely that the factors influencing clarity and separation can be manipulated during the mix process? How do the lower-level parameters of clarity and separation presented previously relate to the factors established in this chapter? The following sections answer these research questions.

3.7.1 Are there any characteristic parameters of clarity and separation that are presented across several areas of research?

The factors influencing clarity and separation can be summarized briefly as follows. Timbral clarity depends on a sound’s spectrum. The number of harmonics, the spectral centroid, the loudness of even harmonics, the spacing of harmonics and the spectral slope appear to play an important role. A greater amount of energy in the high-mid frequency area (2.5—5kHz) appears to lead to greater clarity. This is also the case in sounds that are perceived as bright. In the context of concert halls, reverberation, echoes, noise, tonal distortion, sympathetic ringing tones and non-uniformities of listening conditions counteract clarity, although maximum clarity is not always desirable. Again, the spectrum, as manipulated by tonal distortion, is relevant, as well as spatial factors. In the context of auditory scene analysis, the amplitudes, spatial features and spectra of sonic elements, their onset and offset synchrony, harmonic relations, amplitude and frequency modulation and the centre frequencies of noise bands, as well as the suddenness of the changes of these variables determine whether sonic elements are grouped or segregated. These factors are also likely to determine whether sounds in mixes sound separated. Again, the spectrum of sonic elements and spatial factors are relevant, as well as temporal changes in intensity, frequency and spatial factors. This is also the case in the context of masking where the spectra, modulation, phases of overtones, timing of onsets, reverberation and loudness of sounds determine whether masking occurs. Masking in mixes is likely to impair clarity and separation. The loudness of sounds is linked to their audibility, and therefore clarity. The loudness of sounds depends on their intensity, duration, envelope, amplitude modulation, spectrum, spatial factors and the phases of partials. These factors, again, belong to the categories of the spectrum, spatial factors, intensity and temporal changes of the latter three. Speech intelligibility depends on cues in the dimensions of intensity, frequency and time, where the first two formants (the frequency area between 200kHz to 3kHz) and the temporal fine structure are particularly important. Again, the spectrum and temporal changes therein are
3 Relating psychoacoustic findings to clarity and separation in music mixes

relevant. It is possible that the intelligibility of sung lyrics, or even other sounds in mixes can be related to speech intelligibility.

It can be seen that apart from non-uniformities in listening conditions, all factors can be assigned to the categories *spectrum, spatial factors, intensity* and *temporal changes* (in spectra, spatial factors and intensity), as shown in Table 3-2. Within the scope of the current research project, it is likely that only one of these categories can be investigated in depth. The category ‘spectrum’ is a component of clarity in every research area considered and is therefore likely to be key to clarity in mixes. No other category is covered by all research areas. Increased high-mid frequencies can increase timbral clarity and loudness. Important speech cues also lie in this area. Contrasting spectra are less likely to be grouped or to mask each other. All in all, it appears useful to base further research on the spectral factors influencing clarity and separation. In music mixes, sound spectra are usually altered with EQ. Hence, the impact of EQ on clarity could be tested experimentally. It appears that many of the spectral parameters relate to the relative amount of energy in the different frequency areas of a sound, i.e. the spectrum, increased HF and decreased LF energy, the spectral centroid, spectral slope, the relative energy in the high-mid frequency area, ringing, the loudness of speech cues and tonal distortion. All of these factors can be manipulated with EQ.

Other factors are more difficult to manipulate with EQ, i.e. removing distortion, changing the number of harmonics or altering harmonic relations between instruments. The latter three are usually not manipulated in music mixes at all. The lower level parameters of clarity presented in chapter 2 that could be related to the spectral parameters summarized in the current chapter were mostly based on non-academic literature and often, each author used slightly different words to describe clarity sub-parameters. Therefore, it does not appear useful to test each of these factors individually. By assessing the influence of EQ on clarity, a majority of factors presented in the current chapter can be covered. This is discussed further in chapter 4.
### Relating psychoacoustic findings to clarity and separation in music mixes

<table>
<thead>
<tr>
<th>Timbral clarity</th>
<th>Spectrum</th>
<th>Spatial factors</th>
<th>Intensity</th>
<th>Temporal changes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The Spectrum, increased HF and decreased LF energy</td>
<td>Spatial features</td>
<td>Intensity</td>
<td>Temporal changes</td>
</tr>
<tr>
<td></td>
<td>Number of harmonics</td>
<td>Spectral grouping: Binaural frequency matches</td>
<td>Loudness</td>
<td>Modulation, Timing of onsets</td>
</tr>
<tr>
<td></td>
<td>Spectral centroid</td>
<td>Even harmonics</td>
<td>Intensity</td>
<td>Duration, Envelope, Amplitude modulation, Phases of partials</td>
</tr>
<tr>
<td></td>
<td>Even harmonics</td>
<td>Spacing of harmonics</td>
<td>Spectral grouping: Amplitudes</td>
<td>USAGE</td>
</tr>
<tr>
<td></td>
<td>Evenness of the HF response</td>
<td>Decay time of the midrange frequency response</td>
<td>USAGE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decay time of the midrange frequency response</td>
<td>USAGE</td>
<td>USAGE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spectral slope</td>
<td>USAGE</td>
<td>USAGE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Similar parameters as brightness</td>
<td>USAGE</td>
<td>USAGE</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concert Halls</th>
<th>Noise</th>
<th>Reverberation</th>
<th>Spatial features</th>
<th>Loudness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise, Tonal distortion, Sympathetic ringing tones</td>
<td>Echoes, Flutter echoes</td>
<td>Spectral grouping: Binaural frequency matches</td>
<td>Intensity</td>
<td></td>
</tr>
</tbody>
</table>

| Auditory scene analysis | Sequential grouping: Timbres, Center frequencies of noise bands, Spectral grouping: Harmonic relations, Binaural frequency matches | Sequential grouping: The suddenness of the changes of amplitudes, spatial features, timbres and the centre frequencies of noise bands, Spectral grouping: Onset and offset synchrony, Parallel amplitude modulation, Parallel gliding of spectral components, The principle of common fate, The old plus new heuristic | USAGE |

<table>
<thead>
<tr>
<th>Masking</th>
<th>Spectrum, phases of overtones</th>
<th>Reverberation</th>
<th>Loudness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masking</td>
<td>Reverberation</td>
<td>Loudness</td>
<td>Modulation, Timing of onsets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loudness</th>
<th>Higher energy in the high-mid frequency area generally increases loudness</th>
<th>Spatial factors</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness</td>
<td>Spatial factors</td>
<td>Intensity</td>
<td>Duration, Envelope, Amplitude modulation, Phases of partials</td>
</tr>
</tbody>
</table>

| Speech intelligibility | Cues in the dimensions of frequency, especially in the frequency area between 200 to 3kHz | Cues in the dimensions of intensity | Temporal fine structure |

#### Table 3.2: all parameters that have an impact on clarity and separation can be assigned to the parameters “spectrum”, “spatial factors”, “intensity” and “temporal changes”.  

92
3 Relating psychoacoustic findings to clarity and separation in music mixes

3.7.2 Is it likely that the factors influencing clarity and separation can be manipulated during the mix process?

The four categories shown above can be related to mix parameters. Mix parameters used to alter the spectrum of sounds are equalizers, harmonic exciters (a tool that alters the high frequency content of a sound, e.g. by synthesizing additional harmonics) and distortion (this also produces additional spectral content). Spatial parameters can be altered using reverberation and delay effects, as well as panning. The intensity of sounds can be altered using volume controls, compression and limiters. Temporal changes in frequency and intensity can be controlled using compression, limiters and modulation. Spatial parameters can be manipulated to change over time using automation. Due to the importance of spectral factors shown above, it appears useful to assess the impact of equalization on clarity and separation in mixes as a next step, through experimentation.

How do the lower-level parameters of clarity and separation presented previously relate to the factors established in this chapter?

The present chapter can be related to the previous mix quality chapter, where the parameters of a high quality mix were established. The lower-level parameters of *audibility of dynamic movement, brightness, clear transients, definition, distinctiveness, frequency content, interplay and range, elements conveying instrument timbre and rhythm, intelligibility, lack of masking, localization, lack of clutter, jarring and muddiness, aspects of tonality and timbre, presence and spaciousness* were established to influence clarity. Apart from *jarring*, all parameters can be related to the factors influencing clarity and separation established in this chapter, as shown in Table 3-3.

<table>
<thead>
<tr>
<th>Low-level parameter of clarity</th>
<th>Relation to findings in this chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspects of tonality and timbre</td>
<td>Similarly to <em>frequency content, interplay and range</em>, this appears to be synonymous with the spectra of sonic events (all sections).</td>
</tr>
<tr>
<td>Audibility of dynamic movement</td>
<td>A complex relationship exists between modulation and loudness (section 3.5) and independent source modulation can increase source separation (section 3.3).</td>
</tr>
<tr>
<td>Blend</td>
<td>In mixes, fusion can be used to blend instruments (section 3.3)</td>
</tr>
<tr>
<td>Brightness</td>
<td>Clarity seems to correlate with brightness, as shown in the section on timbral clarity (section 3.1).</td>
</tr>
</tbody>
</table>
## 3 Relating psychoacoustic findings to clarity and separation in music mixes

<table>
<thead>
<tr>
<th><strong>Low-level parameter of clarity</strong></th>
<th><strong>Relation to findings in this chapter</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear transients</td>
<td>The complex relationship between onsets and offsets, loudness and source separation has been established. Sudden rises in loudness attract attention and can hence create separation between sounds (section 3.3). Onset times influence loudness, where sounds with slow onsets and fast offsets are perceived as louder than the opposite in sinusoidal and noise stimuli (section 3.5).</td>
</tr>
<tr>
<td>Definition</td>
<td>Definition may be synonymous with <em>extent to which individual components in a mix can be heard</em> and hence the clarity and separation of sounds in a mix. Definition in concert halls (as measured by the D&lt;sub&gt;50&lt;/sub&gt;) is closely linked to the C&lt;sub&gt;80&lt;/sub&gt; clarity index (section 3.2).</td>
</tr>
<tr>
<td>Distinctiveness</td>
<td>In the same way as definition, distinctiveness may be synonymous to <em>extent to which individual components in a mix can be heard</em>. According the Oxford Thesaurus (Waite, 2008), “distinctive” is linked to “distinguishing”, which could be related to separation between instruments.</td>
</tr>
<tr>
<td>Elements conveying instrument timbre and rhythm</td>
<td>It is assumed that <em>elements conveying instrument timbre and rhythm</em> determine the extent to which instrument timbre and rhythm can be heard which means that this parameter is likely to be a synonym to clarity and separation.</td>
</tr>
<tr>
<td>Frequency content, interplay and range</td>
<td>Frequency content, or spectrum, has been shown to be key to the perceived clarity of single timbres (e.g. higher spectral centroid, correlation with brightness, section 3.1); the combinations of spectra of different sounds can influence masking and auditory scene analysis, where spectral overlap increases the likelihood of masking (section 3.4.2) or spectral grouping (section 3.3.5).</td>
</tr>
<tr>
<td>Fullness</td>
<td>Clarity is opposed to fullness in the context of concert halls (section 3.2).</td>
</tr>
<tr>
<td>Intelligibility</td>
<td>Speech intelligibility (section 3.6) was assessed in detail in this chapter. <em>Intelligibility</em> in general may also be synonymous to the <em>extent to which individual components can be heard</em> and hence <em>clarity and separation</em>.</td>
</tr>
</tbody>
</table>
3 Relating psychoacoustic findings to clarity and separation in music mixes

<table>
<thead>
<tr>
<th>Low-level parameter of clarity</th>
<th>Relation to findings in this chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of “clutter”</td>
<td>The term clutter is interpreted as describing the lack of clarity that can result from sounds with overlapping spectra or sounds that overlap spatially (sections 3.3 and 3.4).</td>
</tr>
<tr>
<td>Lack of masking</td>
<td>Masking was discussed in section 3.4. Masking counteracts clarity and separation and the factors that counteract masking are therefore relevant for the current research project.</td>
</tr>
<tr>
<td>Localization</td>
<td>The influence of spatial factors on clarity and separation was shown in most sections in this chapter (sections 3.2, 3.3, 3.4 and 3.6). Usually, spatial separation, which can only be possible if sounds are localizable, increases clarity and separation.</td>
</tr>
<tr>
<td>Muddiness</td>
<td>According to the Oxford Thesaurus [Waite, 2008] “to muddy” is synonymous to “to make unclear”. Hence, it is likely that muddiness is synonymous with a lack of clarity.</td>
</tr>
<tr>
<td>Presence</td>
<td>Presence may refer to intensity related, spectral or spatial parameters, as discussed in this chapter. When a sound is louder, it is likely to appear both clearer and more present. Spectral factors such as the spectral centroid influence timbral clarity (section 3.1) and spectral overlap increases the likelihood of masking (section 3.4.2) or spectral grouping (section 3.3.5). Here, presence may be affected in similar ways as clarity. Clarity in concert halls (section 3.2), and hence the presence of sounds, can be impaired by reverberation, echoes, noise, tonal distortion, sympathetic ringing tones and non-uniformities of listening conditions.</td>
</tr>
<tr>
<td>Spaciousness</td>
<td>The influence of spatial factors on clarity and separation has been discussed in detail in this chapter (sections 3.2, 3.3, 3.4 and 3.6). Usually, spatial separation increases clarity and separation.</td>
</tr>
</tbody>
</table>

Table 3-3: Previously established low-level parameters of clarity and separation in mixes are related to findings presented in this chapter.

3.8 Summary and conclusions

In the present chapter, the psychoacoustic factors determining clarity and separation have been summarized by consulting literature on timbre (section 3.1), concert halls (section 3.2), auditory scene analysis (section 3.3), masking (section 3.4), loudness (section 3.5) and speech
Relating psychoacoustic findings to clarity and separation in music mixes

Intelligibility (section 3.6). The following research questions, as presented in the introduction, can now be answered.

3.8.1 What factors determining clarity and separation are presented?

Although the definitions of *timbre* and *clarity* are not clearly defined, there seems to be a common agreement on which sounds are timbrally clear. However, it is not specified anywhere what exactly this is. Hence, it is not clear whether timbral clarity is the same as clarity in mixes. Several acoustic parameters influence timbral clarity: the number of harmonics, spectral centroid, relative level of low-order even harmonics, spacing of harmonics, and spectral slope. However, where more than one source dealing with a specific parameter was found, authors come to conflicting conclusions regarding the nature of the relationship, although more HF and less LF frequency content seems to correlate with clarity in the majority of studies. Clarity was also found to be related to brightness, having some of its acoustic parameters in common.

In the context of concert halls, clarity can be described as the audibility of individual elements in musical performances and their separation. This depends on the composition, room acoustics, performers and the auditory acuteness of the listeners. Clarity can be impaired by reverberation, echoes, noise, tonal distortion, sympathetic ringing tones, and non-uniformities of listening conditions, which can be measured to some extent using the C₈₀, C₅₀ and D₅₀ and the EDT. The maximum amount of clarity is usually not desirable as it can reduce fullness and richness.

Auditory scene analysis (ASA) is the process of forming mental representations of individual sounds from the summed waveform that reaches the ears. Cues for the grouping of sonic elements are similarities in their frequencies, pitches, amplitudes, locations, timbres and the centre frequencies of noise bands, the absence of sudden changes of these variables or long durations of silence between sounds, the sound's onset and offset synchrony, the regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation and parallel gliding of spectral components. The absence of these cues leads to the separation of sounds. Attention can also influence auditory scene analysis. In mixes, separation needs to exist between some sounds while others need to blend.

Masking is the process by which the threshold of audibility for one sound is raised by the presence of another, masking sound. It is likely that the clarity and separation in music mixes is related to masking, as masking reduces the extent to which a sound can be heard in a mix. The amount by which the threshold of audibility is raised is expressed in dB. Generally, sounds are likely to be masked when other, louder sounds are present. In the spectral domain, a masker is
most effective when it has energy close to the target's frequency. Narrowband noise can have a stronger effect than harmonic tones. Temporal factors like the comodulation masking release, dip listening, profile analysis and the overshoot effect can create a release from masking. Masking is also possible when the masker occurs before (forward masking) or after the target (backward masking). Spatial cues like interaural time and level differences, spectral differences caused by differences in placement in the median plane, as well as differing temporal or reverberation characteristics of sounds can counteract masking, though this can depend on the use of headphones or speakers. Reverberation can also cause masking.

Loudness is an attribute of auditory sensation describing a sound's position on a scale ranging from quiet to loud. The loudness of sounds is likely to be directly related to the extent to which a sound can be heard, and therefore clarity. It is different from sound pressure or sound intensity and can be measured in e.g. phons or sones. The factors causing increased loudness are fundamental frequencies between 3kHz and 5kHz, an increased intensity, certain temporal changes in intensity; namely longer durations, abrupt onsets (although stimuli with slow onsets and abrupt offsets seem to be louder than stimuli of equal energy featuring reversed on and offsets) and amplitude modulation between 30 and 100Hz. Spectra featuring energy in the high-mid frequency area lead to increased loudness (at the same time, the concept of critical bandwidths is applicable to loudness summation). Further factors include spatial factors (a diotic sound is slightly louder than the same sound presented monaurally and diffuse stimuli are louder than non-diffuse stimuli) and the phases of partials (reverberant sounds, where the partials are out of phase seem to be quieter). Lastly, loudness adaptation, differences between listeners such as fatigue and hearing loss and the presence of visual stimuli (where binaural loudness summation can be reduced) can influence loudness.

Speech consists of harmonic tones featuring resonances (formants) and noises. It is likely that speech can only be clear when it is intelligible. Speech intelligibility increases with the degree to which the linguistic context is known and the degree to which the relevant acoustic cues can be heard. The auditory system uses cues in the dimensions of intensity, frequency and time to understand speech. The manner of phoneme production, timing, intensity, voicing, frequency, noise filtering, rhythm and intonation all contribute. The first two formants and hence the frequency area between 200kHz to 3kHz appears to be particularly important. A familiar linguistic context and visual cues can also aid speech intelligibility. The temporal fine structure information seems to be particularly important and usable over a wide range of frequency areas. Certain cues are redundant, so speech with masked or missing parts can still be understood in many cases.
3.8.2 Are there any characteristic parameters of clarity and separation that are presented across several areas of research? Is it likely that the factors influencing clarity and separation can be manipulated during the mix process?

All parameters that were related to clarity and separation in this chapter can be assigned to the categories *spectrum, spatial factors, intensity* and *temporal changes*. These categories can be related to mix parameters. Mix parameters used to alter the spectrum of sounds are equalizers, exciters and distortion. Spatial parameters can be altered using reverberation and delay effects, as well as panning. The intensity of sounds can be altered using volume controls, compression and limiters. Temporal changes in frequency and intensity can be controlled using compression, limiters and modulation. Spatial parameters can be manipulated to change over time using automation.

As established in the discussion (section 3.7), the spectrum, in particular the higher frequency areas, seems to play a particularly important role in clarity and separation across all areas of literature. Sounds are separated and mask each other less if they have different spectra. Single sounds appear timbrally clearer and louder when they have more high-mid frequency energy and certain important speech cues lie in this area, too. Hence, equalisation and other mix tools that can manipulate the spectrum of a sound are likely to impact on clarity and separation in mixes. It is likely that sounds featuring a presence peak in the high frequency area appear clearer, although separation can be supported by creating contrasting spectra between sounds. Therefore, it seems useful to assess the relationship between clarity and separation and equalisation in music mixes through experimentation.

3.8.3 How do the lower-level parameters of clarity and separation presented previously relate to the factors established in this chapter?

The present chapter can be related to the previous mix quality chapter, where the parameters of a high quality mix were established, as shown in table 3-3. The lower-level parameters of *audibility of dynamic movement, blend, brightness, clear transients, definition, distinctiveness, frequency content, interplay and range, fullness, intelligibility, elements conveying instrument timbre and rhythm, lack of masking, localization, lack of clutter, jarring and muddiness, aspects of tonality and timbre, presence and spaciousness* were established to influence clarity which could be confirmed in this chapter, with the exception of *jarring*. 
4 The influence of dumping bias on spectral clarity ratings (pilot)

Having established the importance of spectra in the context of clarity and separation, it is proposed that listeners rate the effect of boosts and cuts in various frequency areas on spectral clarity, in order to find out more about the relationships between these things. It is likely that some of the factors that affect clarity for single, isolated sources also apply to sounds in mixes. While some of the factors found in the literature are only applicable to sounds in mixes (ASA, masking), there was no indication of any factors that only apply when sounds are played in isolation. Isolated sounds can differ in timbral clarity (chapter 2) and this is likely to be similar to spectral clarity. At the same time, informal experimentation with digital audio workstations suggests that the impact of EQ on the spectral clarity of sounds in mixes can be similar to its impact on the spectral clarity of sounds in isolation. It seems useful to first establish the relationships between particular bands of EQ and the clarity of single sounds (putting to one side the question of separation from other sounds in mixes), before adding the many additional variables that would come with a full mix. Therefore, for the time being, spectral single sound clarity will be investigated, that is the clarity of single sounds, as affected by EQ settings.

In order to evaluate a suitable experimental setup for assessment of the relationship between spectral factors and clarity, a pilot study was carried out, which is summarized here. The goal was to evaluate a suitable experimental setup, and also to test for dumping bias: when subjects are required to rate changes in a single attribute, but also perceive changes in other attributes, this can impact on their reported judgment of the tested attribute (‘dumping bias’, [Bech and Zacharov, 2006] and [Poulton, 1989]).

Dumping biases have primarily been investigated in the context of sensory food evaluation (which can work similarly to auditory perception, where dumping bias has thus far not received sufficient attention [Bech and Zacharov, 2006]). Here, a single negative attribute can influence ratings of other attributes in a negative direction, especially when the negative attribute is not included in the questionnaire [Lawless and Heymann, 2013]. In a study by Wise and Breslin [2011], ratings of ‘sourness’ were influenced by those of ‘saltiness’, especially when test subjects could not explicitly express how they perceived ‘saltiness’.

Dumping bias was thought to be particularly likely in the context of the envisaged experiment since spectral clarity is believed to relate to similar acoustic parameters to those affecting more commonly discussed timbral attributes, e.g. brightness [Schubert et al., 2004, Disley and
The influence of dumping bias on spectral clarity ratings (pilot)

Howard, 2004, von Bismarck, 1974 and Solomon, 1959]. In order to make further experiments run as efficiently as possible, it would therefore be helpful to find out if it is necessary to test not only for clarity, but also other attributes, leading to the research question:

1. Does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity?

Additionally, the pilot study was used to assess whether the proposed methodology was suitable for investigating the impact of boosts and cuts in different frequency areas on spectral clarity, leading to the following additional research questions:

2. Are the chosen stimulus pairs useful and do they have any visible impact on the clarity ratings?
3. Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

In section 4.1, the experimental procedure is outlined, in section 4.2, the results and analysis are presented, focusing on dumping bias. In section 4.3, the other research questions are answered. Section 4.4 concludes and implications for further experiments are discussed. The work on dumping bias presented in this chapter has been published previously (list of publications, Appendix).

4.1 Experimental procedure

In order to assess whether dumping bias can occur when subjects rate the spectral clarity of music stimuli, listeners rated filtered stimuli against an unfiltered reference in terms of clarity only and subsequently in terms of clarity and additional attributes, or in the reverse order (as the stimuli only differed in terms of EQ, the clarity rated by test subjects is synonymous to spectral clarity). Simultaneously, the suitability of the test setup for investigating the relationship between equalisation and clarity was assessed. At the end of both listening tests, people were invited to answer the following four questions:

- Did you find this listening test difficult?
- Did the attributes make sense to you?
- Would you have liked to rate any other attributes that were not there?
- Was the interface easy to use?

The following subsections explain stimulus generation (section 4.1.1), the choice of attributes to be rated (section 4.1.2), the experiment environment and method (section 4.1.3), the user interface (section 4.1.4), and the choice of listening panel (section 4.1.5).
4 The influence of dumping bias on spectral clarity ratings (pilot)

4.1.1 Stimuli

Stimuli were created by equalizing music stimuli. Using Logic Pro 9's parametric equaliser, 9dB boosts and cuts were added to 6 bands centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz respectively (one boost or cut per stimulus), with a Q value of 1.41 (-3dB bandwidth = 1 octave). The resulting stimuli were exported as 44.1kHz 24 bit wave files.

Informal listening indicated that octave bands were fine enough to allow listeners to detect subtle changes in timbre but also wide enough so that no unpleasant or distracting resonances were created. Similarly, 9dB changes provided a good balance between audibility and objectionability. The chosen programme items were a four second guitar loop (“Tell it strumming guitar”, taken from Apple Logic Pro 9) and a nine second piano loop (taken from the author’s composition “Turbulence”): both had dense spectra with energy in all six octave bands.

Ten pairs of these stimuli were randomly selected for use in the pilot study, five piano pairs and five guitar pairs. All possible boosts and cuts were incorporated, comparing boosts with boosts, cuts with cuts and boosts with cuts. The selected stimuli were loudness matched by four audio research professionals. Table 4-1 shows a list of all stimulus pairs.

<table>
<thead>
<tr>
<th>1 lmg4000b.wav</th>
<th>lmg8000c.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 lmg2000b.wav</td>
<td>lmg4000c.wav</td>
</tr>
<tr>
<td>3 lmg8000b.wav</td>
<td>lmg250c.wav</td>
</tr>
<tr>
<td>4 lmg1000b.wav</td>
<td>lmg500b.wav</td>
</tr>
<tr>
<td>5 lmg250c.wav</td>
<td>lmg2000c.wav</td>
</tr>
<tr>
<td>6 lmp4000b.wav</td>
<td>lmp8000c.wav</td>
</tr>
<tr>
<td>7 lmp2000b.wav</td>
<td>lmp4000c.wav</td>
</tr>
<tr>
<td>8 lmp8000b.wav</td>
<td>lmp250c.wav</td>
</tr>
<tr>
<td>9 lmp1000b.wav</td>
<td>lmp500b.wav</td>
</tr>
<tr>
<td>10 lmp250c.wav</td>
<td>lmp2000c.wav</td>
</tr>
</tbody>
</table>

Table 4-1: the stimulus pairs used in the experiment. lmg = loudness-matched guitar, lmp = loudness-matched piano, 250/500/1000/2000/4000/8000 = boost/cut centre frequency (Hz), b = boost and c = cut.
The influence of dumping bias on spectral clarity ratings (pilot)

4.1.2 Attributes to be rated

Additional attributes were chosen that were likely to be familiar to listeners and to be affected by the EQ. Brightness was chosen as it appears to correlate with spectral centroid [Schubert et al., 2004, Disley and Howard, 2004 and von Bismarck, 1974], similarly to clarity [Schubert et al., 2004 and Solomon, 1959]. Warmth appears to correlate with spectral slope and is often considered to be negatively correlated with brightness [Brookes and Williams, 2010]. Fullness can be affected by low-frequency spectral fluctuations [Alluri and Toiviainen, 2010].

4.1.3 Experiment environment and method

The experiment took place in a listening room compliant with ITU-R BS.1116, using Bowers & Wilkins Nautilus 801 loudspeakers. The playback level was adjusted to be comfortable, and produced an average SPL of approximately 69dB Leq (A-weighted). As a first step, test subjects were presented with a familiarization page featuring all stimuli. Next, in order to assess the impact of dumping bias on clarity ratings, the experiment was split into two test halves. In test ‘C’, only clarity was rated. In test ‘F’, the full set of attributes—clarity, brightness, warmth and fullness—were rated. Half of the listeners performed C followed by F; half performed F followed by C. Although there is no universally accepted definition of spectral clarity, a common understanding appears to exist [Disley and Howard, 2004] and therefore, no definition was given to listeners. Similarly, listeners were not provided with definitions of brightness, warmth or fullness.

4.1.4 User interface

A familiarization page and a test interface were created in Max MSP. The familiarization page featured twelve buttons that triggered playback of the twelve stimuli, as well as stop and start buttons (Fig. 4-1). The user interface for the experiment consisted of ten test pages for each of the C and F test halves (Fig. 4-2 and Fig. 4-3). After pressing play, the two stimuli on each page could be auditioned by pressing the buttons labelled ‘A’ and ‘B’. A stop button would reset the stimuli to their beginning positions. Listeners moved sliders to indicate which stimulus sounded clearer, warmer, brighter or fuller and by how much. A ‘next’ button took the listener to the next page of the test.

It is unclear whether spectral clarity can be expressed as an absolute value, i.e. maximum or minimum clarity may not exist. Therefore, listeners were only asked to report relative clarity changes in stimulus pairs. The chosen method did not only allow listeners to report which stimulus they perceived clearer, warmer, brighter or fuller but the magnitude of this difference...
The influence of dumping bias on spectral clarity ratings (pilot)

was also recorded. This made it easier to test whether the clarity differences between stimuli were significant. The slider positions could also be used in a listener consistency check (section 4.2.2). By only presenting two stimuli per page, the complexity of the test could be kept low.

The order of stimulus pairs as well as the order of the stimuli in each pair were automatically randomised to counteract sequential bias and bias caused by the potentially changing motivation of test subjects during the test. The familiarization page also helped to minimize these potential biases [Bech and Zacharov, 2006]. All listeners were provided with instructions prior to the test (Appendix, Figs. A-1 and A-2).

![Familiarization Interface](image)

Fig. 4-1: the familiarization interface.
4.1.5 Listening panel

Bech and Zacharov recommend that at least ten to fifteen suitably trained listeners are employed for listening tests (2006). Therefore, the chosen listening panel comprised 15 male and female undergraduate students and postgraduate researchers of the University of Surrey’s Institute of Sound Recording. The participants were aged between nineteen and twenty-seven years. Experienced listeners are more likely to understand the terminology and to hear small
4 The influence of dumping bias on spectral clarity ratings (pilot)

nuances in the spectra than inexperienced listeners: all participants were experienced in listening tests, and in verbalising sensations of timbre. The undergraduate students had received extensive technical ear training, which included the blind identification of EQ changes. None of the participants reported having any hearing damage. Listeners sat the test individually.

4.2 Results and analysis

Section 4.2.1 determines whether parametric or non-parametric statistical methods will be more appropriate for analysis. Section 4.2.2 then addresses the impact of dumping bias on clarity ratings.

4.2.1 Suitability of parametric statistics

To decide whether to use parametric or non-parametric methods, the following conditions had to be tested:

1. Independence (the data for each level of each variable should be collected independently from that for any other level of the variable)
2. Interval data (all points along the rating scale should be equally spaced)
3. Normally distributed data (the data for each variable should be normally distributed which resembles a bell shape with the mean/median as maximum)
4. Homogeneity of variance (the variance is consistent for each level of each independent variable)

It is unclear whether independence of ratings was fulfilled: the data obtained for each level of each independent variable was obtained independently as participants were asked not to talk about the listening test with anybody until it had been completed. On the other hand, listeners may have compared their answers from different pages. The prevention of other bias has been covered in preceding sections.

The data is interval data, as the distance of the slider from the middle position corresponded with the extent to which one stimulus was clearer than the other.

In order to assess whether the data were normally distributed, histograms and quantile-quantile plots were analysed for all stimulus pairs, ordered by test type. The data were mostly normally distributed, but with some exceptions. For datasets smaller than 2000 elements, it is suggested that the Shapiro-Wilk test is used, otherwise, the Kolmogorov-Smirnov test should be used [Field, 2013]. The current dataset comprises 720 values, hence the Shapiro-Wilk test bears
4 The influence of dumping bias on spectral clarity ratings (pilot)

greater significance: three of the significance values in the Shapiro Wilk test were below 0.05. Similarly, some of the plots indicated a normal distribution of the data, while others did not.

A Levene's test shows that the distribution of data is homogenous with a probability of 0.142. All in all, the data were roughly normally distributed with some exceptions. As a result, it was decided that parametric methods alongside some non-parametric methods could be used. Where values such as the R squared value in variance tests are given, it was further possible to assess the suitability of the chosen parametric method.

4.2.2 The impact of rating other attributes on clarity ratings

The research question "does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity?" can be broken down as follows:

a) Does rating other attributes affect clarity ratings?

b) How does the impact of previous exposure compare to that of simultaneous exposure?

c) Does rating clarity only first bias subsequent full-test ratings?

Clarity ratings will be categorized according to the test half (C or F) and the current order of test completion, as follows:

- CF-C: C completed before F; ratings from C
- FC-C: F completed before C; ratings from C
- CF-F: C completed before F; ratings from F
- FC-F: F completed before C; ratings from F.

The research questions will be addressed by comparing these four groups of results.

Does rating other attributes affect clarity ratings?

In order to answer this question, it would be useful to compare

- CF-C: no exposure to other attributes, with
- CF-F, FC-F, FC-C: exposure to other attributes.

CF-C is the only group of clarity ratings without previous or simultaneous exposure to rating other attributes. Therefore, if dumping bias did occur in this group, this should become apparent when comparing it to the other three. CF-C and CF-F were compared in a paired t-test, comparing the two clarity ratings each listener provided in the two test sections (line 1 in Table 4-2). Three further t-tests were then carried out, for the same pair (line 2) and for the remaining two pairs CF-C vs. FC-F (line 3) and CF-C vs. FC-C (line 4). The means for each stimulus pair
The influence of dumping bias on spectral clarity ratings (pilot) were compared. For all t-tests, the p-value was greater than 0.05, indicating that the mean values of clarity ratings were not affected significantly by dumping bias. Since the conditions for parametric statistical methods were not quite fulfilled, Wilcoxon Rank Sum tests were carried out in addition to the t-tests. The results for this are shown underneath the t-test results in table 4-2. All p-values were greater than 0.05 here also. In notched box plots (Fig. 4-4—Fig. 4-6), the notches always overlap, which means that medians are never significantly different between CF-C and the respective other groups. In the boxplots, the box is the interquartile range and the whiskers include the rest of the data except from outliers (crosses). The notches are the confidence intervals. The notch is centred on the median (line in the middle) and extends to \[ \pm 1.58 \cdot \frac{IQR}{\sqrt{N}} \] where \( N \) is the sample size and \( IQR \) is the interquartile range. The limits that define outliers (crosses) are defined as \( Q_1 - 1.5 \cdot IQR \) and \( Q_2 + 1.5 \cdot IQR \), where \( Q_1 \) is the 25th percentile and \( Q_2 \) is the 75th percentiles.

### Table 4-2: paired t-test and Wilcoxon Rank Sum results: no exposure to other attributes (CF-C) vs. exposure to other attributes.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 CF-C vs. CF-F, on per listener basis</td>
<td>( t(79) = -0.647, p=0.520 )</td>
</tr>
<tr>
<td>2 CF-C vs. CF-F, comparing means for each stimulus pair</td>
<td>( t(9) = -0.819, p=0.434 )</td>
</tr>
<tr>
<td></td>
<td>( W = 6135.5, p=0.299 )</td>
</tr>
<tr>
<td>3 CF-C vs. FC-C, comparing means for each stimulus pair</td>
<td>( t(9) = 0.726, p=0.486 )</td>
</tr>
<tr>
<td></td>
<td>( W = 6203.5, p=0.539 )</td>
</tr>
<tr>
<td>4 CF-C vs. FC-F, comparing means for each stimulus pair</td>
<td>( t(9) = 1.461, p=0.178 )</td>
</tr>
<tr>
<td></td>
<td>( W = 6211.5, p=0.519 )</td>
</tr>
</tbody>
</table>
The influence of dumping bias on spectral clarity ratings (pilot)

Fig. 4-4: notched box plot for CF-C (no exposure to other attributes) and CF-F (simultaneous exposure). ‘b’= boost, ‘c’= cut.

Fig. 4-5: notched box plots for CF-C (no exposure to other attributes) and FC-F (simultaneous exposure). ‘b’= boost, ‘c’= cut.
The influence of dumping bias on spectral clarity ratings (pilot)

It can be concluded that previous or simultaneous exposure to rating other attributes does not affect mean or median clarity ratings significantly. It is, however, possible that the noise or confidence in the ratings may differ. In order to assess this, Table 4-3 shows all average interquartile ranges for the four groups CF-C, CF-F, FC-F and FC-C.

<table>
<thead>
<tr>
<th>Group of ratings</th>
<th>Avg. interquartile range</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF-C</td>
<td>21.5250</td>
</tr>
<tr>
<td>CF-F</td>
<td>14.6833</td>
</tr>
<tr>
<td>FC-F</td>
<td>14.8500</td>
</tr>
<tr>
<td>FC-C</td>
<td>17.7833</td>
</tr>
</tbody>
</table>

Table 4-3: average interquartile ranges for the ratings. (Full ratings range was 0 to 100.)

The interquartile range belonging to CF-C (no exposure to other attributes) is notably bigger than all other interquartile ranges. Hence, rating other attributes seems to have affected the noise in the clarity ratings.

In summary, the exposure to rating the other attributes did not affect mean or median clarity ratings significantly but it did lead to a reduction of the noise in the ratings. As explained further
The influence of dumping bias on spectral clarity ratings (pilot)

in section 4.3.2, the selection of additional attributes appeared to be suitable and there was no missing attribute that all listeners agreed on. Ratings from trials with simultaneous and previous exposure are compared in the next section.

How does the impact of previous exposure compare to that of simultaneous exposure?

This question can be answered by comparing

- FC-C: previous exposure to other attributes, with
- FC-F, CF-F: simultaneous exposure to other attributes.

As above, the two means for each stimulus pair are compared in t-tests and Wilcoxon Rank Sum tests (Table 4-4). Once again, each p value is greater than 0.05. In the notched box plots, all notches overlap again (Fig. 4-7 and Fig. 4-8). Therefore, simultaneous exposure to rating the other attributes does not deliver significantly different results from those obtained from a test with previous exposure.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 FC-C vs. FC-F, comparing means for each stimulus pair</td>
<td>t(9) = 0.816, p=0.436</td>
</tr>
<tr>
<td></td>
<td>W = 4932.5, p=0.993</td>
</tr>
<tr>
<td>2 FC-C vs. FC-F, on per listener basis</td>
<td>t(69) = 0.966, p=0.338</td>
</tr>
<tr>
<td>3 FC-C vs. CF-F, comparing means for each stimulus pair</td>
<td>t(9) = -1.368, p=0.204</td>
</tr>
<tr>
<td></td>
<td>W = 4874, p=0.122</td>
</tr>
</tbody>
</table>

Table 4-4: paired t-test and Wilcoxon Rank Sum results: previous exposure to other attributes (FC-C) vs. simultaneous exposure.

In terms of noise in the ratings, it can be seen in Table 4-3 that FC-F and CF-F (simultaneous exposure) have average interquartile ranges noticeably lower than that for FC-C (previous exposure). However, this is still considerably lower than that for CF-C (no exposure to other attributes). Hence, of the two, simultaneous exposure to other attributes appears to be a more beneficial tool for noise reduction in clarity ratings, but previous exposure is also useful.

The effects discussed above can be observed in a different way: for each of the two groups of listeners, listener consistency can be tested. CF-C and CF-F: listeners in this group had no exposure to other attributes in their first test, which means that this group of ratings is the noisiest in this experiment. In their second test they had simultaneous exposure to other attributes, leading to the least noisy data set. Clarity ratings from this group's C (first) and F
The influence of dumping bias on spectral clarity ratings (pilot)

(second) tests were plotted against each other in a scatter plot (Fig. 4-9), which shows the resultant lack of inter-test correlation.

FC-F and FC-C: listeners in this group had simultaneous exposure to other attributes in their first test and previous exposure in their second test which means that the ratings for these two tests have similar noise levels—both lower than the levels for ratings with no exposure to other attributes. Plotting clarity ratings from this group’s F (first test half) and C (second test half) against each other (Fig. 4-10) shows much greater inter-test correlation than in the previously tested pair. This once again shows that simultaneous and previous exposure to other attributes impact similarly on clarity ratings.

Fig. 4-7: notched box plots for FC-C (previous exposure) and CF-F (simultaneous exposure). ‘b’= boost, ‘c’= cut.
4 The influence of dumping bias on spectral clarity ratings (pilot)

Fig. 4-8: notched box plots for FC-C (previous exposure) and FC-F (simultaneous exposure). ‘b’ = boost, ‘c’ = cut.

Fig. 4-9: scatter plot comparing CF-C and CF-F.
Does rating clarity only first bias subsequent full-test ratings?

In the previous sections, previous or simultaneous exposure to rating other attributes was shown to reduce the noise in clarity-only ratings. It is also possible that previous exposure to rating clarity in isolation may create biases that affect subsequent full-test ratings. This is tested in this section. Here, it would be useful to compare

FC-F: full-test ratings, with no previous clarity-only rating; with

CF-F: full-test ratings, after clarity has been rated in isolation.

These two data sets contain ratings given by the same listeners. A t-test shows that ratings are significantly different between FC-F and CF-F: t(9)=-2.684, p=0.025 and when presented in a box plot (Fig. 4-11), the FC-F medians appear to be lower than the CF-F medians. This seems to suggest that stimulus A was rated as clearer on average in the FC-F sample. However, as all notches overlap, this observation is unlikely to be significant. The significance found in the t-test may be due to the fact that the data were not quite normally distributed. The Wilcoxon Rank Sum test does not indicate a significant difference (W=4864.5, p=0.113).

The interquartile ranges (Table 4-3) for FC-F and CF-F are almost identical, indicating no impact of previous exposure to clarity-only ratings on statistical noise.
4 Factors that need to be considered before planning experiments

4.3 Further aims of the pilot study

Apart from assessing the impact of dumping bias on clarity ratings, the pilot study aimed to answer the following questions:

1. Are the chosen stimulus pairs useful and do they have any visible impact on the clarity ratings? Are the boosts and cuts and durations of the stimuli and the programme items useful?

2. Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

These questions can now be answered.

4.3.1 Are the chosen stimulus pairs useful and do they have any visible impact on the clarity ratings?

In order to answer this question, an ANOVA test was conducted in SPSS, using the stimulus pairs and listeners as factors (Table 4-5). The resulting R squared value was large (0.811, adjusted 0.622) and the significance was around 0.000. It can also be seen that the stimulus pairs had a much bigger effect on clarity ratings than differences between listeners or a combination
Factors that need to be considered before planning experiments

between listeners and stimulus pairs. This means that while listeners are not in complete agreement, the effects of EQ are observable and significant and exceed the effect of differences between listeners. Therefore, subjects understood the term 'clarity' and the chosen stimulus pairs appear suitable to detect differences in perceived clarity. In the various boxplots presented in section 4.2.2, it can be seen that the guitar and piano ratings differed slightly, so it is possible that the programme items had an influence on the clarity ratings. Hence, it appears useful to include both in further listening tests.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>64970.187</td>
<td>149</td>
<td>436.042</td>
<td>4.308</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>393276.813</td>
<td>1</td>
<td>393276.813</td>
<td>3885.367</td>
<td>.000</td>
</tr>
<tr>
<td>StimulusPair</td>
<td>23251.787</td>
<td>9</td>
<td>2583.532</td>
<td>25.524</td>
<td>.000</td>
</tr>
<tr>
<td>Listener</td>
<td>9690.887</td>
<td>14</td>
<td>692.206</td>
<td>6.839</td>
<td>.000</td>
</tr>
<tr>
<td>StimulusPair * Listener</td>
<td>32027.513</td>
<td>126</td>
<td>254.187</td>
<td>2.511</td>
<td>.000</td>
</tr>
<tr>
<td>Error</td>
<td>15183.000</td>
<td>150</td>
<td>101.220</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>473430.000</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>80153.187</td>
<td>299</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-5: ANOVA test with the stimulus pair and the listeners as factors. R Squared = .811 (Adjusted R Squared = .622)

4.3.2 Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

The Max/MSP interface and reproduction equipment all worked properly. Differences between the stimulus pairs were audible in the given setup; hence their impact on the clarity ratings could be observed as shown in previous sections. The duration of the entire listening test was on average half an hour per test subject, which was short enough to ensure that clarity ratings were not significantly affected by ear fatigue or a loss of motivation but also long enough to be able to include a suitable number of stimuli. In the following, the answers given by test subjects to the questions at the end of the listening test are presented.

- ‘Did you find this listening test difficult?’

Occasionally, listeners commented on the fact that they found it difficult to decide how to translate what they perceived into exact slider positions, as there was no reference. Some listeners commented on the fact that they found it difficult to rate fullness, and occasionally to
Factors that need to be considered before planning experiments

tell warmth and fullness, as well as clarity and brightness apart, as they were each correlated. However, in general, test subjects reported that they found the task easy.

- ‘Did the attributes make sense to you?’
All listeners stated that they knew what the attributes meant to them.

- ‘Would you have liked to rate any other attributes that were not there?’
When asked if they would have liked to rate any additional attributes, individual listeners mentioned annoyance, preference, audio quality, hollowness, harshness and muddiness (one listener each). Each listener had to think about their response and most appeared to name further attributes in order to help, not because these particularly stood out as missing. There was no missing attribute that any two listeners agreed on.

- ‘Was the interface easy to use?’
All participants found the interface easy and intuitive to use.

4.4 Conclusion

Having established the importance of spectra in the context of perceived clarity and separation, it was suggested to assess the impact of boosts and cuts in different frequency regions on the spectral clarity of single sounds in a listening test. In order to prepare for such a listening test, a pilot study was conducted to address the following research questions:

- Does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity?
- Are the chosen stimulus pairs useful and do they have any visible impact on the clarity rating?
- Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

Two loudness-matched programme items (guitar and piano) featuring 9dB boosts and cuts in the 6 octave bands centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz were presented to listeners via high quality studio loudspeakers in a treated listening room. After a familiarization stage, listeners were asked to rate pairs of stimuli in two separate tests. In the first test, only clarity was rated. In the second test, clarity, warmth, fullness and brightness were also rated in order to investigate the possibility of dumping bias. One half of the listeners undertook the clarity-only test first, followed by the full test; the other half undertook the tests in the opposite order.
4 Factors that need to be considered before planning experiments

4.4.1 Does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity?

Dumping bias is an effect whereby listeners’ ratings of a specific attribute are affected by an inability to rate other changes that are perceived. One aim of the pilot study was to establish whether ratings of spectral clarity would be affected by asking listeners to also rate other attributes.

The mean and median clarity ratings did not depend on previous or simultaneous exposure to rating the other three attributes; however, the noise in the clarity ratings was reduced if other attributes were rated simultaneously or had been rated previously. Simultaneous exposure to rating other attributes reduced the noise a little more than previous exposure.

Therefore, in future single-attribute rating experiments, it could be beneficial to ask subjects to also rate additional attributes, or even just to ask them to engage with other attributes, in order to reduce the resulting statistical noise in the single-attribute ratings. If all possible pairs of stimuli were compared for all attributes, listening tests would be too long. Concentration and motivation levels would drop and ear fatigue may occur. If, however, the tests are carried out in several sessions, it is possible that subjects are not available for each session. Hence, it would be useful to allow subjects to rate warmth, fullness and brightness only once or to merely let them engage with the other attributes to make the listening test more efficient. It has been shown that the noise in the ratings can still be reduced in this way as opposed to only rating clarity.

4.4.2 Are the chosen stimulus pairs useful and do they have any visible impact on the clarity rating?

Significant differences were observed between the clarity ratings for the different stimulus pairs and therefore, it appeared that the stimulus pairs were suitable to investigating the impact of boosts and cuts in different frequency areas on single sound clarity. The effect of the stimuli on clarity significantly exceeded that of differences between listeners. The guitar and piano ratings differed slightly, so it is possible that the programme items had an influence on the clarity ratings. Hence, it appears useful to include both in further listening tests.

4.4.3 Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?

The listening test took on average half an hour. In general, test subjects reported that they found the task easy and the interface easy to use. The attributes were meaningful to them and there
4 Factors that need to be considered before planning experiments

was no common agreement on any specific further attributes that were missing. Hence, it was concluded that the chosen stimuli and the setup were suitable to the research aim. The interface, facilities and reproduction equipment worked fine.
5 Assessing the contribution of different octave bands to the single sound spectral clarity of piano and guitar stimuli

The importance of spectra in the context of the perceived clarity of sounds was established in chapter 3. In order to investigate the impact of spectral equalisation on clarity, it was decided to first establish whether single sounds could be equalised in order to sound clearer (chapter 4). The current listening test aimed to answer the research question: how do boosts and cuts in different frequency regions influence single sound spectral clarity?

In section 5.1, the setup for the new listening test is presented and in section 5.2 the results are introduced. Drawing on the literature, the findings are used to discuss implications for further listening tests in section 5.3. Section 5.4 is a conclusion.

5.1 The setup

In order to assess the influence of spectral equalisation on clarity, listeners rated the clarity of filtered stimuli against an unfiltered reference. Section 5.1.1 presents the stimuli used, the inclusion of further attributes is explained in section 5.1.2, section 5.1.3 introduces the user interface, the experiment environment and method are introduced in section 5.1.4 and the listening panel is presented in section 5.1.5.

5.1.1 Stimuli

The four-second guitar loop ("Tell it strumming guitar", taken from Apple Logic 9) and the nine second piano loop (taken from the author's composition “Turbulence”) that were used in the previous pilot study (chapter 4) were also used in the listening test. Both programme items are suitable for adding boosts and cuts across the six octave bands; both have fundamentals low enough to cover the 250Hz band and span several registers. The waveforms and spectra of the programme items are shown in Figs. 5-1 – 5-4. Long-term average spectra were calculated from the average power spectral density (PSD) obtained from a series of overlapping Fast Fourier Transforms (FFT). The FFT length was 4096, and the hop size was 2048. The segments of the signal were Hann-windowed. The average PSD was Gaussian-smoothed to 1/6-octave resolution.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

Fig. 5-1: smoothed long-term average spectrum of the unequalised guitar programme item.

Fig. 5-2: smoothed long-term average spectrum of the unequalised piano programme item.
Assessing the contribution of different octave bands to clarity (guitar, piano)

No equalisation was added to the reference stimuli. Stimuli were created by equalizing music stimuli. For each equalized stimulus, a single 9dB boost or cut was added to one of the six octave bands centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz respectively (Fig. 5-5). This was done using Logic Pro 9’s parametric equaliser, with a Q value of 1.40 (-3dB bandwidth = 1 octave). The resulting stimuli were exported as 44.1kHz 24 bit wave files. Informal listening indicated that these octave bands were fine enough to allow listeners to detect resulting changes in timbre but also wide enough so that no unpleasant or distracting resonances were created. Similarly, it was observed that 9dB boosts and cuts were enough to produce detectable changes without becoming unpleasant. The fact that the chosen boosts and cuts lead to useable results was confirmed in the pilot study (Chapter 4). The stimuli were loudness-matched by four audio professionals, as in the pilot study.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

Fig. 5-5: creating boosts and cuts in Logic 9.

5.1.2 The choice of additional attributes

In addition to clarity, listeners also rated stimuli in terms of warmth, brightness and fullness. The motivation for including these other attributes came from the finding that previous or simultaneous exposure to rating these other attributes can reduce the statistical noise in clarity ratings (chapter 4). Brightness was chosen due to the existing evidence that it correlates with spectral centroid [Brookes and Williams, 2010] which might also be the case for clarity (chapter 3). Warmth appears to correlate with spectral slope and is often considered to be negatively correlated with brightness [Brookes and Williams, 2010]. Fullness can be affected by low-frequency spectral fluctuations [Alluri and Toiviainen, 2010].

5.1.3 User interface

Listeners rated the filtered stimuli against the unfiltered reference in terms of clarity, fullness, warmth and brightness. In this way, the ratings of different stimulus pairs could be compared to each other. On each GUI page, listeners could trigger the equalized stimulus and reference by pressing buttons labelled ‘A’ and ‘B’ as often as required after pressing ‘play’. A ‘stop’ button could be used to reset the stimuli to their beginning position. The stimuli were compared in terms of the perceptual attribute given at the top of each page. Clarity, fullness, warmth or brightness ratings were recorded as slider positions. The position on the sliders indicated if and by how much one of the stimuli evoked the greater sensory strength for each of the perceptual attributes. Fig. 5-6 shows an example of a test page.
Each rating was undertaken twice for both programme items. Nine pages were added to assess the additional perceptual attributes (three pages each), resulting in fifty-seven pages. Each listener was given an instruction sheet before undertaking the listening test (Appendix, Fig. A-3). The test was preceded by the same familiarization interface (Fig. 5-7) as was used in the pilot test (chapter 4), allowing each subject to audition all stimuli. Afterwards, they were given the opportunity to try out the test interface before starting the test. Like in the pilot test, no definition was given for clarity because, although there is no commonly accepted definition for clarity, a common understanding appears to exist of what clarity means (chapter 3). Furthermore, test subjects had agreed that they knew what the attributes meant in the pilot study.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

5.1.4 The presentation of the stimuli to test subjects

In the same way as in the pilot test, the test interface and incorporated stimulus pairs were presented at a comfortable listening level in an international-standard listening room built to ITU-R Standard BS1116, using Bowers & Wilkins Nautilus 801 speakers. Pink noise was loudness-matched to the average stimulus loudness by four audio professionals, resulting in an SPL of 69dB $L_{eq}$ [A-weighted average sound pressure level, IEC 61672, 2013] at the chosen listening level. The lack of extraneous noises ensured that important nuances in the spectra of the stimuli were audible.

5.1.5 Type of listeners

As in the pilot study, 19 students of the Surrey University Institute of Sound Recording (some undergraduates and some postgraduate researchers) were recruited for the listening test. The benefit of using professional subjects, rather than untrained listeners is the fact that audio professionals are likely to understand the terminology and to hear small nuances in the spectra. Olive [2003] states that audio professionals could discriminate between loudspeakers better and were more critical than inexperienced listeners. Similarly, subjects in an experienced group of listeners were more in agreement than subjects in a group of inexperienced listeners in an experiment involving loudspeaker evaluation conducted by Bech [1992]. Bech concludes that audio professionals have a higher ability to repeat ratings of the same stimulus than untrained listeners.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

5.2 Presentation of the results

In the following, the results of the listening test will be presented. Section 5.2.1 shows that the prerequisites for using parametric methods of statistics are not quite fulfilled. In section 5.2.2 the contribution of boosts and cuts across the six octave bands is shown in boxplots.

5.2.1 Prerequisites for using parametric methods

To decide whether to use parametric on non-parametric methods, the following conditions had to be tested, as explained further in chapter 4:

1. Independence
2. Interval data
3. Normally distributed data
4. Homogeneity of variance

Independence of clarity ratings is unlikely to have been fulfilled. The data obtained for each level of each independent variable was obtained independently as participants were asked not to talk about the listening test with anybody until it had been completed, but listeners most likely compared their answers from different pages. The data is interval data, as the distance of the slider from the middle position corresponded with the extent to which one stimulus was clearer than the other. In this way, the distance between two possible fader positions corresponded to a fixed difference in clarity ratings.

In order to assess whether the data were normally distributed, histograms were analysed for all stimulus pairs. The data were mostly normally distributed, but with some exceptions. For datasets smaller than 2000 elements, it is suggested that the Shapiro-Wilk test is used as normality test, otherwise, the Kolmogorov-Smirnov test should be used [Field, 2013]. The current dataset comprises 912 values, hence the Shapiro-Wilk test was used. One of the significance values in the Shapiro-Wilk test were below 0.05, indicating a mostly normal distribution. However, while some of the plots indicated a normal distribution of the data, others did not (Fig. 5-8).
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

Fig. 5-8: two examples of the histograms used to assess normality of the data. The clarity ratings are approximately normally distributed for stimulus pair 11 but slightly less for stimulus pair 12.

As a next step, it was assessed whether the elimination of potentially inconsistent listeners would lead to a more normal distribution. As outlined in section 5.1, each listener undertook each rating twice. The notched box plot in Fig. 5-9 shows the distribution of differences between the ratings for all stimulus pairs and for each listener.
It appears that the ratings given by listeners four, five and six are notably less consistent than the other listeners’ ratings. For all other listeners, most pairs are no further apart than about 26 data points but listeners four, five and six have several pairs that are further apart by at least 10 points. At the same time, the medians for their distributions of differences are noticeably higher than for all other listeners.

The distribution of the data was tested again for normality after removing the ratings belonging to listeners four, five and six. Presumably due to the reduced number of data points, three of the significance values in the Shapiro Wilk test were now below 0.05, more than in the previous analysis.

All in all, the data were not quite normally distributed. Levene’s test shows that the variance of the data is not homogenous, with a significance of 0.000. Hence, mostly non-parametric methods were used with the current data set (plots).

5.2.2 The contribution of octave bands to clarity

In order to assess the impact of boosts and cuts across all octave bands to clarity, the data are presented in four notched boxplots. As the data do not appear to be normally distributed or homoscedastic, the medians, rather than the means of the clarity ratings for each pair are compared. First, the distribution of all ratings (piano and guitar) for each boost and cut are
Assessing the contribution of different octave bands to clarity (guitar, piano)

presented in a box plot (Fig. 5-10). Here, a clear trend is evident as many confidence intervals (notches) do not overlap and the medians differ considerably for each stimulus pair. When 4kHz and 8kHz are boosted, clarity increases significantly and the notch lies completely above 0. For boosts in all other frequency bands, the reference was perceived to be clearer, as the notches lie below 0. Similarly, 250 and 500Hz cuts contribute to clarity, while all other cuts reduce clarity.

When the ratings belonging to the three least consistent listeners are removed (Fig. 5-11), the conclusion is the same, although most of the interquartile ranges are smaller. Hence, there is less noise in the ratings. It appears that high frequencies contribute to clarity positively and low frequencies contribute to clarity negatively. The cutoff point is somewhere between 2kHz and 4kHz, at least for the chosen programme items.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

Fig. 5-11: distribution of clarity ratings for all boosts (white) and cuts (grey) in comparison to the unequalised reference stimuli, excluding the three least consistent listeners. Here, both piano and guitar ratings are included in each distribution.

In higher frequencies, boosts contribute to clarity positively and lower frequencies negatively; the opposite is the case for cuts. The medians for both the boost and cut boxes are arranged in a diagonal line that becomes flatter at both ends. Hence, a 250Hz boost seems to reduce clarity more than a 500Hz boost, while respective cuts in these areas have roughly the same impact as each other. 2kHz and 4kHz cuts reduce clarity considerably more than 8kHz cuts and 4kHz and 8kHz boosts seem to contribute equally to clarity.

Although higher frequencies generally seem to contribute to single sound spectral clarity while lower frequencies reduce clarity, it is possible that the exact impact of each frequency region depends on the programme item. The bandwidth of the spectrum could influence the impact each frequency region has on clarity and other factors like the spectral centroid or the fundamental frequency could alter the cutoff point between frequencies that contribute to or reduce clarity. In order to find out more about such possible phenomena, the long-term average spectra of the unequalised piano and guitar programme items are compared (Fig. 5-12). Long-term average spectra were calculated in the same way as explained in section 5.1.

The two programme items have similar spectral slopes but the piano programme item has a slightly higher range of fundamental frequencies and less energy above about 4kHz, especially in the 8kHz octave band. Hence, it appears useful to assess the contribution of the different octave bands to spectral clarity separately for piano and guitar stimuli.
Fig. 5-12: smoothed long-term average spectra of the unequalised guitar and piano programme items.

As above, assessments showed that perceived clarity is similar with and without the less consistent listeners but the noise in the ratings is reduced slightly. Some differences exist between the guitar and piano ratings, however: for the piano programme item, a 4kHz boost contributes to clarity which is not the case in the guitar programme item (Fig. 5-13), maybe due to the wider range of fundamental frequencies in the piano. The contribution of the 8kHz octave band on clarity is greater for the guitar programme item, possibly because it has more energy in that area. Here, both the extent to which the boost adds to clarity and the extent to which the cut reduces clarity (Fig. 5-14) are greater than in the piano programme item. There is less noise in the piano ratings overall, so it appears that listeners find it easier to rate clarity in pianos, especially for the low frequency cuts.
5 Assessing the contribution of different octave bands to clarity (guitar, piano)

Fig. 5-13: distribution of clarity ratings for all boosts in comparison to the unequalised reference stimuli, plotted separately for piano (grey) and guitar (white) stimuli.

Fig. 5-14: distribution of clarity ratings for all cuts in comparison to the unequalised reference stimuli, plotted separately for piano (grey) and guitar (white) stimuli.

5.3 Relation to the current literature and implications for future listening tests

As mentioned in chapter 3, the spectral clarity of single sounds has been found to depend on the balance between high and low frequencies in previous studies. However, authors come to
5.4 Conclusion

Having established the importance of spectra in the context of perceived clarity and separation, the impact of boosts and cuts in different frequency regions on the spectral clarity of single sounds was assessed in a listening test, where test subjects rated stimulus pairs for clarity, warmth, fullness or brightness. Two programme items featuring 9dB boosts and cuts across the 6 octave bands centered at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz were compared to unequalised reference stimuli. The aim was to answer the research question: how do boosts and cuts in different frequency regions influence single sound spectral clarity?

As the data gathered were not quite normally distributed, nor independent, conclusions were drawn by assessing four notched box plots, rather than by using parametric statistical methods. When removing the three least consistent listeners, the noise in the ratings was reduced slightly but the conclusions were the same.

Generally, when the 4kHz and 8kHz bands are boosted, clarity increases significantly. For boosts in all other frequency bands, the reference is perceived to be clearer. Hence, it appears that high frequencies contribute to clarity positively and low frequencies negatively. The point from
which the contribution becomes positive is somewhere between 2kHz and 4kHz. This is in line with the findings summarized in the literature review (chapter 3).

Some differences exist between the guitar and piano ratings. For the piano programme item, a 4kHz boost contributes to clarity positively but this is not the case in the guitar programme item. The contribution of the 8kHz octave band to clarity is greater for the guitar programme item, possibly because it has more energy in that area. Although the two programme items have similar fundamental frequencies and spectral slopes, the piano programme item has less energy above about 4kHz, especially in the 8kHz octave band. Here, both the extent to which the boost adds to clarity and the extent to which the cut reduces clarity are greater than in the guitar programme item. There is also less noise in the piano ratings overall, so it appears that listeners are in closer agreement in their piano ratings, or that they might be more confident about rating the piano stimuli, especially for the low frequency cuts.

It is suggested the impact of differences in the spectra of programme items on the relationship between EQ and clarity ratings is assessed in more detail in future listening tests. It is suggested that stimuli with different fundamental frequencies and spectral centroids are used and equalized in the same way as in the current experiment.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Spectra play an important role in the context of clarity. In a previous listening test, it was shown that boosts to higher octave bands contribute to clarity positively, while boosts to lower octave bands contribute negatively. The point at which the nature of the contribution changed from negative to positive differed between the programme items, however. It was concluded that differences in the spectra of sounds—such as their spectral slopes, their fundamental frequencies or the bandwidths of the spectra—are likely to influence the way in which boosts and cuts impact on clarity. Hence, it was decided to repeat the experiment, assessing the impact of the fundamentals and initial spectral centroids of programme items on the relationship between EQ and clarity. The spectral centroid is also a commonly accepted measure of brightness [e.g. Schubert, 2006] which has been shown to correlate with clarity [Disley and Howard, 2004].

The current listening test aimed to answer the research question: how can changes in the spectral clarity of equalized programme items with differing fundamentals and original spectral centroids be predicted?

In section 6.1, the setup for the new listening test is presented and in section 6.2, the results are introduced. In section 6.3, the established predictors of single sound clarity are tested on the data gathered in the previous experiment. Section 6.4 summarizes the implications of the current findings and literature for further experiments. Section 6.5 is a conclusion.

6.1 The setup for the new listening test

Listeners were asked to rate the clarity of stimuli featuring boosts and cuts in one of six octave bands, each against an unequalised reference, by adjusting a slider. The setup was nearly identical to the previous listening test but four new programme items were used, in order to test the influence of sounds’ original spectra on clarity. Following the results of a pilot study, warmth, fullness and brightness were rated occasionally to try to reduce the noise in the clarity ratings (chapter 5).

Each test page contained two stimuli (one was the unequalised reference stimulus) that could be triggered by pressing buttons labelled ‘A’ and ‘B’ as often as required after pressing ‘play’. A
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

‘stop’ button could be used to reset the stimuli to their beginning position. The stimuli were compared in terms of the perceptual attribute given at the top of each page. Clarity, fullness, warmth or brightness ratings were undertaken by positioning a slider on each of the fifty-seven pages (three pages each for the additional attributes). Fig. 6-1 shows an example of a test page. The interface differed slightly from that for the previous listening test. Additional, crossed out sliders were added for warmth, brightness and fullness in order to make it even more obvious that these should not be rated on the clarity pages.

![Diagram](image)

**Fig. 6-1: an example of the test page**

The entire listening test was repeated for each subject in a separate session in order to test listener consistencies. This resulted in two listening tests of about half an hour each. In order to counteract sequential bias, the order of the pages in each test was randomised. The assignment of reference and equalized stimulus to buttons A and B was also randomised. In this way, listeners did not know which of the stimuli the reference was.

Each listener was given an instruction sheet before undertaking the listening test (Appendix A-4). The test was preceded by a familiarization interface (Fig. 6-2) similar to the one in the pilot test and previous listening test, allowing each subject to audition all stimuli. Afterwards, listeners were given the opportunity to try out the test interface before starting the test. Like in the pilot test and previous listening test, no definition was given for clarity.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

In section, 6.1.1, stimulus generation is explained. In section 6.1.2, the presentation of the stimuli to test subjects is introduced. In section 6.1.3, the listening panel is presented.

Fig. 6-2: the familiarization interface.

6.1.1 Stimuli

In the following, the programme items and added equalisation are introduced.

Programme items

As outlined in chapter 5, the contribution of different octave bands to single sound clarity depended on the sound’s spectral shape in the last listening test. Where there was little to no energy in an octave band, boosts and cuts were less effective. Both programme items had a similar range of fundamental frequencies and a similar spectral slope. It is possible that sounds with much lower or higher fundamental frequencies, as well as strongly differing spectral bandwidths (i.e. spectral slopes and spectral centroids) may lead to different clarity ratings. Hence, the same listening test was repeated with other programme items, covering a wider range of spectra typically found in commercial music mixes.

The spectra of typical instruments vary in their spectral shapes, (i.e. spectral slopes, fundamental frequencies and spectral centroids) and their spectral detail (smoothness,
irregularity, the strength of even or odd harmonics and the presence of inharmonic partials such as distortion etc). However, following the previous literature review, there is no evidence that the impact of equalisation on clarity would change depending on the spectral detail: the number of harmonics [Ethington and Punch, 1994], strength of the lower even harmonics [Jeans, 1937], spacing of harmonics [Ethington and Punch, 1994; Plomp, 1976], spectral slope [Solomon, 1959] and spectral centroid [Disley and Howard, 2004] were previously found to influence clarity. The strength of the lower even harmonics, spectral slope and spectral centroid can be altered by boosting and cutting different octave bands. The spacing of harmonics can also be increased and the number of harmonics can be reduced by cutting certain frequency areas. There was no indication in the literature that the initial status of these attributes would influence the relative contribution of octave bands to single sound clarity. Instead, it appears useful to assess the spectral shapes (range of fundamental frequencies and spectral bandwidths) of typical instruments. As partials typically gradually decay towards the higher frequencies, the spectral centroid appeared to be a good measure of the spread of energy across all octave bands.

*Apple Loops* of fifteen instruments typically used in professional mixes, playing generic pop and classical motifs, were analysed. The approximate range of fundamental frequencies of the loops was analyzed and the average spectral centroids and range of possible fundamental frequencies of the instruments in the loops were investigated through a literature search [Blood, 2012; Everest, 2009; Sandell, 1991; Sutton et al., 2013], as shown in Table 6-1.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Rough fundamental range of programme item lies around</th>
<th>Average spectral centroid</th>
<th>Possible fundamental range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electric Bass</td>
<td>200–400Hz</td>
<td>588Hz</td>
<td>Around 30–440Hz</td>
</tr>
<tr>
<td>Brass ensemble</td>
<td>400–700Hz</td>
<td>3.034kHz</td>
<td>E.g. horns: around 70–900Hz</td>
</tr>
<tr>
<td>Synthesized Dance bass</td>
<td>200Hz or less</td>
<td>2.145kHz</td>
<td>Synthesized: entire audible range possible (20Hz–20kHz), but usually LF is used</td>
</tr>
<tr>
<td>Drum kit</td>
<td>Kick around 100Hz, snare around 200Hz, hi hats around 1k (highest peaks in the spectra were observed)</td>
<td>3.413kHz</td>
<td>Kick around 100Hz, snare around 100–200Hz, hi hat around 3kHz</td>
</tr>
<tr>
<td>Female pop vocal</td>
<td>300Hz–400Hz</td>
<td>3.231kHz</td>
<td>Alto: around 190–690Hz, Soprano: around 260Hz–1kHz</td>
</tr>
<tr>
<td>Piccolo flute</td>
<td>1–2kHz</td>
<td>1.871kHz</td>
<td>587Hz–3.729kHz</td>
</tr>
<tr>
<td>Male pop vocal</td>
<td>200–300Hz</td>
<td>2.413kHz</td>
<td>Baritone: around 140–600Hz</td>
</tr>
<tr>
<td>Organ</td>
<td>100–500Hz</td>
<td>1.430kHz</td>
<td>Around 16Hz – 8kHz</td>
</tr>
<tr>
<td>Synthesized pad sound</td>
<td>100–200Hz</td>
<td>1.55kHz</td>
<td>Synthesized: entire audible range possible</td>
</tr>
<tr>
<td>Previously used guitar</td>
<td>70–200Hz</td>
<td>3.604kHz</td>
<td>Around 70Hz–1.300kHz</td>
</tr>
<tr>
<td>Previously used piano</td>
<td>100–500Hz</td>
<td>985Hz</td>
<td>27Hz–4.186kHz</td>
</tr>
<tr>
<td>Rock guitar</td>
<td>Around 200Hz</td>
<td>1.685kHz</td>
<td>Similar to acoustic guitar</td>
</tr>
<tr>
<td>Saxophone</td>
<td>Around 200Hz</td>
<td>2.153kHz</td>
<td>Baritone: 70–400Hz, Alto: 130–830Hz</td>
</tr>
<tr>
<td>String quartet</td>
<td>100–500Hz</td>
<td>1.536kHz</td>
<td>70Hz–3kHz</td>
</tr>
<tr>
<td>Violin</td>
<td>Around 500Hz</td>
<td>3.527kHz</td>
<td>261Hz–3kHz</td>
</tr>
</tbody>
</table>

Table 6-1: the approximate range of the fundamental frequencies, average spectral centroids and the range of possible fundamental frequencies of 15 musical instruments.

As can be seen in the table, all instruments have a large range of possible fundamental frequencies, often spanning over three octaves. The average spectral centroids range from 600Hz to 2.5kHz (about two octaves). Sandell [1991] points out that the spectral centroid of an instrument depends on the pitch played. Each instrument has an individual range of spectra and
Assessing the contribution of different octave bands to single sound clarity, depending on programme items

playing styles or added devices like mutes can influence these further. Hence, the spectrum of a single instrument can be hard to quantify.

When investigating the long-term average spectra of the fifteen corresponding Apple loops, the harmonics mostly decayed evenly towards the higher frequencies. Where there were troughs in the slope, this was individually different for all instruments. Hence, none of the instruments seemed to share significant, unusual spectral details that were deemed likely to impact on the relationship between EQ and clarity. Some instruments such as piccolo flutes and electric basses have a lower spectral centroid relative to their fundamental on average than e.g. rich, synthesized dance basses. At the same time, although the range of possible fundamental frequencies is large for most instruments, some instruments like the piccolo flute emit much higher fundamental frequencies than basses.

Informal listening of the 15 instrument samples in Table 6-1 indicated that the cut-off point from which octave bands contributed to clarity positively ranged from about 1kHz to 5kHz, depending on the fundamental frequency of each sound. For monophonic instruments, it seemed easier to identify the cut-off point than for polyphonic instruments. In polyphonic instruments and mixtures of sounds (e.g. brass ensemble and string quartet), the cut-off point seemed to correlate with the lower instruments. For now, it appeared useful to assess the clarity of monophonic programme items suggested above in the same way as in previous listening tests.

It is possible that test subjects may find it easier to judge the clarity of familiar instruments, rather than newly synthesized timbres. Furthermore, in order to ensure that the impact of EQ in the different frequency bands on clarity really differs between the stimuli only due to differences between their spectral shapes, rather than any spectral detail, instruments of similar timbres, in this case string instruments, were used. Based on all these considerations, four 4-second programme items were selected: two cello and two violin stimuli with a bowed and a plucked version for each. The plucked programme items have lower spectral centroids than the bowed items and at the same time, the cello has a lower fundamental than the violin, hence covering several combinations of spectral centroids and fundamentals. The bowed samples stemmed from the Logic 9 Apple loops library and the plucked samples were created using the East West Composer’s Complete Gold Edition sample library. For environmental validity, the programme items were short musical phrases with typical temporal variation, rather than single note stimuli. The melodies in the programme items were mostly scales without large jumps in pitch: the pitch ranges were a 5th for the bowed cello, a minor 10th for the plucked cello, a major 6th for the bowed violin and a major 6th for the plucked violin.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

In order to assess the impact of the spectral shape on the relationship between equalisation and clarity, the position of the octave bands relative to the fundamental and spectral centroids of all four programme items was considered before running the experiment. Ideally, the octave bands should cover a wide range of harmonics, as well as distances from the spectral centroid. The range of octave bands that the harmonics contained in all programme items lie in were investigated. Table 6-2 shows the approximate range of harmonics affected by boosts and cuts in each octave band. Each harmonic up to the 200th harmonic is featured at least twice (for example, the first harmonic of the bowed and plucked violin programme items lie in the 500Hz band). Informal listening revealed that boosts and cuts in all octave bands were audible for all programme items, although they were less apparent in the 8kHz octave band for the cello programme items. In Table 6-3, the distance of the centre of each octave band from the spectral centroid of each programme item is given. Here, it can be seen that the chosen programme items cover a large number of distances, ranging from -3277Hz to 7527Hz.

<table>
<thead>
<tr>
<th>Programme items</th>
<th>250Hz</th>
<th>500Hz</th>
<th>1kHz</th>
<th>2kHz</th>
<th>4kHz</th>
<th>8kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cello bowed</td>
<td>1–3</td>
<td>4–5</td>
<td>5–15</td>
<td>10–30</td>
<td>20–60</td>
<td>40–130</td>
</tr>
<tr>
<td>Cello plucked</td>
<td>1–3</td>
<td>2–3</td>
<td>3–18</td>
<td>5–28</td>
<td>12–55</td>
<td>20–110</td>
</tr>
<tr>
<td>Violin bowed</td>
<td>0</td>
<td>1</td>
<td>1–3</td>
<td>2–5</td>
<td>4–15</td>
<td>8–20</td>
</tr>
<tr>
<td>Violin plucked</td>
<td>0</td>
<td>1</td>
<td>1–3</td>
<td>2–5</td>
<td>3–13</td>
<td>6–23</td>
</tr>
</tbody>
</table>

Table 6-2: the harmonics present towards the middle of each octave band (approximate for higher harmonics). In each cell, the harmonic numbers are given (e.g. “3” denotes “third harmonic”).

<table>
<thead>
<tr>
<th>Programme items with relative spectral centroids in brackets</th>
<th>250Hz</th>
<th>500Hz</th>
<th>1kHz</th>
<th>2kHz</th>
<th>4kHz</th>
<th>8kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cello bowed (1.232kHz)</td>
<td>-982</td>
<td>-732</td>
<td>-232</td>
<td>768</td>
<td>2768</td>
<td>6768</td>
</tr>
<tr>
<td>Cello plucked (473Hz)</td>
<td>-223</td>
<td>27</td>
<td>527</td>
<td>1527</td>
<td>3527</td>
<td>7527</td>
</tr>
<tr>
<td>Violin bowed (3.527kHz)</td>
<td>-3277</td>
<td>-3027</td>
<td>-2527</td>
<td>-1527</td>
<td>473</td>
<td>4473</td>
</tr>
<tr>
<td>Violin plucked (1.419kHz)</td>
<td>-1169</td>
<td>-919</td>
<td>-419</td>
<td>581</td>
<td>2581</td>
<td>6581</td>
</tr>
</tbody>
</table>

Table 6-3: the distance of the center of each octave band from the spectral centroid of each programme item in Hz.

Below are the long time average spectra of all programme items, again showing the variation in fundamentals and spectral centroids in the programme items.
Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Added equalisation

For each equalized stimulus, a single 9dB boost or cut was added to one of the six octave bands centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz respectively, with a Q value of 1.41 (-3dB bandwidth = 1 octave). This was done using Logic Pro 9’s parametric equaliser, like in the
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

previous listening test (chapter 5). The resulting stimuli were exported as 44.1kHz 24 bit wave files. As mentioned above, the boosts and cuts were audible for all programme items, although they were less apparent in the 8 kHz octave band for the cello.

The stimuli were loudness-matched by four audio professionals, like in the pilot study and previous listening test.

6.1.2 The presentation of the stimuli to test subjects

In the same way as in the pilot test and previous listening test, the test interface and incorporated stimulus pairs were presented at a comfortable listening level in an international-standard listening room built to ITU-R Standard BS1116, using Bowers & Wilkins Nautilus 801 speakers. The lack of extraneous noises ensured that important nuances in the spectra of the stimuli were audible.

6.1.3 Type of listeners

It is recommended that at least ten to fifteen suitably trained listeners are employed for listening tests [Bech and Zacharov, 2006], hence 17 male and female students (undergraduates and postgraduate researchers) of the University of Surrey's Institute of Sound Recording took part. The participants were aged between nineteen and twenty-seven years old. All subjects were experienced in listening tests, as well as in verbalising sensations of timbre. As part of their degree course, the undergraduate students were receiving extensive technical ear training, which included the blind identification of EQ changes. None of the participants reported having any hearing damage. Only one listener undertook the test at a time.

6.2 Presentation of the results

In the following, the results of the listening test will be presented. Section 6.2.1 shows that the prerequisites for using parametric methods of statistics are not quite fulfilled. In section 6.2.2, the contribution of boosts and cuts across the six octave bands is shown in boxplots. In section 6.2.3, the relationship between the spectral centroid and clarity ratings is assessed. In section 6.2.4, the boosted and cut harmonic numbers and their influence on clarity are tested. In section 6.2.5, the harmonic centroid is introduced.

6.2.1 Prerequisites for using parametric methods

To decide whether to use parametric on non-parametric methods, the following conditions had to be tested:
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

1. Independence
2. Interval data
3. Normally distributed data
4. Homogeneity of variance

Independence was most likely not fulfilled for the same reasons as in previous listening tests. The data is interval data, as all points along the slider were equally spaced, rather than making a qualitative observation about the data elicited from listeners. Hence, this condition for using parametric tests was fulfilled.

In order to assess whether the data were normally distributed, histograms were analysed for all stimulus pairs. For datasets smaller than 2000 elements, it is suggested that the Shapiro-Wilk test is used as normality test, otherwise, the Kolmogorov-Smirnov test should be used [Field, 2013]. The current dataset comprises 1632 clarity values, hence the Shapiro-Wilk test was used. It was concluded that the data were mostly normally distributed, but with some exceptions: in the Shapiro-Wilk test, four of the 48 significance values were below 0.05. Similarly, some of the plots indicated a roughly normal distribution of the data, while others did not (Fig. 6-5).
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

![Histogram for StimPair 7.00](image1)

![Histogram for StimPair 25.00](image2)

Fig. 6-5: Two examples of the histograms used to assess normality of the data. The clarity ratings are approximately normally distributed for stimulus pair 7 but slightly less for stimulus pair 25.

As a next step, it was assessed whether the elimination of potentially inconsistent listeners would lead to a more normal distribution. As outlined in section 6.1, each listener undertook each rating twice. The notched box plot in Fig. 6-6 shows the distribution of differences between the ratings for all stimulus pairs and for each listener. It appears that the ratings given by listener three are less consistent than all other listeners' ratings, although not significantly. There is no clear cut-off point between consistent and inconsistent listeners, as consistencies
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

are quite varied overall. Hence, it would be arbitrary to remove certain listeners. Without listener three, the data were still not quite normally distributed.

![Chart showing differences between clarity ratings](image)

Fig. 6-6: distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.

All in all, the data were not quite normally distributed with some exceptions. Levene’s test shows that the variance of the data is not homogenous, with a significance of less than 0.000. As the prerequisites for using parametric statistical methods are not fulfilled, mostly non-parametric methods were used with the current data set, such as interpreting plots. Both Pearson (parametric) and Spearman (non-parametric) correlations were considered. The Pearson coefficient assesses the degree to which the relationship between two data sets (in this case spectral centroid deviations and clarity ratings) is linear, whereas the Spearman coefficient assesses the degree to which the relationship is monotonic. The current experiment assesses relative changes of clarity. The aim is to establish how reliably raising e.g. the spectral centroid increases clarity, rather than by how much. Hence, the degree to which the different spectral centroids have a monotonic relationship with clarity is more important than the linearity of that relationship. Hence, the Spearman coefficient appears to be a better measure and was used to compare the clarity predictors introduced in the following. As the data were not quite normally distributed, the correlation between the median clarity ratings and spectral centroids will be assessed (rather than the means).

Conetta et al. [2015] present objective models of sound quality (Table 6-4). For each model, the strength of correlation and RMSE, measured against listening test data, is presented. The smallest correlation, belonging to the widely accepted PEAQ [ITU-R BS.1387, 1998] model is
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

0.67. Hence, a correlation of that or more is likely to indicate that a predictor is useful in the current research project.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Purpose</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEAQ [ITU-R BS.1387]</td>
<td>Adopted standard algorithm for the objective assessment of perceived audio quality</td>
<td>The six models that form part of PEAQ have correlations between r=0.67 and r=0.86</td>
</tr>
<tr>
<td>BAQ [Zieliński et al., 2004, 2005]</td>
<td>Parametric model for predicting the Basic Audio Quality of a multichannel audio system</td>
<td>The &quot;quality advisor&quot;, developed for validation, measures a correlation of r=0.93 and an RMSE of 9%</td>
</tr>
<tr>
<td>Choi et al. [2008]</td>
<td>Multichannel addition to the PEAQ standard with additional spatial metrics</td>
<td>The two models belonging to the model addition achieve r=0.85 (RMSE of 5.09%) and r=0.79 (RMSE of 5.44%)</td>
</tr>
<tr>
<td>Seo et al. [2010, 2013]</td>
<td>Improvement of PEAQ model with a neural network model</td>
<td>r=0.88 and RMSE of 5.18%</td>
</tr>
<tr>
<td>George [2009]</td>
<td>Objective evaluation models for predicting process-induced impairment to the frontal spatial fidelity, surround spatial fidelity, and timbral fidelity of 5-channel audio recordings</td>
<td>Frontal spatial fidelity: r=0.91, RMSE of 9.33 (calibration); r=0.88, RMSE of 15.45 (validation). Surround spatial fidelity: r=0.95, RMSE of 8.87 (calibration); r=0.87, RMSE of 14.19 (validation). Timbral fidelity: r=0.95, RMSE of 7.72 (calibration); r=0.92, RMSE of 8.37 (validation)</td>
</tr>
</tbody>
</table>

Table 6-4: correlations and RMSE's for measured listening test data and predicted data

6.2.2 The impact of octave bands on clarity

In order to assess the impact of boosts and cuts across the octave bands on clarity, the distribution of all clarity ratings is presented in two box plots, for all boosts (Fig. 6-7) and for all cuts (Fig. 6-8). A clear trend is evident, as many notches do not overlap and for each programme item, the medians differ considerably between several stimulus pairs.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

**Fig. 6-7:** Distribution of clarity ratings for all boosts in comparison to the unequalised reference stimuli.

**Fig. 6-8:** Distribution of clarity ratings for all cuts in comparison to the unequalised reference stimuli.

It can be seen that higher frequencies contribute to clarity positively (or at least less negatively than low ones) and low ones contribute negatively. Overall, HF boosts seem to increase clarity and LF boosts reduce it, and conversely for cuts. On average, it seems that boosts around 500Hz make the strongest negative contribution, while cuts in this area make the strongest positive contribution. However, the zero-crossing point and slope of a curve imagined through the medians seems to vary with programme item which may be due to the fundamental frequencies and initial spectral slopes of the programme items, as explained in the following.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

When comparing the impact of boosts on the two bowed programme items, both seem to be arranged in a curve resembling a parabola facing upward, but the curve for the violin lies further to the right and spans a smaller range of clarity ratings. The octave bands that make the most negative contribution to the bowed cello programme item are 250Hz, 500Hz and 1kHz, but for the bowed violin, 2kHz and 4kHz also contribute negatively. Even the 8kHz band does not contribute significantly positively to the violin programme item, whereas in the cello both 4kHz and 8kHz contribute positively. When comparing cuts, the situation is similar, but the parabola faces downward. On average, boosts and cuts seem to have a stronger impact on the cello, where particularly the absence of low frequency energy contributes to clarity. The bowed violin never becomes clearer in the presence of boosts and just slightly clearer in the presence of cuts.

The situation is similar when comparing the plucked cello and violin programme items, although cuts in the plucked cello make no significant contribution to clarity and in the violin, boosts in higher octave bands now contribute positively to clarity. It is concluded that the fundamental of programme items plays an important role, as the violin has a much higher average fundamental than the cello.

When comparing programme items with similar average fundamentals but differing centroids, differences can also be observed. The boosts and cuts contribute similarly for bowed and plucked violin programme items, but cuts in low frequencies contribute positively for the plucked items which is not the case for bowed items. For higher frequencies, both are more similar. For the boosts, the curves have similar shapes but for the plucked violin programme item, the curve seems to be more compressed horizontally. This is also true for cello programme items: for bowed programme items, boosts and cuts seem to have a bigger impact than for plucked items and in the bowed programme item, cuts can have a significant positive impact to clarity (500Hz). Cuts in the highest two octave bands in the plucked cello programme item do not seem to make a big difference, presumably because the programme item has little energy in this area.

In conclusion, higher octave bands still contribute more positively to clarity than lower octave bands, but the spectral centroids and fundamentals of the programme items seem to play an important role. At the same time, boosts and cuts cannot make a significant difference where programme items do not have much energy in these areas. In the following, spectral centroids and the fundamentals are investigated in more detail.

6.2.3 Spectral centroids

As mentioned in the previous section, it appears that the original spectral centroids of programme items influence the contribution of octave bands to clarity. Whether each boost or
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

cut raises or lowers the spectral centroids differs between programme items. Therefore, it appears useful to investigate the contribution of differences in spectral centroid, induced through EQ, on clarity. For each programme item, the spectral centroids of the unprocessed version and all twelve equalised versions were calculated.

To do this, the stimuli were split into a number of Hamming windowed segments, such, that each segment length did not exceed an 8th of the overall stimulus length or 2084 samples, whichever the greater. The segments overlapped by 50%. For each segment, a spectrogram was calculated. By doing so, it could be ensured that intra-stimulus centroid variation was not too large. The overall spectral centroid was established by calculating the mean spectral centroid of all segment centroids.

The centroids for the unequalised reference stimuli were deducted from the equalised centroids. These differences were plotted against the clarity data in a line plot (Fig. 6-9). The lines in the plots are the non-parametric confidence intervals, with the medians shown in the centre. The plot clearly shows that raising the spectral centroid by more than 200Hz increases clarity. When raising the centroid by more than about 400Hz, clarity drops again. Lowering the spectral centroid always makes a significant negative contribution to clarity. The Spearman correlation coefficients indicate a significant positive correlation between the difference in spectral centroid and clarity ratings. Here, the Spearman correlation coefficient is 0.806 with a p-value < 0.001. The fact that clarity drops again when the centroid is raised by more than about 400Hz may be due to the fact that the fundamental also needs to be considered: when the average fundamental of a sound is already high (the stimuli on the far right of the plot are violin stimuli), the spectral centroid may not have moved up a large enough number of overtones, even if the difference in spectral centroids is high. To investigate this further, the harmonic centroid is introduced in section 6.2.5. Before moving on, however, the spectral centroid was calculated in three further ways, taking the auditory perception of pitch into consideration.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

The traditional spectral centroid that Fig. 6-9 is based on is calculated using linearly spaced frequencies. However, frequencies are not spaced linearly across the basilar membrane [Moore, 2012] and musical notes are not spaced linearly in frequency. Hence, it was decided to calculate the spectral centroid musically and perceptually, using Cent, Mel and Equivalent Rectangular Bandwidth (ERB) scales.

The Mel Scale is a perceptual scale of pitch. Stevens et al. [1937] asked test subjects to adjust the frequency of a comparison tone until its pitch appeared to be twice or half or that of a given reference. Here, the pitch value of a 1000Hz tone is defined as 1000 Mels. A tone twice as high has a pitch of 2000 Mels. Frequencies in Hz ($f$) are converted to Mels ($m$) as follows:

$$ m = 1127 \log_e \left( 1 + \frac{f}{700} \right) $$

(Modeling auditory filters, Fletcher [1940] proposed specifying the bandwidth of a rectangular filter passing equivalent energy, for simplification. He thus established the concept of an “equivalent rectangular bandwidth” (more detail in chapter 3). Moore [2012] introduces a frequency scale related to $ERB_N$ (mean value for moderate sound levels and young people with normal hearing), where the value of the $ERB_N$ is used as the unit of frequency. Here, $f$ is the frequency in Hz.

$$ ERB_N(f) = 21.4 \log_{10}(1 + 0.00437f) $$
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

The Cent scale is not an absolute measure of frequency, but a logarithmic ratio of two frequencies [Walker and Don, 2013]. There are 100 cents in a semitone, and 12 semitones in an octave. An octave interval between two frequencies is a doubling of frequency. As this is a relative, rather than an absolute measure, the distance of each frequency value from the arbitrary value 27.5Hz (A0, bottom 'A' on a piano) was calculated.

\[ c = 1200 \log_2 \left( \frac{f}{27.5} \right) \]  \hspace{1cm} (6.3)

The spectral centroids of all stimuli were calculated in ERBs, Mels and Cents. The resulting values were used to plot the data in the same way as before three times (Fig. 6-10, Fig. 6-11 and Fig. 6-12) but using these three new types of centroids. The ERB, Mel and Cent centroids were also reverted back to Hz, resulting in three additional plots (Fig. 6-13, Fig. 6-14 and Fig. 6-15).

The calculation of the ERB, Mel and Cent centroids was similar to that of the spectral centroid calculation detailed above. As before, spectrograms were calculated for stimulus segments. Through a Fourier transform, the original audio signals were converted to a magnitude \((x(n))\) and a frequency vector \((f(n))\). The frequency vector contained the linearly spaced frequency bin values (in Hz) and the corresponding magnitude values were in \((x(n)).\) The original vector of linearly spaced frequency bins \((f(n))\) was replaced by three new vectors. In this way, three new vectors were calculated: \(f_{\text{Mel}}(n), f_{\text{Cent}}(n)\) and \(f_{\text{ERB}}(n)\):

\[ f_{\text{Mel}}(n) = 1127 \log_e \left( 1 + \frac{f(n)}{700} \right) \]  \hspace{1cm} (6.4)

\[ f_{\text{Cent}}(n) = 1200 \log_2 \left( \frac{f(n)}{27.5} \right) \]  \hspace{1cm} (6.5)

\[ f_{\text{ERB}}(n) = 21.4 \log_{10} (1 + 0.00437f(n)) \]  \hspace{1cm} (6.6)

The new amplitude and frequency vectors were inserted into the spectral centroid formula to calculate the spectral centroids using Mels, Cents, and ERBs, as shown below.

\[ C_{\text{Mel}} = \frac{\sum_{n=0}^{N-1} f_{\text{Mel}}(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \]  \hspace{1cm} (6.7)

\[ C_{\text{Cent}} = \frac{\sum_{n=0}^{N-1} f_{\text{Cent}}(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \]  \hspace{1cm} (6.8)

\[ C_{\text{ERB}} = \frac{\sum_{n=0}^{N-1} f_{\text{ERB}}(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \]  \hspace{1cm} (6.9)
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

For each stimulus pair, the differences between the original and processed spectral centroid calculated in Mels, Cents and ERBs were calculated and graphs were plotted in the same way as before (Fig. 6-13, Fig. 6-14 and Fig. 6-15). Next, all Mel, Cent and ERB spectral centroid values were reverted to Hz values, using the inverse of the equations presented above. Differences between each stimulus and the reference were calculated again and graphs were plotted like before (Fig. 6-10, Fig. 6-11 and Fig. 6-12).

The observed relationship between the differences in centroids (e.g. \(C_{\text{Mel original}} - C_{\text{Mel equalized}}\)) and clarity ratings is similar to before and each has a significant positive correlation. The coefficients are given in Table 6-5. As can be seen in the plots, the cut-off points where the contribution of centroid changes to clarity becomes positive and then negative again differ slightly between plots.

<table>
<thead>
<tr>
<th>Correlation between ratings and...</th>
<th>Spearman</th>
<th>Spearman p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in spectral centroid (linear)</td>
<td>0.806</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Mel)</td>
<td>0.745</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (ERB)</td>
<td>0.720</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Mel-&gt; Hz)</td>
<td>0.701</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Cent)</td>
<td>0.673</td>
<td>0.000</td>
</tr>
<tr>
<td>Harmonic number boosted</td>
<td>0.657</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (ERB-&gt; Hz)</td>
<td>0.651</td>
<td>0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Cent-&gt; Hz)</td>
<td>0.638</td>
<td>0.000</td>
</tr>
<tr>
<td>Harmonic number cut</td>
<td>-0.621</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

Table 6-5: all correlation coefficients, sorted descending by the Spearman coefficient. All correlate significantly positively, except “harmonic number cut” (significant negative correlation). All p-values are lower than 0.05

Compared to linear frequency spacing, the new scales are stretched at their LF ends and compressed at their HF ends. In this way, the scaled spectral centroid calculations are affected more by changes at low frequencies than by changes at high frequencies. This reflects the auditory system’s nonlinear perception of pitch as outlined above. The Mel scale is the closest to linear spacing and the cent scale is the most different. As a result of the scaling, the rank order of centroid shift magnitudes between stimulus pairs changed. Changes in Mel and linear spectral centroids appear to be the most useful predictors of clarity change. However, the linear spectral centroid still has the highest Spearman coefficient (0.806). Other means of improvement will therefore be sought. In the following sections, the influence of the average fundamental of each programme item is taken into consideration.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Fig. 6-10: differences in spectral centroids, calculated in Cents and plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-11: spectral centroids, calculated in Mels, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

![Graph showing spectral centroids, calculated in ERBs, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.](image1.png)

**Fig. 6-12:** spectral centroids, calculated in ERBs, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

![Graph showing differences in spectral centroids, calculated in Cents, and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.](image2.png)

**Fig. 6-13:** differences in spectral centroids, calculated in Cents, and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Fig. 6-14: differences in spectral centroids on clarity, calculated in Mels and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

Fig. 6-15: differences in spectral centroids on clarity, calculated in ERBs and converted back into Hz, plotted against clarity. Each line represents the non-parametric confidence interval of the ratings of the corresponding stimulus pair and each circle is the median.

6.2.4 Harmonics Boosted and cut

As established in the last section, the average fundamental of each programme item seems to play an important role for the impact of equalisation on clarity. Each octave band contains some of the harmonics of each programme item, but the harmonic numbers differ for each programme item. In order to establish which harmonics were boosted or cut in each stimulus
Assessing the contribution of different octave bands to single sound clarity, depending on programme items

pair, the octave band centre frequencies were divided by the median fundamental of each programme item. This was defined as the frequency exactly in the middle between the highest and lowest note played by each instrument. The melodies in the programme items were mostly scales and did not contain large jumps in pitch: the pitch range was a 5th for the bowed cello, an octave and a minor 3rd for the plucked cello, a major 6th for the bowed violin and a major 6th for the plucked violin.

The impact of boosts and cuts of different harmonics on clarity is shown in Fig. 6-16 and Fig. 6-17. It appears that boosting higher harmonics or cutting lower ones increases clarity, while cutting higher harmonics or boosting lower ones reduces clarity. The graphs indicate a possible linear relationship, although it is difficult to tell where exactly the cut-off point is from which the contribution to clarity becomes positive. Fitting lines to the plots, it appears that cutting frequencies below the second harmonic or boosting above the 17th harmonic increases clarity. In contrast to raising the spectral centroid, boosting increasingly high harmonics seems to increasingly raise clarity, rather than beginning to increase it to a lesser extent, as was the case when raising the spectral centroid by more than around 400Hz. It is likely that this was because the stimuli where the spectral centroid was raised by more than 400Hz also had high fundamentals (violin programme items) and therefore the centroid change was smaller in respect to the item's overall position in the frequency spectrum. The harmonic number boosted and clarity ratings have a significant positive correlation: the Spearman correlation coefficient is 0.657 with a p-value < 0.001. Similarly, the harmonic number cut and clarity ratings have a significant negative correlation: the Spearman correlation coefficient is -0.705 with a p-value < 0.001.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Fig. 6-16: harmonic number boosted, plotted against clarity ratings. Harmonics are shown on a logarithmic scale for an even spread of x-axis values. This reflects the fact that pitches are arranged logarithmically on the basilar membrane.

Fig. 6-17: harmonic number cut, plotted against clarity ratings. Harmonics are shown on a logarithmic scale for an even spread of x-axis values. This reflects the fact that pitches are arranged logarithmically on the basilar membrane.

6.2.5 Harmonic centroid

In the previous sections, it was suggested that both the spectral centroid and the average fundamental of the programme items influence the way in which boosts and cuts in different octave bands impact on clarity. In this section, the harmonic centroid is defined as the harmonic
Assessing the contribution of different octave bands to single sound clarity, depending on programme items

number where the spectral centroid is situated. It is calculated by dividing the spectral centroid by the mean fundamental. Unlike the spectral centroid of a sound, which is largely dependent on the sound’s fundamental frequency, the harmonic centroid is largely independent of fundamental frequency and is, instead, primarily affected by spectral shape. The harmonic centroid is the same as the ‘unitless spectral centroid’ introduced by Kendall and Carterette [1996] as a useful predictor of brightness. Schubert and Wolfe [2006] state that the spectral centroid appears to be a more useful measure of brightness than the unitless spectral centroid. However, Marozeau et al. [2003] found that differences in timbre depended little on pitch when the pitch difference was either 2 semitones or 11 semitones. Therefore, it is possible that spectral shape may be more important than spectral position. Since brightness appears to be perceptually similar to clarity [Disley et al., 2006], the harmonic centroid is worth investigating in this context.

Differences in harmonic centroid, induced by equalisation, are plotted against clarity in Fig. 6-18, Fig. 6-19 and Fig. 6-20. Here, we can see that raising the harmonic centroid increases clarity. This is the case also when boosts or cuts were plotted individually. The relationship is significantly positive with a Spearman correlation coefficient of 0.818 (p < 0.001). These values are similar when regarding boosts and cuts in isolation, as shown below in Fig. 6-19 and Fig. 6-20.

The extent to which clarity increases eventually reaches a saturation point: clarity cannot be increased by more than about 20% of the scale, and this point is reached after raising the harmonic centroid by just a few harmonics. This may be due to only applying 9dB boosts and cuts, or it could be a feature of the particular programme items chosen, or a general rule but the current data set cannot explain this phenomenon.

The plots also show which octave band was boosted or cut in each case. This confirms once again that the contribution of the octave bands depends on the input signal and is not a consistent value. The correlation coefficients for changes in all centroids introduced in this report, as well as the harmonics boosted or cut are summarized in table 6-6. As explained in section 6.2.3, the Spearman coefficient was chosen as a measure to establish the best predictor of clarity and this is the harmonic centroid. However, the difference between the Spearman coefficients of e.g. harmonic centroid and spectral centroid is small and the rank order in table 6-6 may be specific to the chosen programme items. In order to test whether the harmonic centroid can predict clarity well for other programme items, the most useful clarity predictors found so far are compared in their suitability for predicting guitar and piano stimulus clarity in section 6.3.
Assessing the contribution of different octave bands to single sound clarity, depending on programme items

<table>
<thead>
<tr>
<th>Correlation between ratings and...</th>
<th>Spearman</th>
<th>Spearman p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change in harmonic Centroid</td>
<td>0.818</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in harmonic Centroid, boosts only</td>
<td>0.813</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (linear)</td>
<td>0.806</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in harmonic Centroid, cuts only</td>
<td>0.749</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Mel)</td>
<td>0.746</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (ERB)</td>
<td>0.729</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Mel&gt; Hz)</td>
<td>0.701</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Cent)</td>
<td>0.673</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Harmonic number boosted</td>
<td>0.657</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (ERB&gt; Hz)</td>
<td>0.651</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Change in spectral centroid (Cent&gt; Hz)</td>
<td>0.638</td>
<td>&lt; 0.000</td>
</tr>
<tr>
<td>Harmonic number cut</td>
<td>-0.621</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

Table 6-6: all correlation coefficients, sorted descending by the Spearman coefficient. All correlate significantly positively, except "harmonic number cut" (significant negative correlation). All p-values are lower than 0.05.

Fig. 6-18: changes in harmonic centroid, plotted against clarity ratings.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

![Graph showing changes in harmonic centroid due to boosts and cuts in different octave bands.](image)

**Fig. 6-19:** changes in harmonic centroid, plotted against clarity ratings (boosts only).

**Fig. 6-20:** changes in harmonic centroid, plotted against clarity ratings (cuts only).

6.3 Clarity predictors for guitar and piano stimuli

High frequency boosts and low frequency cuts increased the clarity of the piano and guitar stimuli in the last experiment (chapter 5). In the current experiment, the harmonic centroid was found to correlate positively with the clarity change in plucked and bowed violin and cello stimuli. As the harmonic centroid proved to be a good predictor of single sound clarity for these stimuli, it would likely form a useful part in an overall clarity model. In the current section, the most useful clarity predictors found so far are compared in their suitability for predicting guitar
Assessing the contribution of different octave bands to single sound clarity, depending on programme items and piano stimulus clarity. The following line plots (Fig. 6-21—Fig. 6-27) show the relationship between clarity and the spectral centroid, the Mel, ERB, Cent centroids, the harmonic number boosted and cut and lastly, the harmonic centroid. The piano and guitar stimuli were polyphonic and spanned a very large range of pitches. Therefore, the fundamental of the lowest note in each stimulus was used for the harmonic centroid.

---

**Fig. 6-21:** differences in spectral centroids, plotted against clarity ratings.

---

**Fig. 6-22:** differences in Mel centroids, plotted against clarity ratings.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

**Fig. 6-23** differences in ERB centroids, plotted against clarity ratings.

**Fig. 6-24**: differences in Cent centroids, plotted against clarity ratings.
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Fig. 6-25: boosts on harmonic numbers, plotted against clarity ratings

Fig. 6-26: cuts on harmonic numbers, plotted against clarity ratings
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

The plots show that the further centroids are raised, or the higher the level of higher harmonics is relative to lower harmonics, the clearer stimuli become. Table 6-7 shows all Spearman correlation coefficients for all predictors. All are significant. The harmonic centroid appears to be the best predictor, after the harmonic number boosted. As the harmonic number cut does not correlate well with clarity, the harmonic centroid appears to be the most suitable predictor. This confirms the assumption that an overall measure of clarity should most likely include the harmonic centroid.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Spearman R</th>
<th>Spearman P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmonic number boosted</td>
<td>0.837</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>HC</td>
<td>0.775</td>
<td>0.001</td>
</tr>
<tr>
<td>Linear</td>
<td>0.750</td>
<td>0.001</td>
</tr>
<tr>
<td>Harmonic number cut</td>
<td>-0.734</td>
<td>0.009</td>
</tr>
<tr>
<td>Mel</td>
<td>0.721</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>ERB</td>
<td>0.642</td>
<td>&gt; 0.001</td>
</tr>
<tr>
<td>Cent</td>
<td>0.601</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table 6-7: Spearman correlation coefficients for all clarity predictors for piano and guitar stimuli

6.4 Relation to the literature and implications for further listening tests

The current experiment shows that raising the harmonic centroid of recorded programme items increase spectral clarity. This is in line with the majority of studies in the context of timbral clarity [Disley and Howard, 2004 and Solomon, 1959]. A positive correlation between the
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Spectral centroid and clarity was found. In the context of string instruments, Fritz et al. [2012] found violin clarity to correlate with an increase in level between 1520–6080 Hz but this is not in accordance with an earlier study by Dünnwald [1991], where clarity was associated with a lower level between 4200 and 6400 Hz. This indicates that further factors may need to be considered in addition to the harmonic centroid. For example, it is possible that the violins tested by Dünnwald had unpleasant resonances in the clarity reducing frequency area while this was not the case in Fritz et al.’s study. In order to investigate this further, it would be useful to test the findings summarized in this chapter on further programme items.

6.5 Conclusion

Having established the importance of spectra in the context of clarity and separation, it was shown that the spectra of the programme items appeared to have an influence on the contribution of each octave band to the clarity of single sounds. Hence, the present listening test aimed to answer the research question: how can changes in the spectral clarity of equalized programme items with differing fundamentals and original spectral centroids be predicted?

Seventeen test subjects compared forty-eight stimuli featuring one boost or cut each in one of six octave bands to unequalised reference stimuli for four programme items (plucked and bowed violin, and plucked and bowed cello programme items), using a paired comparison test. The octave bands were centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz. The four programme items were chosen as they differ in their fundamental frequencies and spectral centroids, while still sharing similar timbres.

Overall, a significant positive correlation exists between the difference in processed and original centroids and clarity. This was also true when calculating spectral centroids perceptually, using Mel, ERB and Cent scaled frequencies. The Mel and linear centroids had the highest correlation with clarity. Raising the linear spectral centroid by more than 200Hz raised clarity. The same is true for raising the centroid by more than about 50 Mels. Within the current dataset, clarity appeared to fall as the centroid increased by more than 400Hz, 150 Mels or 2 ERBs. However, the data in this area came from programme items with high fundamental frequencies and so the effect of the fundamental was evaluated by analysing the data in terms of harmonic numbers rather than frequencies.

Boosting higher harmonics or cutting lower ones increased clarity. The point at which frequency areas started to contribute to clarity positively was between the second and the 17th harmonic. This seemingly linear relationship appeared to stay constant even for higher
6 Assessing the contribution of different octave bands to single sound clarity, depending on programme items

harmonics. Again, a significant positive correlation could be observed. When cutting higher harmonics or boosting lower ones, a significant negative correlation could be observed.

Lastly, Raising the harmonic centroid (the spectral centroid divided by the average fundamental) increased clarity, although a saturation point was reached after about 1.5 harmonics. Following the Spearman correlation coefficient, a measure used to determine to what degree the relationship between the predictors and clarity is monotonic, the harmonic centroid is the most useful predictor for the impact of equalisation on single sound clarity. The found clarity predictors were also tested on the guitar and piano stimuli found in the previous experiment and here, the harmonic centroid was also the best predictor. In order to assess whether the harmonic centroid alone can successfully predict clarity, it is suggested that the same EQ treatments are applied to further programme items and tested for clarity.
7 The single sound spectral clarity of vocal and bass stimuli

In previous experiments, the harmonic centroid (spectral centroid divided by a sound’s average fundamental) was established to be a useful predictor of single sound clarity, followed by Mel, linear, ERB and Cent centroids, as well as the harmonic number boosted or cut. The predictors were tested on monophonic string stimuli and polyphonic guitar and piano stimuli. It is possible that the string stimuli were perceived similarly as they belong to the same instrument group. The piano and guitar stimuli were polyphonic which made it slightly more difficult to determine the average fundamental for harmonic centroid calculation. In order to test how reliably the above metrics can predict single sound spectral clarity, it would be useful to test them on further, harmonic, monophonic programme items. Programme items where clarity might work differently would be useful here, as this might show any limitations of the previously tested predictors. The aim of the current listening test, therefore, is to test the predictor of single sound spectral clarity on a new set of stimuli.

Many commercial music mixes contain vocals and at the same time, it is likely that the perception of voices differs from that of other sounds. Speech intelligibility may also influence the perceived clarity of singing. Therefore, it is suggested that vocal stimuli are tested for clarity. Another harmonic, monophonic instrument frequently found in mixes is the bass. It is possible that low frequencies play a more important role in bass clarity than the clarity of other sounds, as basses often fulfil the role of adding low end to mixes. Therefore, in addition to the vocal stimuli, bass stimuli are tested. For consistency, the same experiment design should be used. The following research question is answered:

Does raising the harmonic, linear, Mel, ERB and Cent centroids, boosting higher harmonics or cutting lower ones increase spectral clarity for vocal and bass stimuli, when the same EQ settings are used as in the last experiment?

In section 7.1, the setup for the new experiment is outlined and in section 7.2, the results are summarized. Section 7.3 is a discussion and section 7.4 concludes.

7.1 The setup for the listening test

The following paragraphs provide an overview of the listening test setup. Sixteen male and female listeners were asked to rate the clarity of vocal and bass stimuli featuring boosts and cuts in one of eight octave bands each against an unequalised reference by adjusting a slider.
The single sound spectral clarity of vocal and bass stimuli

Warmth, fullness and brightness were rated occasionally to try to reduce the noise in the clarity ratings, like in previous listening tests. Fig. 7-1 shows an example of a test page. The interface was identical to the previous listening test.

Each stimulus featuring a boost or cut was compared to the unequalised reference stimulus in terms of clarity. Three pages were added to assess the additional perceptual attributes (one page each), resulting in thirty-three pages. The entire listening test was repeated for each subject in a second test half in order to test listener consistencies. This resulted in an overall session of about half an hour. The stimulus pairs were presented at a comfortable listening level in a treated, quiet listening room, using a VRM box interface and Sennheiser HD 600 headphones. The virtual reference monitoring function on the VRM box was switched off.

The stimuli in the first three listening tests were presented at a comfortable listening level in an international-standard listening room built to ITU-R Standard BS1116, using Bowers & Wilkins Nautilus 801 speakers, which delivered useable results also. The switch from loudspeaker presentation in the previous listening tests to headphone presentation in the current test was due to availability of facilities. It is possible that headphones might deliver a clearer reproduction than loudspeakers (e.g. due to the absence of room reflections). It is expected, however, that since the two reproduction systems have similar frequency responses, perception of the change to clarity resulting from spectral filtering will be similar.
Each listener was given an instruction sheet before undertaking the listening test (Appendix A-5). The test was preceded by a familiarization interface (Fig. 7-2), allowing each subject to audition all stimuli. Afterwards, they were given the opportunity to try out the test interface before starting the test. Like in all previous listening tests, no definition was given for clarity. In the next section, stimulus generation is introduced.

![Image of stimuli](image)

**Fig. 7-2**: the familiarization interface.

### 7.1.1 Stimulus generation

A 4 second plucked electric bass stimulus was created using the East West Composer’s Complete Gold Edition sample library and a 4 second vocal stimulus was recorded using an SE2200A microphone and Focusrite Saffire interface. Herein, a male singer sang a short melody on the syllable “la”. In this way, the listeners could not be biased in their ratings by the lyrics, and at the same time, the formant spectrum was consistent across all notes. If there had been changes in the vocal formant spectrum throughout, it is possible that clarity would have been perceived differently for the different notes, making it difficult to test the overall predictors. For
environmental validity, the programme items were short musical phrases with typical temporal variation, rather than single note stimuli. The melodies in the programme items were mostly scales without large jumps in pitch. The pitch ranges were a 4th for the vocal and a 5th for the bass.

For each equalized stimulus, a single 9dB boost or cut was added to one of the eight octave bands centred at 62Hz (bass only), 125Hz, 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz respectively, with a Q value of 1.41 (-3dB bandwidth = 1 octave). This was done using Logic Pro 9’s parametric equaliser), like in the previous listening test (chapter 5). The resulting stimuli were exported as 44.1kHz 24 bit wave files and loudness-matched by four audio professionals.

7.2 Presentation of the results

In the following, the results of the listening test will be presented. Section 7.2.1 shows that the prerequisites for using parametric methods of statistics are not quite fulfilled. In section 7.2.2, the influence of spectral centroid deviations on spectral clarity is assessed. In section 7.2.3, the influence of harmonics boosted and cut is assessed.

7.2.1 Prerequisites for using parametric methods

Before deciding whether parametric statistical methods could be used, the conditions specified in previous experiment reports were tested (interval data, independence, normal distribution and homoscedasticity). The data gathered in the experiment were interval data. Like before, the fact that listeners most likely compared ratings to each other means that independence could not be guaranteed.

In order to assess whether the data were normally distributed, histograms were analysed for all stimulus pairs. For datasets smaller than 2000 elements, it is suggested that the Shapiro-Wilk test is used, otherwise, the Kolmogorov-Smirnov test should be used [Field, 2013]. The current dataset comprises 960 values, hence the Shapiro-Wilk test was used. It was concluded that the data were mostly normally distributed, but with some exceptions: 8 of the 30 significance values were below 0.05 in the Shapiro-Wilk test. Some of the plots indicated a roughly normal distribution of the data, while others did not (Fig. 7-3).
The single sound spectral clarity of vocal and bass stimuli

As a next step, it was assessed whether any listeners were considerably less consistent than the rest, and may therefore have contributed to the uneven distribution of the data. Each listener undertook each rating twice. The notched box plot in Fig. 7-4 shows the distribution of absolute differences between the ratings for all stimulus pairs and for each listener. By considering not only the medians of the differences but also the non-parametric confidence intervals, it can be assessed whether significant differences exist between the consistency of different listeners.

Listener fifteen appears to have changed their mind the most and there is a larger median difference between his or her ratings than other listeners despite not using the scale more than other listeners: the median difference between the two stimuli on a given page was 13 for this
The single sound spectral clarity of vocal and bass stimuli

participant, the median of all differences for all listeners was 17 (values up to 50 were possible). However, there is no clear cut-off point between consistent and inconsistent listeners, as consistencies are quite varied overall. Hence, as it would be arbitrary to remove any listeners, all data are kept for the following analysis.

![Fig. 7-4: distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.](image)

All in all, the data were not quite normally distributed with some exceptions. Levene's test shows that the variances of the data are not homogenous, with a significance of less than 0.001. As the prerequisites for using parametric statistical methods are not fulfilled, mostly non-parametric methods are used with the current data set, such as interpreting plots. For the calculation of correlations, the medians of the data sets were used, rather than the means.

### 7.2.2 Centroid deviations

In previous experiments, several variations of the spectral centroid appeared to be useful predictors of single sound spectral clarity change. The harmonic centroid was the best predictor. Differences in spectral centroids (Hz), calculated linearly and in Mels, ERBs and Cents and the harmonic centroid are plotted against reported clarity in Fig. 7-5—Fig. 7-9. The lines in the plots are the non-parametric 95% confidence intervals, with the medians shown in the centre. The confidence intervals are centred on the median and extend to $\pm 1.58 \cdot \text{IQR} / \sqrt{N}$, where $N$ is the sample size and IQR is the interquartile range. The harmonic centroids for all stimuli are shown in Table 7-1.
The single sound spectral clarity of vocal and bass stimuli

Surprisingly, most equalized stimuli were judged to be less clear than the references. Clarity was reduced for all vocal stimuli, and in all but one case, the change was significant. For the bass stimuli, clarity was only increased in four cases, while ten stimuli became significantly less clear. In four bass stimuli, clarity remained unchanged.

The extent to which clarity decreases still diminishes as the centroids are raised, especially for the bass stimuli. For the vocal, the more the centroids are decreased, the more clarity is reduced. However, most confidence intervals overlap in the vocal and hence, this relationship is not significant. Both vocal and bass medians appear to be arranged in a diagonal line but with some exceptions. The corresponding correlation coefficients are shown in Table 7-2. According to the Spearman correlation coefficient, the Mel centroid is the best predictor of spectral clarity in this case. However, due to the large negative offset in the clarity ratings, the centroids can no longer predict whether clarity was increased or decreased. As mentioned earlier, many confidence intervals overlap, and hence, many clarity ratings do not differ significantly. At the same time, there are several cases, especially in the vocal, where e.g. the harmonic centroid is altered by an equal number of harmonics but in opposing directions, leading to the same resulting reduction in clarity. Equal increases in centroids also led to opposing clarity ratings in several cases.

![Graph](image)

Fig. 7-5: differences in spectral centroids, plotted against clarity ratings.
The single sound spectral clarity of vocal and bass stimuli

Fig. 7-6: differences in Mel centroids, plotted against clarity ratings.

Fig. 7-7: differences in ERB centroids, plotted against clarity ratings.
The single sound spectral clarity of vocal and bass stimuli

Fig. 7-8: differences in Cent centroids, plotted against clarity ratings

Fig. 7-9: differences in harmonic centroids, plotted against clarity ratings.
### Table 7-1: Harmonic Centroids of All Stimuli

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Harmonic Centroid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocal reference</td>
<td>20.37</td>
</tr>
<tr>
<td>Vocal 125Hz boost</td>
<td>17.59</td>
</tr>
<tr>
<td>Vocal 125Hz cut</td>
<td>22.35</td>
</tr>
<tr>
<td>Vocal 250Hz boost</td>
<td>17.01</td>
</tr>
<tr>
<td>Vocal 250Hz cut</td>
<td>23.44</td>
</tr>
<tr>
<td>Vocal 500Hz boost</td>
<td>15.72</td>
</tr>
<tr>
<td>Vocal 500Hz cut</td>
<td>25.06</td>
</tr>
<tr>
<td>Vocal 1kHz boost</td>
<td>15.84</td>
</tr>
<tr>
<td>Vocal 1kHz cut</td>
<td>25.53</td>
</tr>
<tr>
<td>Vocal 2kHz boost</td>
<td>18.61</td>
</tr>
<tr>
<td>Vocal 2kHz cut</td>
<td>22.26</td>
</tr>
<tr>
<td>Vocal 4kHz boost</td>
<td>21.99</td>
</tr>
<tr>
<td>Vocal 4kHz cut</td>
<td>19.37</td>
</tr>
<tr>
<td>Vocal 8kHz boost</td>
<td>25.69</td>
</tr>
<tr>
<td>Vocal 8kHz cut</td>
<td>17.08</td>
</tr>
<tr>
<td>Bass reference</td>
<td>18.32</td>
</tr>
<tr>
<td>Bass 62Hz boost</td>
<td>15.12</td>
</tr>
<tr>
<td>Bass 62Hz cut</td>
<td>21.36</td>
</tr>
<tr>
<td>Bass 125Hz boost</td>
<td>13.07</td>
</tr>
<tr>
<td>Bass 125Hz cut</td>
<td>24.08</td>
</tr>
<tr>
<td>Bass 250Hz boost</td>
<td>14.01</td>
</tr>
<tr>
<td>Bass 250Hz cut</td>
<td>22.76</td>
</tr>
<tr>
<td>Bass 500Hz boost</td>
<td>16.52</td>
</tr>
<tr>
<td>Bass 500Hz cut</td>
<td>19.84</td>
</tr>
<tr>
<td>Bass 1kHz boost</td>
<td>19.17</td>
</tr>
<tr>
<td>Bass 1kHz cut</td>
<td>17.50</td>
</tr>
<tr>
<td>Bass 2kHz boost</td>
<td>22.47</td>
</tr>
<tr>
<td>Bass 2kHz cut</td>
<td>15.26</td>
</tr>
<tr>
<td>Bass 4kHz boost</td>
<td>22.76</td>
</tr>
<tr>
<td>Bass 4kHz cut</td>
<td>15.35</td>
</tr>
<tr>
<td>Bass 8kHz boost</td>
<td>21.64</td>
</tr>
<tr>
<td>Bass 8kHz cut</td>
<td>16.31</td>
</tr>
</tbody>
</table>
The single sound spectral clarity of vocal and bass stimuli

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Spearman Value</th>
<th>Spearman p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel Centroid</td>
<td>0.557</td>
<td>0.001</td>
</tr>
<tr>
<td>Cent Centroid</td>
<td>0.533</td>
<td>0.002</td>
</tr>
<tr>
<td>ERB Centroid</td>
<td>0.530</td>
<td>0.003</td>
</tr>
<tr>
<td>Harmonic centroid</td>
<td>0.525</td>
<td>0.003</td>
</tr>
<tr>
<td>Linear Centroid (Hz)</td>
<td>0.487</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 7-2: Spearman correlation coefficients relating harmonics boosted and cut and centroid deviations to median clarity ratings. Although all EQ settings reduced clarity, all predictors correlate strongly positively correlation with clarity ratings.

7.2.3 Harmonics boosted and cut

The harmonic number boosted or cut was a good predictor for single sound clarity in previous studies. In the current experiment, no EQ treatments could increase clarity for the vocal and the same was the case for most bass stimuli (section 2.1). However, it is still possible that the extent to which clarity was reduced diminishes for boosts in higher harmonics and cuts in lower harmonics. The contribution of each harmonic number to target clarity is assessed in this section. In order to establish which target harmonics were boosted or cut in each stimulus pair, the octave bands were divided by the average fundamental of each programme item. The impact of boosts and cuts in different harmonics on clarity is shown in figs. Fig. 7-10 and Fig. 7-11. In previous single-sound experiments (experiments 2 and 3), boosting higher harmonics or cutting lower ones in the target increased clarity, while cutting higher harmonics or boosting lower ones reduced clarity. This is not the case here. Both boosts and cuts in most frequency areas contribute negatively. However, there appears to be a near-linear correlation between all cuts, as well as bass boosts and clarity. When the lowest two harmonics are cut in the bass, clarity increases. Cuts in higher harmonics contribute negatively to clarity increasingly much up to about the 40th harmonic. After that, the negative impact on clarity is less pronounced, presumably because the bass did not have much energy in this area. Although all boosts impact negatively on clarity, the extent to which this is the case gradually diminishes towards higher harmonics. The relationship appears to be linear with the exception that boosts near the bass fundamental reduce clarity considerably less than boosts around the 50th harmonic. Spearman correlation coefficients were calculated in order to assess the relationship between harmonics boosted and cut and clarity (Table 7-3). It can be seen that the harmonic number boosted correlates significantly positively with clarity change. However, the relationship between cut harmonics and clarity is not significant.
The single sound spectral clarity of vocal and bass stimuli

Fig. 7-10: the influence of cutting harmonics on clarity

Fig. 7-11: the influence of boosting harmonics on clarity
The single sound spectral clarity of vocal and bass stimuli

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Spearman p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmonic number boosted</td>
<td>0.629</td>
</tr>
<tr>
<td>Harmonic number cut</td>
<td>-0.089</td>
</tr>
<tr>
<td>Harmonic number boosted (bass)</td>
<td>0.539</td>
</tr>
<tr>
<td>Harmonic number cut (bass)</td>
<td>-0.548</td>
</tr>
<tr>
<td>Harm number boosted (vocal)</td>
<td>0.75</td>
</tr>
<tr>
<td>Harmonic number cut (vocal)</td>
<td>-0.321</td>
</tr>
</tbody>
</table>

Table 7-3: Spearman correlation coefficients relating harmonics boosted and cut to median clarity ratings. Although all EQ settings reduced clarity, the harmonic number boosted has a strong positive correlation with clarity ratings: the negative contribution to clarity diminishes towards higher frequencies.

7.3 Discussion

Contrary to previous experiments, it was not possible to increase the clarity of the vocal stimuli and most bass stimuli with the EQ treatments used. Clarity may have been reduced by most EQ treatments because it can perhaps only be increased up to a maximum value, after which it decreases again. The vocal stimuli may already have had maximum clarity. Another possible reason is the fact according to informal listening, most EQ treatments sounded unnatural and caused unpleasant resonances. This is likely to be an important factor, as it was mentioned by several test subjects in informal conversations: many of the listeners mentioned that they could hear which EQ band had been boosted or cut and that the EQ treatments altered the character of the sound in a way that reduced clarity. In previous experiments, the chosen EQ treatments were a good balance between audibility and objectionability. For comparability, the same treatments were used in the current experiment but it appears as though they sounded objectionable in this case. Several test subjects mentioned that to them, clarity meant the extent to which the true character of an instrument is audible. When unnatural resonances are added to a sound, however, this can no longer be the case. Therefore, the chosen EQ treatments were probably not ideal for an increase in clarity. Lastly, it is possible that the acoustic parameters of clarity differ between instruments and that different instruments need to be equalized differently for clarity.

All in all, it appears that the harmonic centroid cannot reliably predict clarity, either because it only applies to some instruments or because other important factors must be considered additionally. Therefore, it would be useful to carry out a verbal elicitation task where listeners
are asked to define what clarity means in the context of EQ adjustments and explain what key factors spectral clarity depends on. This could be combined with a clarity adjustment EQ task, where listeners increase the clarity of a range of different instruments. In this way, it is possible to see what EQ treatments exactly are useful for increasing clarity.

### 7.4 Conclusion

In previous experiments, the harmonic centroid (spectral centroid divided by a sound’s average fundamental) was established to be a useful predictor of single sound clarity, followed by the Mel, Cent, ERB and spectra centroids, as well as the harmonic number boosted or cut. Previous experiments used monophonic string programme items, as well as polyphonic guitar and piano programme items. In order to test the ability of the above metrics to predict clarity in a larger group of programme items, a similar experimental setup was used as before, this time using bass and vocal programme items.

The following research question was answered: does raising the harmonic, linear, Mel, ERB and Cent centroids, boosting higher harmonics or cutting lower ones increase spectral clarity for vocal and bass stimuli, when the same EQ settings are used as in the last experiment?

None of the EQ treatments made the vocal clearer and only two of the cuts increased bass clarity significantly. Therefore, the previous predictors can no longer successfully predict clarity. Despite this, the extent to which clarity was reduced still seemed to correspond to the extent to which linear, Mel, ERB, Cent and harmonic centroids were lowered, especially for bass programme items. The further the centroids are raised, the more the negative impact on clarity diminishes. When the lowest two harmonics are cut in the bass, clarity increases. Cuts in higher harmonics contribute negatively to clarity increasingly up to about the 40th harmonic. After that, the negative impact on clarity is less pronounced, presumably because the bass did not have much energy in this area.

It can be concluded that while clarity was not increased in most cases, the predictors found in previous single sound experiments may still be useful. It is possible that the programme items used already had maximum clarity before applying EQ, and that therefore most EQ treatments reduced clarity. Secondly, the EQ treatments used may have created unnatural sounding resonances in the sound that made the stimuli appear less clear. Positive centroid shifts may only correlate positively with clarity when the EQ treatments lie within certain gains or Q values. Lastly, it is possible that clarity works differently for different instruments. All in all, it can be concluded that while harmonic numbers boosted and cut, as well as Mel, ERB, Cent, linear and harmonic centroids can be used to predict clarity change with limited accuracy, other
The single sound spectral clarity of vocal and bass stimuli

factors may need to be taken into consideration. These may be elicited from listeners in a verbal elicitation task. A clarity EQ matching task may help establish better suited gains, frequencies and bandwidths.
8 Single sound spectral clarity adjustment task

In previous experiments, the harmonic centroid was established to be the most useful predictor for guitar, piano and string spectral clarity. However, when the spectral clarity of vocal and bass stimuli was assessed in chapter 7, spectral clarity was reduced for nearly all stimuli even when the harmonic centroid was significantly raised.

It was concluded that the harmonic centroid can predict the single sound spectral clarity of harmonic, monophonic sounds with limited accuracy and that further factors may need to be considered. These additional factors may have been moved in a negative direction by the equalisation in the last experiment. There may be certain EQ centre frequencies, gains and bandwidths that reduce spectral clarity, or combinations thereof. At the same time, spectral clarity has thus far not been clearly defined in the context of EQ adjustments. An exploratory listening test was carried out with the aim to answer the following three research questions:

1. What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ?
2. How can spectral clarity be defined in the context of EQ adjustments?
3. What additional factors are likely to be important for spectral clarity?

The new listening test was designed as a combined matching and verbal elicitation task, where participants increased the clarity of programme items with EQ controls, commented on the difficulties they experienced and provided their definitions of clarity. In this way, additional factors for inclusion in an improved spectral clarity predictor, as well as an overall definition for spectral clarity, were sought. The favoured centre frequencies, gains and bandwidths for increasing clarity could be elicited from the settings that participants used. In section 8-1, the setup for the new listening test is explained. In section 8-2, the quantitative and qualitative data area analysed. Section 8-3 is a discussion and section 8-4 is a conclusion.

8.1 The setup for the new listening test

21 listeners were asked to maximise the spectral clarity of programme items in a GUI that was prepared in MaxMSP (Fig. 8-1), using parametric EQ with a frequency, gain and
8 Single sound clarity adjustment task

bandwidth control each (three rotary controls). An EQ stimulus volume control, a rating slider, a text box for entering comments and buttons for playing and stopping audio and navigating pages were also included.

Fig. 8-1: an example of a test page

Each programme item was presented on a separate page featuring play and stop buttons, as well as a reference button, allowing listeners to turn off their EQ settings for comparison. Stimulus order was randomized in order to mitigate sequential biases. The stimuli played repeatedly until stopped or until a new page was loaded. Once listeners had completed the task, they could move to the next page by pressing the “>” button. Each software button was mapped to a hardware button on a Novation Launch Control MIDI controller. The LED lights on the hardware controller were programmed to indicate whether the R or EQ stimulus was selected, whether the audio was playing, paused or stopped and whether the “next page” button was active.

On each new page, all hardware controls had to be reset to the middle position before audio was enabled. A reminder to reset the controls and a red overlay were shown in the software GUI. This ensured that listeners did not merely adopt their previous EQ settings. At the same time, healthy listening levels were maintained in this way. Each listener was provided with detailed instructions at the beginning of the test (Appendix A1, fig. A6). They were given the opportunity to try out the test interface before starting.
8 Single sound clarity adjustment task

the test, completing two or more pages to practice. All controls are explained in more
detail in the following: in section 8.1.1, more information is given about the EQ controls.
In section 8.1.2, loudness matching is discussed. In section 8.1.3, the ratings slider and
text box are introduced. Section 8.1.4 clarifies the use of additional attributes. Section
8.1.5 mentions the pilot study. In section 8.1.6, the stimuli are introduced. In section
8.1.7 the presentation of the stimuli to the test subjects is explained and in section 8.1.8
the type of listeners is introduced.

8.1.1 EQ controls

As mentioned above, one control each was included for gain, frequency and Q. Previous
listening tests have shown that one boost or cut can be used to vary spectral clarity
significantly. At the same time, more controls would make the analysis more
complicated. As shown in section 8.2, it was possible to answer the above research
questions without further EQ controls. In order to avoid visual bias, no spectrogram was
included and no calibration markings in Hz or dB were shown. This was due to the fact
that listeners may have read recommendations regarding the ‘correct’ equalisation of
sounds and that they may therefore have followed the visual markings rather than using
their ears.

The range for the frequency control was roughly 20Hz to 20kHz, similarly to typical
parametric EQs. An informal search of commercial EQ plugins showed that typical gain
ranges usually do not exceed +/-30dB, hence this approximate gain range was used for
the current listening test. Possible Q values ranged from 0.01 to 11.22, which offered a
wide enough range of bandwidths while still allowing for fine Q adjustments. As
mentioned above, the EQ controls had to be reset to their centre position on each new
page in order to ensure that listeners did not merely adopt their previous settings.
Additionally, the movements on all controls were recorded in order to ensure that
listeners did try out different settings. Fig. 8-2 shows a typical example of fader
movement. As can be seen here, the listener tried different combinations of parameter
settings. This was the case for most pages and listeners.
8.1.2 Loudness matching

It is possible that loudness correlates with clarity, as both relate to the audibility of a stimulus. In order to remove the possibility of participants merely boosting the frequency area the ear is most sensitive to for increased loudness (3kHz–5kHz, chapter 3), participants were asked to adjust the loudness of the equalised version to be same of the reference stimulus with a volume fader (mapped to a rotary control on the MIDI controller). The references were loudness matched in advance in order to ensure comfortable listening levels overall. Like all controls, the loudness fader had to be reset to its middle position on each new page before audio could be played. In this way, unhealthy volume levels, caused by a gain setting that was appropriate for the preceding, but too loud for the current page, could be avoided.

8.1.3 Rating and text box

Listeners were asked to rate how much clearer they perceived their version to sound in comparison to the original, using a slider: it is possible that some stimuli already had maximum clarity to begin with, or that the EQ controls were not sufficient to make the sounds clearer which could be assessed in this way. It is also possible that by inviting listeners to assess their settings, this may have provided additional motivation to do
8 Single sound clarity adjustment task

better. On each page, listeners were asked to enter text into a box, explaining what stopped them from making the sound even clearer. In this way, the factors that clarity depends on and any limitations of the experimental procedure could be elicited. Listeners could only move to the next page once text had been entered. In the cases where the clearest imaginable version of the sound had been achieved, listeners were asked to enter “n/a” in the text box.

At the end of the test, listeners were asked to provide their definition of clarity. The definitions were collected at the end of the test as at this point, listeners likely had a clearer idea of what clarity meant to them than at the beginning. In order to ensure that listeners did adjust clarity consistently throughout the test, they were also asked whether they knew what clarity meant to them before starting the test.

8.1.4 Additional attributes

In previous listening tests, listeners were asked to rate the clarity of stimuli. Brightness, warmth and fullness ratings were included occasionally in order to try to reduce the noise in clarity ratings. It is possible that mixing engineers do not usually attempt to make single sounds as clear as possible, as this may reduce e.g. warmth or overall quality. It is important that this is explained to the test participants. Therefore, additional, crossed out attributes were included at the top of each test page, like in previous listening tests, namely brightness, warmth and quality. Quality was included as this may be what test subjects usually try to achieve when equalising sounds.

8.1.5 Pilot Study

The functionality and feasibility of the test setup were assessed in a pilot study, using three participants. In this way, it was ensured that the number of stimuli was suitable to the average time spent on each test page. The interface worked properly and the participants found the task manageable.

8.1.6 Stimuli

The nature of the current study was exploratory, i.e. the aim was to elicit further important factors of single sound clarity and clarity definitions. In order to elicit as many important factors as possible, stimuli were chosen that differed in the factors mentioned below.
Clarity reducing resonances

As mentioned above, it is possible that the EQ caused unpleasant resonances for the previously tested vocal and bass stimuli, leading to a reduction in clarity. Therefore, resonant stimuli were created where a sharp peak had been inserted with EQ. The following considerations informed the choice of programme items:

- In order to test whether a given EQ adjustment was influenced by resonances in the sound and not by other characteristics of the programme item, several versions of the same stimulus were included for two programme items (vocal and violin), with and without resonances.

- It is possible that resonances only reduce clarity when they are not naturally part of the sound. For this reason, an Erhu sound (Chinese violin) was included, as it features a natural, characteristic resonance around 500Hz. The Erhu sound was taken from a professional film score to ensure high quality. In this way, the presence of other clarity reducing factors was ruled out.

- As resonances seem to have played an important role for vocal and bass stimuli, the same voice and bass, recorded under the same conditions, were included again. The verbal elicitation task could then offer further clues as to what may have caused a degradation in clarity.

Maximum centroids

Some programme items were processed to have very high harmonic centroids before EQ in order to test whether a saturation point may be reached, where other clarity factors become more important. In order to ensure that this did not cause unnatural resonances, spectral tilt was used to maximise original centroids. Like for resonances, several versions of the same programme items, with and without added “brightness EQ”, were included to be able to elicit further clarity factors independently from the original centroid position.

Variation in instrument types

Eight different instruments were used, including harmonic, monophonic sounds and one inharmonic sound (crash cymbal). In this way, the clarity factors that may differ between instruments could be varied.

Previously, listening test subjects mentioned in informal conversations that for them, the audibility of the natural characteristics of sources contributes to spectral clarity. Hence, it appeared useful to include sources with such important characteristics, such as
string vibration and bow noise in cellos, air vibration and turbulence in brass instruments or consonant sounds in voices. Sounds without articulation sounds were also included. An interesting consideration is the clarity of newly synthesized sounds: here, it is unclear which characteristics are particularly important or 'natural'. Therefore, one of the stimuli was synthesized in Sylenth (Lennar Digital), such, that it did not resemble any typical instrument timbre.

In order to test to which extent the harmonic centroid is a useful predictor of single sound spectral clarity, different combinations of high and low fundamentals and bandwidths were also included. Lastly, the sounds differed in their temporal behaviour, e.g. some plucked and some bowed sounds were included. (Following informal discussions with test participants, it is e.g. possible that the spectrum at the beginning of notes is the most important for clarity).

Little variation over time

Each instrument played only a single note. If clarity had varied too much over the course of an excerpt it would have been unclear in which way listeners made clarity judgments and adjustments as a result. At the same time, the ideal EQ so far seemed to depend on the fundamental (harmonic centroid).

Stimulus sources

In order to be sure that certain EQ adjustments took place because of the factors mentioned above (e.g. resonances), rather than because of other clarity degrading factors, high quality microphone recordings were used. Most of the stimuli were Apple loops, some were high quality samples from the East West Composers Complete library and one stemmed from a professional film score (Erhu). Only the vocal was recorded under less than ideal conditions (small, untreated room), as this was the case in the last experiment also. The stimuli were exported as 44.1kHz and 24 bit wave files after being processed with Logic Pro 9's parametric equaliser where relevant and loudness-matched by four audio professionals. The level was set to minimize the possibility of clipping during the EQ adjustment task.

Final choice of stimuli

The following list of programme items was used:

- Plucked, electric bass (the same bass as in chapter 7)
8 Single sound clarity adjustment task

- Two versions of a bowed cello sound, one of which was equalized to sound ‘too bright’ for comparison

- A crash cymbal (inharmonic sound)

- An Erhu (sound featuring a natural, strong resonance)

- The sung syllable "sit", sung by the same singer and recorded under the same conditions as in chapter 7, this time including high frequency articulation sounds (the two consonants). Two additional versions were created in Logic Pro 9, one sounding ‘too bright’ and one sounding resonant. For the bright version, a shelf filter was added, boosting frequencies above 1kHz by about 9dB and cutting frequencies below 9dB by 24dB. For the resonant version, a sharp, thin 19 dB boost was inserted at 3.4kHz. These settings fulfilled the requirements according to informal listening by the researcher, without causing clipping.

- A non time-varying synthesized sound (in order to test how clarity is judged when no natural reference exists)

- A trumpet (including HF and LF ‘ripping’ articulation sounds)

- A plucked violin sound, including one additional version that was equalized to have an unpleasant resonance, similarly to the vocal. The violin has a lower harmonic centroid than the cello.

The chosen stimuli are related to the above list of factors in Table 8-1.
<table>
<thead>
<tr>
<th></th>
<th>Has AS (high)</th>
<th>Has AS (low)</th>
<th>Has no AS</th>
<th>Variation: high and low FOs (and high and low reference HCs)</th>
<th>Constant spectrum</th>
<th>Time var. spectrum</th>
<th>Natural, strong res.</th>
<th>No res., not too bright</th>
<th>Too bright</th>
<th>Unpleas. res.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bass</td>
<td>x (pluck noise)</td>
<td>65.4 Hz (7.4)</td>
<td>x (more or less)</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowed cello, too bright</td>
<td>x (bow noise)</td>
<td>131 Hz (36.6)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowed cello, no EQ</td>
<td>x (bow noise)</td>
<td>131 Hz (12.4)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crash cymbal</td>
<td>x (hit)</td>
<td>Broad spectrum, goes down to 150 Hz (33.7)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erhu</td>
<td>x (bow noise)</td>
<td>246 Hz (7.8)</td>
<td>x (except slight vibrato)</td>
<td>X around 500Hz and 1 khz</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Sit”, too bright</td>
<td>x (‘s’, ‘t’)</td>
<td>344 Hz (14.1)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Sit”, resonant</td>
<td>x (‘s’, ‘t’)</td>
<td>344 Hz (10.4)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Sit”, no EQ</td>
<td>x (‘s’, ‘t’)</td>
<td>344 Hz (11.0)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static synth</td>
<td>x</td>
<td>659.25 Hz (9.7)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trumpet</td>
<td>x (ripping sound)</td>
<td>139 Hz (11.9)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plucked violin, resonant</td>
<td>x (low pluck sound)</td>
<td>523 Hz (2.2)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plucked violin, no EQ</td>
<td>x (low pluck sound)</td>
<td>523 Hz (2.8)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-1: the chosen programme items vary in the factors that may be important for clarity. ‘AS’ stands for ‘articulation sounds’ and ‘HC’ stands for ‘harmonic centroid’
8 Single sound clarity adjustment task

8.1.7 The presentation of the stimuli to test subjects

The test interface and incorporated stimulus pairs were presented via high quality headphones, using a VRM box interface like in the previous two listening tests. The stimuli in the first three listening tests were presented at a comfortable listening level in an international-standard listening room built to ITU-R Standard BS1116, using Bowers & Wilkins Nautilus 801 speakers. However, by using headphones, the possibility of room reflections and inter-speaker comb-filtering introducing listener-position-dependent spectral irregularities could be removed completely.

8.1.8 Type of listeners

Students of the Surrey University Institute of Sound Recording (some undergraduates and some postgraduate researchers) were recruited for the listening test. Additionally, some more experienced mixing engineers were employed. It is possible that more experienced mixing engineers are influenced by the way in which they usually equalize sounds and literature they have read, increasing sound quality, rather than making sounds as clear as possible (this may be different from what is usually required in a mixing context). Less experienced listeners, however, may not know how to achieve the desired sound. Hence, a mixture of both was deemed useful.

8.2 Results and analysis

In the following, the results of the study are presented. In section 8.2.1, the possibility of using parametric statistics methods is assessed. In section 8.2.2, the adjustment ratings data is presented. In section 8.2.3, the favoured EQ settings, as chosen by the test subjects, are presented. In section 8.2.4, the approach to analysing the qualitative data is presented. In section 8.2.5, the clarity definitions are examined and in section 8.2.6, the important factors of spectral clarity are presented.

8.2.1 Requirements for parametric statistics

The nature of the data suggested the use of mostly non-parametric statistical analysis methods. For parametric tests, the following conditions had to be fulfilled:

1. Independence
2. Interval data
8 Single sound clarity adjustment task

3. Normally distributed data

4. Homogeneity of variance

It is unclear whether independence was fulfilled for the same reason as in previous experiments. In order to assess whether the data were normally distributed, histograms and quantile-quantile plots were analysed for all EQ controls (the histograms for the controls are shown in the next section). The Q and ratings histograms are clearly skewed which points towards non-normality.

Altogether 252 EQ adjustments were undertaken, including three EQ adjustments and the ratings and volume slider. Hence, a Shapiro-Wilk test was carried out, as recommended by Field [2013]. Out of the 12 values for each stimulus pair, 6 frequency significance values, 12 Q values, 6 gain values, 7 ratings values and 2 volume slider values were below 0.05. It was concluded that the data were not normally distributed and therefore, mostly boxplots, scatter plots and histograms were used for analysis.

8.2.2 Ratings data

On each test page, listeners rated on a slider how much clearer they managed to make the sound. By inspecting a histogram of the ratings data (Fig. 8-3), it can be concluded that most subjects were successful in their EQ tasks and that clarity could be increased in most cases. Over 50 adjustments were rated at a 100 and less than 10 were rated around 0. This shows that it is possible to increase clarity with just three EQ controls.

The ratings cannot be used to measure the extent to which clarity was increased, as no two adjustments were identical, and therefore only one rating exists for each adjusted EQ/R stimulus pair. Furthermore, listeners appear to have used the scale differently. Scatter plots of the ratings, plotted against the resulting change in spectral centroid show that each listener used a different part of the scale (Fig. 8-4, Fig. 8-5 and Fig. 8-6).

Listeners were asked to give a rating between “not clearer” and “this is the clearest version of the sound I can imagine”. Some seem to have interpreted the endpoints of the scale as the extreme values possible within the constraints of the experimental setup and their own skill level, whereas others were influenced by their negative opinion of their skill level, rating their adjustments around the bottom half of the scale. Again others used the top half of the scale, presumably as they felt content with their ratings in all cases. Some listeners also rated all adjustments similarly to each other, possibly because they put the same amount of effort into the adjustment task on each page. Another difficulty lay in the fact that as opposed to previous experiments, where
8 Single sound clarity adjustment task

listeners were presented with a familiarization page, listeners were not aware of the full range of possible clarity changes until after the test, when all adjustments had been completed.

All in all, the fact that listeners were rating their own EQ adjustments seems to have biased their ratings. As all listeners essentially rated different stimuli, the ratings cannot be normalized. It can be seen however, that nearly all ratings are greater than 0 and that therefore, clarity was nearly always increased.

![Histogram showing the ratings given by listeners](image1)

**Fig. 8-3:** histogram showing the ratings given by listeners

![Scatter plot of linear spectral centroid shifts](image2)

**Fig. 8-4:** scatter plot of linear spectral centroid shifts, plotted against ratings for one listener. This listener uses the top part of the slider only (grey points correspond to stimuli with unnaturally high initial centroids).
8.2.3 What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ?

It was previously hypothesized that the EQ treatments on the vocal and bass may have lowered clarity because they created unpleasant resonances. For the experiments using...
8 Single sound clarity adjustment task

piano, guitar and string stimuli, 9dB boosts and cuts provided a good balance between audibility and objectionability. For consistency, the same EQ treatments were used on the vocal and bass. However, in this case, the EQ treatments may have sounded objectionable.

In the following, histograms are shown for the preferred Q (Fig. 8-7 and Fig. 8-8), gain (Fig. 8-9) and frequency settings (Fig. 8-10), as chosen in the adjustment task. In most cases, the widest possible bandwidths, that is the smallest possible Q values (0.01), were used. The biggest Q value (around 11) was used slightly more often than slightly wider bandwidths and this may have been due to the fact that some participants tried to target the thin, unpleasant resonances in some stimuli.

As can be seen in the gain plot (Fig. 8-9), it appears that boosts were preferred to cuts, with a median for boosts around 9dB and the median for cuts around -7dB. This is fairly close to the gains employed in previous listening tests; hence the previously used gains alone could not have led to a clarity reduction. Frequencies between 1.7kHz and 7.4kHz were most often treated; hence high frequency boosts were preferred to low frequency cuts for raising the centroids, using wide bandwidths. This is in line with Pestana and Reiss’ findings where wide Qs are preferred for boosts [2014]. So far, the only control that differed significantly from previous settings was the Q control, where much wider bandwidths were preferred for clarity.

![Fig. 8-7: histogram showing the Q values used. The smallest Q values (widest bandwidth) are used most often.](image-url)
Fig. 8-8: histogram showing the smaller Q values used only.

Fig. 8-9: histogram showing typical gain settings. Boosts are preferred.
Fig. 8-10: histogram showing the frequencies most often targeted

It is possible that the preferred treatments differ between stimuli and that thinner bandwidths (greater Q values) are more likely to occur for the resonant stimuli. The boxplots below show Gain (Fig. 8-11), Q (Fig. 8-12) and frequency settings (Fig. 8-13 and Fig. 8-14), separately for each stimulus pair. Almost all confidence intervals overlap, hence, there do not seem to be significant differences between stimuli. Listeners did not seem to try to remove the added resonances, apart from a few exceptions: the means for the resonant stimuli lie slightly higher than the other means but confidence intervals still overlap. Resonances were mentioned frequently in the qualitative data, however (sections 8.2.5 and 8.2.6), and it is likely that their removal was merely too difficult. The mean for the Erhu stimulus is not significantly higher than that of the other stimuli but the interquartile range appears to extend further upwards than that of other stimuli. It is possible that therefore, some listeners did attempt to remove the natural resonance in the Erhu. At the same time, listeners reported that they found it difficult to remove the resonances they could hear and that they would have preferred to use more EQ controls, presumably in order to be able to tackle both resonances and HF/LF balance.
8 Single sound clarity adjustment task

Fig. 8-11: preferred gain settings for each programme item, plotted separately for boosts and cuts.

Fig. 8-12: preferred Q settings, plotted separately for each programme item. Some outliers are not shown as they lie around the biggest possible Q setting and would hence cause the boxes to be compressed vertically. One outlier each exists for ‘sit bright’ and ‘sit rez’.

In the frequency settings box plot, most notches seem to overlap for the boosts and the medians for the ‘brighter’ stimuli are not significantly lower than those for the other
stimuli. The median for the plucked violin stimuli appears to be a little higher than the other medians, presumably as the violin had a higher fundamental than other stimuli. It appears that therefore, a further increase in harmonic centroids still contributed positively to clarity even when the harmonic centroids were already high.

For the cuts, larger differences seem to exist between stimuli. Most notably, the medians for the 'bright' stimuli and the synthesized sound are higher than the medians for the corresponding boosted frequency. No participants appeared to boost lows for the bright stimuli apart from the bright version of the cello. Apart from that, cut frequencies appear to correspond to the fundamental frequency.

Fig. 8-13: preferred boosted and cut frequencies, plotted separately for each programme item. Another box plot was created for the harmonics boosted and cut. In most cases, the harmonics cut lie below the harmonics boosted except for the 'too bright' stimuli and the synth. This confirms once again that increased HF content and decreased LF content corresponds to an increase in clarity.
8 Single sound clarity adjustment task

So far it has been shown that high frequency boosts are preferred to low frequency cuts, that wide bandwidths are most often used and that for most stimuli, the ratio between high and low frequencies was increased. It has not yet been tested whether the preferred EQ controls correlate with each other, e.g. whether larger Q values are more likely for cuts as Reiss and Pestana claim [2014]. At the same time, it may be possible that larger Q values may correspond to smaller gain values, as resonances may reduce clarity less when they are quieter. Therefore, the three EQ controls were compared in pairs in scatter plots (Fig. 8-15, Fig. 8-16 and Fig. 8-17). No correlation can be seen in any pair of controls. The three controls were also examined in a three-dimensional scatter plot (not included here) and no correlation was found there, either. The correlation coefficients (Pearson) for all pairs are shown below.

Gain–Frequency: $r=0.361\, (=p<0.001$

Q–Frequency: $r=-0.018\, =p=0.773$ (not significant)

Gain–Q: $r=-0.194\, p=0.001$

A multiple correlation was carried out for all three controls and again, no correlation was found. All in all, it can be concluded that the preferred settings for the EQ controls do not depend on each other. The main difference between the adjusted settings and settings used for previous experiments was the fact that preferred Q values were much
8 Single sound clarity adjustment task

smaller (and therefore, bandwidths much larger) than the previously chosen octave bands.

![Graph showing no correlation between preferred gain and frequency settings.](image1)

Fig. 8-15: no correlation exists between preferred gain and frequency settings.

![Graph showing no correlation between preferred gain and Q settings.](image2)

Fig. 8-16: no correlation exists between preferred gain and Q settings.
8.2.4 Approach to analyzing the qualitative data

For the analysis of the qualitative data collected during the listening test, a researcher based at another university was contacted. They were given the clarity definitions and qualitative data collected on each test page and asked to find and highlight all clarity attributes they could find in the text. The researcher had experience in verbal profile analysis and knew nothing about the current project. In this way, lone researcher bias was counteracted [Burnard et al., 2008]. The researcher was not biased by prior knowledge of clarity factors. He also had no personal "investment" in a particular outcome or finding [Simmons and Gregory, 2003]. This made it possible to use the data itself to derive at the structure of analysis (inductive approach, [Burnard et al., 2008]), allowing for an unbiased elicitation of clarity attributes.

The method employed was thematic content analysis, where themes and categories that emerge from the data are identified [Burnard et al., 2008]. This is essentially a grounded theory approach [Bryman, 2008, and Simmons and Gregory, 2003], as the researcher was not given any information about the listening test or research questions. As suggested by Simmons and Gregory [2003], no a priori formulations of problems, issues, hypotheses, or theories were included and no presumption of the relevance of a
particular type of factor was made. Simmons and Gregory [2003] stress the usefulness of methodological pluralism, i.e. in addition to establishing clarity definitions and factors, the most important factors can be used for inclusion in a mathematical model that can be tested on previous data in future chapters.

8.2.5 How can spectral clarity be defined in the context of EQ adjustments?

In order to understand spectral clarity more fully, it would be useful to define it. All subjects were asked to write down their definition of clarity at the end of the test. The external researcher examined all data and structured the definitions into prominent themes, categorized into factors relating to amplitude, envelope, expectations, frequency content, “readability”/intelligibility [sic], recording artefacts, reverb and hedonic responses (included in Table 8-2). All categories for all qualitative data are shown in appendix 2, including definitions.

Presumably due to the nature of the task being an EQ task, the category that was mentioned most often in the clarity definitions was frequency content (22 instances). Many of the phrases within this category appear to be directly related to the harmonic centroid, namely HMF and HF content (12) and frequency balance (4). Other factors may not be measurable by the harmonic centroid alone, such as resonance (3 mentions), and boosts and cuts to specific frequency areas (1). Both appear to be related to clarity reducing lumps or ‘holes’ in the spectrum, rather than overall frequency balance. This seems to support the assumption that clarity reducing resonances may have been caused by the EQ in chapter 7.

In many definitions, ‘clarity’ is linked to naturalness or expectations of how the instrument should sound (7) and it is likely that gentle boosts to HF boosts and LF cuts can increase this (since most sounds have less energy in the upper parts of their spectra). Perhaps similarly to this, factors related to the “readability” of an instruments’ characteristics were elicited from 13 participants. It is assumed that elements that are not naturally part of the sound, such as unpleasant resonances, are likely to reduce clarity, hence this may be indirectly related to the point above.

Seven listeners mentioned that artefacts, distortion and spill from other instruments can reduce clarity. This seems to support the fact that high quality mixes should be free from technical faults (literature review). Three listeners each mentioned the amplitude envelope and reverberation as important clarity factors. In has been shown in the
literature review that both can indeed be related to clarity, in the context of loudness and concert halls. More information about the relation of the elicited clarity factors with previously found literature can be found in Table 8-3. It is unlikely that the EQ had a significant impact on either reverb or amplitude envelopes, and hence they appear less useful in an overall clarity model than the other factors mentioned above. Lastly, pleasantness was mentioned once. It is assumed that pleasantness depends on the lower level factors listed above, (e.g. resonances might lead to unpleasantness) and possibly personal taste.

Considering the factors that were mentioned most often, spectral clarity in this context could be defined as the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard. The following listener definitions support this.

“Clarity is for me how well you can hear the intended, or what one perceives to be the intended, sound”. Clear sounds contain “partials in amplitudes that I’m expecting to hear”. The sound is not “timbrally wrong”. “I would describe clarity as a sound that has a natural sound”. It is easy to “hear sounds that are part of character” (of the instrument). “Clarity is when you can hear all the nuances of the actual sound, without the sound becoming too sharp or unpleasant to listen to”. “A sound is heard with as much or enhanced definition as it would in a real live situation”. “Clarity means the instrument has a clear identity, [...] its harmonic characteristics are apparent”. It is “easy to pick out what you want to hear from the desired source”, i.e. clarity is the “complete intelligibility of all elements of a sound”. “The property of a specific voice or musical instrument to be recognised” is audible. In clear sound, the “character of the instrument can be heard”. “Clarity means being able to hear all the components of a sound or mix individually without any of them getting in the way of any others, i.e. without any two sounds occupying the same space in terms of frequency. It also tends to mean something with a lot of HF and not much lower mid or sub bass frequencies, which may just be a product of the first definition”.

It seems likely that the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard, i.e. spectral clarity, might be increased by gentle HF boosts or LF cuts (since most sounds have less energy in the upper parts of their spectra), and avoidance of potentially distracting unnatural/unpleasant resonances and other artefacts.
8.2.6 What additional factors are likely to be relevant for spectral clarity?

From both the clarity definitions and the text entered on each clarity adjustment page, important clarity factors could be elicited. Similarly to the definitions, most factors were related to the overall frequency content of the sound (Table 8-2, mentioned altogether 158 times). Again, this is most likely due to the fact that spectral clarity was investigated and that therefore, spectral factors are likely to be the most important. As in the last section, most of the frequency related factors are likely to correlate with the harmonic centroid, namely bandwidth, brightness, frequency balance, HMF, HF, LMF and LF, warmth and thinness. It is therefore likely that the harmonic centroid is still useful in an overall clarity model. Interestingly, where harmonic centroids had been raised prior to the experiment, listeners were unsure whether to increase them even further or whether to lower them. In the 'bright' version of the vocal, one listener wrote “seems like there is a roll off of the low frequencies” and another wrote “stimuli [sic] needs more bass content” while another said that “removing all of the hiss in the recording would make the balance too LF heavy.” Some listeners tried to apply EQ that would raise the centroids even further but disliked the resulting effect: “adding even more of the frequency I boosted started to make the sibilance uncomfortable”. A similar disagreement could be found for the 'bright' cello, where some found that “the original stimulus lacks bass” and that the “sound source was too thin”, while five listeners thought that the sound was “already quite clear”. Again others even considered raising the centroids further. One listener felt like he needed “another EQ band to perhaps boost something in the highs for a little more clarity”. Overall, factors relating to the overall frequency balance still appear to be important.

Additional spectral factors were found that do not necessarily correlate with the harmonic centroid. These are cuts and boosts to specific frequencies, harmonic content, muddy/muffled, resonances and thinness. These factors may be related to the amount of energy in specific frequency bands, such as the presence or absence of strong peaks in the spectrum. Cuts and boosts to specific frequencies had 49 mentions, which is a fairly large number; hence the presence of 'lumps' in the spectrum seems to be a useful factor to investigate. Listeners commented on resonance most often for the resonant version of the vocal, stating that it was difficult to remove, writing e.g. “couldn’t remove HF resonance”, “there's a certain frequency also bugging me that I'd like to remove but can’t”. Similarly, ringing was mentioned four times for the vocal. It is possibly that
listeners are more sensitive to distortions in the human voice than other sounds because they have a clearer internal reference for its natural sound. Within the recording artefacts category (49 mentions), ringing was mentioned altogether 9 times which may be related to resonance. Sibilance was mentioned 4 times, which may be related to uncomfortable resonances near the ‘s’ consonant in the vocal. All these factors appear to be related to clarity reducing ‘lumps’ in the spectra of sounds. Further additional clarity factors are summarized in the following paragraphs.

Other factors within the recording artefacts category are likely to be difficult to tackle with EQ, such as artefacts, distortion, Doppler, noise and hiss and spill from other instruments and were also mentioned less often than the factors mentioned above. Interestingly, performance artefacts were most often found in the bass, despite the fact that this was a high quality sample. It is likely that listeners were referring to the string plucking sound here.

Factors related to “readability” (term used by the external researcher) and intelligibility were mentioned altogether 14 times. Intelligibility was only mentioned for the voice. Listeners pointed out that to them, a clear version of the sound sounded as “intended” and that it was possible to “hear all the components of the sound” and “the character” of the sound. These factors appear to be related to an internal reference for the sound, possibly related to the “expectations and naturalness” related factors, which were elicited altogether 9 times. Interestingly, many listeners did not appear to be familiar with the Erhu sound and hence expressed that they found it difficult to adjust its clarity. Many assumed that it was intended to be a low quality sound, although it stemmed from a high quality film score recording. One listener said that it was “hard to cut a previously boosted frequency accurately”, probably referring to the natural, strong resonance in the Erhu. Another listener wrote that it sounded “like it has been low pass filtered and has some weird high resonant reverb on it”. A third listener said: “unsure what I am listening to so I do not know how to make it clearer, it could be a person singing nasally, a goat bleating or a bad synth violin.” In all these cases, the natural, internal reference for the sound was lacking, making it more difficult to equalize it and therefore, the resonance was perceived as unnatural and unpleasant.

Hedonic factors relating to quality and pleasantness were mentioned 40 times. As argued in the last section, it is possible that these factors depend at least in part on the lower level perceptual attributes elicited here, i.e. ringing could lead to unpleasantness. Reverb and room artefacts were mentioned 14 times and it is likely that these factors are also difficult to remove with just one EQ control.
8 Single sound clarity adjustment task

Envelope related factors were elicited 8 times, loudness related factors were found 3 times and in 20 instances, listeners felt that the sound was already clear and needed little further adjustment. Many of the factors can be related to the clarity factors previously found in the literature (chapter 3) and this is shown in Table 8-3 for each factor, although some of the factors elicited in the current experiment were not explicitly mentioned in the literature, i.e. the spectral factors that appear to be related to clarity reducing ‘lumps’ in the spectra of sounds (i.e. resonances and ringing), factors relating to expectations and naturalness, “Doppler effect” and pleasantness/hedonic responses.
## Single sound clarity adjustment task

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amplitude</strong></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Loudness</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Envelope</strong></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amplitude envelope</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attack of transients</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Expectations</strong></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>7</td>
<td>7</td>
<td>9</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corresponds with expectation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naturalness</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td><strong>Frequency content</strong></td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>14</td>
<td>11</td>
<td>20</td>
<td>8</td>
<td>11</td>
<td>10</td>
<td>12</td>
<td>8</td>
<td>22</td>
<td>158</td>
<td></td>
</tr>
<tr>
<td><strong>Bandwidth</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Brightness</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Cuts/boosts/specific frequencies</td>
<td>3</td>
<td></td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>Frequency balance</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td><strong>Harmonic Content</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HMF&amp;HF content</strong></td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td><strong>LF content</strong></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td><strong>LMF content</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td><strong>Muddy/muffled</strong></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Resonances</strong></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
<td>1</td>
<td>3</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thin</strong></td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td><strong>Warmth</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>&quot;Readability&quot;</strong>/intelligibility</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td><strong>&quot;Readability&quot;</strong></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intelligibility</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recording Artifacts</strong></td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>47</td>
</tr>
<tr>
<td><strong>Artifacts</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Distortion</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Doppler</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Noise &amp; hiss</strong></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance artifacts</strong></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ringing</strong></td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sibilance</strong></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Spill</strong></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reverb</strong></td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>13</td>
<td></td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reverb</strong></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Room artifacts</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hedonic</strong></td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td></td>
<td>1</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-2: elicited clarity factors found in the clarity definitions and the qualitative data collected on each test page. Sums are shown per factor and programme item and overall.
<table>
<thead>
<tr>
<th>Spectral single sound clarity adjustment task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TIMBRAL CLARITY</strong></td>
</tr>
<tr>
<td>Amplitude/Loudness</td>
</tr>
<tr>
<td>Envelope/amplitude envelope, attacks</td>
</tr>
<tr>
<td>Expectations/naturalness</td>
</tr>
<tr>
<td>Frequency content</td>
</tr>
<tr>
<td>Bandwidth</td>
</tr>
<tr>
<td>Brightness</td>
</tr>
<tr>
<td>Cuts/boosts/specific frequencies</td>
</tr>
<tr>
<td>Frequency balance</td>
</tr>
<tr>
<td>Harmonic Content</td>
</tr>
<tr>
<td>HMF&amp;HF content</td>
</tr>
<tr>
<td>LF content</td>
</tr>
<tr>
<td>LMF content</td>
</tr>
<tr>
<td>Resonances</td>
</tr>
<tr>
<td>Thin</td>
</tr>
<tr>
<td>Warmth</td>
</tr>
<tr>
<td>“readability”/intelligibility</td>
</tr>
<tr>
<td>Recording artefacts</td>
</tr>
<tr>
<td>Distortion</td>
</tr>
<tr>
<td>Doppler</td>
</tr>
<tr>
<td>TIMBRAL CLARITY</td>
</tr>
<tr>
<td>-------------------------</td>
</tr>
<tr>
<td>Noise &amp; hiss</td>
</tr>
<tr>
<td>Performance artefacts, room artefacts</td>
</tr>
<tr>
<td>Ringing</td>
</tr>
<tr>
<td>Sibilance</td>
</tr>
<tr>
<td>Spill</td>
</tr>
<tr>
<td>Reverb</td>
</tr>
<tr>
<td>Pleasantness/quality</td>
</tr>
</tbody>
</table>

Table 8-3: relation of elicited clarity factors to factors previously found in the literature

### 8.3 Discussion

In order to improve the existing clarity model, it is necessary to investigate why the harmonic centroid was a poor predictor of spectral clarity for the vocal and bass stimuli. One or more additional factors may need to be considered. The harmonic centroid may only be useful for string, piano and guitar stimuli and not generally applicable. One of the main aims of this chapter was to elicit the most important factors of clarity, which made it possible to investigate this further.

Spectral factors were mentioned most often in the qualitative data (158 factors) and many of these factors can be related to the overall frequency balance, as measured by the harmonic centroid (91 of the written phrases). Hence, the harmonic centroid still appears to be an important parameter of spectral clarity. All other elicited spectral factors appear to be related to clarity reducing ‘lumps’ or ‘holes’ in the spectrum, i.e. cuts and boosts to specific frequencies (48), harmonic content (8), muddy and muffled sound (2) and resonances (16). Ringing (9) and sibilance (4, recording artefacts category) may also be related to this phenomenon.

It is possible that there are specific frequency areas that contribute more to clarity than others, independently from the fundamental or centroid. At the same time, clarity reducing resonances can also only be tackled with boosts and cuts in specific frequency areas. The latter is in line
with informal discussions with test subjects in the experiment summarized in chapter 7, where it was stated that unpleasant resonances had been caused by the inappropriate EQ settings. Since the relationship between resonances and unpleasantness was derived mostly from informal conversation, further research in this area, i.e. a formal verification of this relationship would be useful. However, these resonances seem to dominate the audible changes in the most unclear stimuli, compared to the reference. This is further supported by the fact that in the current adjustment task, listeners preferred to boost and cut wider frequency bands than the previously chosen octave bands in order to avoid introducing resonances. It is therefore likely that sharp resonances are a likely cause unexpected clarity reductions.

Factors outside the frequency content category were mentioned slightly less often: amplitude or loudness were elicited 3 times, envelope related factors were elicited 8 times, expectations and naturalness 9 times, intelligibility and “readability” 14 times, recording artefacts 46 times, reverb 14 times and hedonic responses, i.e. quality and pleasantness 40 times.

Among these factors, most are difficult to manipulate with EQ. For example, amplitude envelopes may impact on clarity but no correlation between attack time and clarity could be found for the data thus far collected in this thesis because the EQ could not alter amplitude factors significantly enough for a change in clarity. Loudness was matched prior to all experiments, hence it is unlikely that this could have had a great impact on clarity. Expectations and naturalness would be difficult to measure and are likely to be individually different (for instance, an Erhu player would have probably recognized the sound of the Erhu and perceived the strong resonance as natural). Recording artefacts are highly likely to influence clarity but through informal listening, these were less audible than unpleasant resonances in the previous vocal and bass experiment. The bass was a high quality sample from a sample library. Recording artefacts would be difficult to measure, as CASA would need to be employed in order to establish which parts of the sound belong to the instrument and which are unwanted. At the same time, it is likely a difficult task to remove artefacts like distortion with EQ. Reverberation (mentioned 14 times) does influence clarity, as established in the literature review (chapter 3) but is also unlikely to have been changed significantly through the EQ. Where there was reverberation in the sound, it appeared to span a wide range of frequencies. Hedonic responses, i.e. quality and pleasantness (40) may also differ between test subjects and are likely to be influenced by the lower level parameters listed above, therefore, those should be understood fully before assessing hedonic responses further.

Therefore, unpleasant peaks in the spectra of sounds appear to be the most useful factor of single sound clarity for further investigation. It would be useful to assess whether such peaks
lead to the reduction in clarity for the previously tested vocal and bass stimuli. This will be assessed in chapter 10.

8.4 Conclusion

In previous experiments, the harmonic centroid was established to be the most useful predictor for guitar, piano and string spectral clarity. However, when the spectral clarity of vocal and bass stimuli was assessed in chapter 7, spectral clarity was reduced for nearly all stimuli even when the harmonic centroid was significantly raised, hence this predictor no was no longer able to predict spectral clarity reliably. An exploratory listening test was carried out with the aim to answer the following three research questions:

1. What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ?

2. How can spectral clarity be defined in the context of EQ adjustments?

3. What additional factors are likely to be relevant for spectral clarity?

Listeners were asked to maximise the spectral clarity of programme items, using parametric EQ with a frequency, gain and bandwidth control. On each page, they were asked to write down what difficulties they experienced. At the end of the test, they were asked to provide their definitions of clarity. In this way, important clarity factors could be elicited.

Stimuli were chosen to vary in those factors that were deemed to be likely to influence clarity, namely unpleasant resonances, spectral balance, different combinations of high and low fundamentals and bandwidths, sounds that listeners were more or less likely to be familiar with and harmonic and inharmonic sounds. Most adjustment tasks were rated highly, hence listeners were able to significantly improve clarity with just one EQ control. The research questions can now be answered.

8.4.1 What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ?

In general, boosts were preferred to cuts. The favoured centre frequencies spanned 1kHz to 7kHz. Preferred gains were around -7dB for cuts and 9dB for boosts, which is similar to previously used gains. The Q settings differed considerably from those used in previous experiments. In general, much wider bandwidths were preferred. The preferred settings did not depend heavily on each other or on the stimuli.
8 Spectral single sound clarity adjustment task

8.4.2 What additional factors are likely to be relevant for spectral clarity?

For an unbiased analysis of the qualitative data, an external researcher was asked to establish a list of clarity factors from the clarity definitions and text box input.

Factors relating to amplitude, envelope, expectations, frequency content, “readability”/intelligibility, recording artefacts, reverb and hedonic responses were found in the definitions and text box entry. Presumably due to the nature of the task being an EQ task, the category that was mentioned most often was frequency content (22 instances). Many of the phrases within this category appear to be directly related to the harmonic centroid, namely bandwidth, brightness, frequency balance, the energy in specific frequency bands, warmth and thinness. It is therefore likely that the harmonic centroid is still useful in an overall clarity model.

Additional spectral factors were found. These appear to be related to clarity reducing lumps or 'holes' in the spectrum, i.e. cuts and boosts to specific frequencies, harmonic content, muddy and muffled sound and resonances. All factors and their overall number of instances were presented in Table 8-3. While all factors can be related to the literature, some were not explicitly mentioned in the sources consulted in chapter 3: the spectral factors that appear to be related to clarity reducing ‘lumps’ in the spectra of sounds (i.e. resonances and ringing), factors relating to expectations and naturalness, “Doppler effect” and pleasantness/hedonic responses.

8.4.3 How can spectral clarity be defined in the context of EQ adjustments?

Considering the definitions given by test subjects, in particular the factors that were mentioned most often, spectral clarity in this context could be defined as the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard. It appears that their audibility might be increased by gentle high frequency boosts and low frequency cuts (since most sounds have less energy in the upper parts of their spectra), and avoidance of potentially distracting unnatural/unpleasant resonances and other artefacts.
9 An improved spectral clarity predictor

The spectra of sounds play an important role in the context of perceived clarity. In previous experiments using piano, guitar and string stimuli, harmonic centroid change, as altered by 9dB octave band boosts and cuts, was found to correlate positively with clarity change. In chapter 7, the clarity of vocal and bass stimuli with the same EQ treatments was assessed. Surprisingly, almost all EQ treatments led to a reduction in the clarity of the target sounds, even when the harmonic centroid was increased. Despite this, a decrease in the linear, Mel, ERB, Cent and harmonic centroids still corresponded with a further decrease in clarity change, especially for bass programme items. At the same time, the further the centroids were raised, the more the negative impact on clarity diminished. It was concluded that harmonic numbers boosted and cut, as well as Mel, ERB, Cent, linear and harmonic centroids can be used to predict clarity change with limited accuracy. In order to predict clarity change more accurately, further factors that clarity could depend on in the context of EQ adjustments were established in a verbal elicitation experiment (chapter 8). It was established that in addition to the overall HF and LF balance, unpleasant peaks in the spectra of sounds, e.g. resonances or ringing, appear to be important. It was concluded that a measure of these unwanted, sharp peaks would be a useful addition to the harmonic centroid based clarity model. The current chapter aims to answer the research question: can the harmonic centroid based predictor of single sound clarity be improved by including a metric for potentially unpleasant peaks in the spectra?

In section 9.1, a computational model of mid-range spectral peakiness is presented and added to the clarity model. Section 9.2 is a conclusion.

9.1 A computational model of mid-range spectral peakiness

In section 9.1.1, a computational model of mid-range spectral peakiness is presented. In section 9.1.2, this is integrated into the existing harmonic centroid based clarity model.

9.1.1 Devising the model

In order to test whether potentially unpleasant peaks are visible in the LTAS spectra, piano, guitar, string, vocal and bass stimuli with significantly raised harmonic centroids (more than 1.5 harmonics) were split into two groups: those, where clarity was lowered significantly (group 1) and those, where clarity was increased significantly (group 2). The two groups were compared by inspecting long-term average spectra and by auditioning the stimuli once again. Examples for both groups are shown in Figs. 9-1 – 9-3 (group 1) and Figs. 9-4 – 9-6 (group 2).
9 an improved clarity predictor

Fig. 9-1: bass, 2000Hz boost. Clarity has decreased significantly despite a raised harmonic centroid (group 1)

Fig. 9-2: bass, 500Hz cut. Clarity has decreased significantly despite a raised harmonic centroid (group 1)
9 an improved clarity predictor

Fig. 9-3: vocal, 500Hz cut. Clarity has decreased significantly despite a raised harmonic centroid (group 1)

Fig. 9-4: bowed cello, 250Hz cut: the harmonic centroid and clarity both increased significantly (group 2)
Fig. 9-5: piano, 500Hz cut: the harmonic centroid and clarity both increased significantly (group 2)

Fig. 9-6: piano, 8kHz boost: the harmonic centroid and clarity both increased significantly (group 2)
It was observed that the stimuli in group 1 had a peak that became more pronounced between the fundamental and highest frequency, following a steep drop in amplitude from the fundamental up to the point just before the peak (mid-range spectral peakiness). This phenomenon is missing or less noticeable in group 2. This is likely perceived as a clarity reducing resonance in the sound.

By further analysing the long-term average spectra and by listening to the stimuli, it was noted that a dip only appeared to affect clarity negatively if it served to enhance a nearby peak. Peaks towards the middle of the spectra appeared to affect clarity the most: enhanced peaks close to the fundamental appeared to add fullness and peaks at very high frequencies appeared to make the sound more detailed or ‘airy’ rather than decreasing clarity. Following this observation, a computational model was devised with the ability to detect peaks in the centre of spectra. Peaks close to or below the fundamental or near high frequencies were paid less attention. The model consisted of the following steps:

1. The LTAS of each stimulus was calculated in dB with 1/6 octave band smoothing. Smoothing was used because the overall shape of the spectrum is more important than the fine detail and 1/6 octave band smoothing provided a useful balance between removing unnecessary fine detail while still keeping the overall shape of the spectrum.

2. The LTAS was interpolated to comprise equidistant points on a logarithmic scale (so that e.g. each octave band contained the same number of points)

3. A curve was fitted to the LTAS that lies below those peaks that are likely to affect clarity (peaks towards the centre of the spectra) but above all other frequency areas. By investigating the portion of the LTAS above the curve, changes in mid-range spectral peakiness could be measured. The curve was fitted to the LTAS as follows.
   - The curve fits closely at low frequencies, up to the fundamental frequency.
   - From the fundamental frequency onwards, the curve fits loosely when falling (this means that most LTAS data points now lie below the curve).
   - With increasing frequency, the curve fits better when the LTAS is falling.
   - At the same time, with increasing frequency, the curve fits more loosely when the LTAS is rising. This means that towards the middle of the spectrum, most data points belonging to peaks lie above the curve.
   - The curve fits closely again at high frequencies.

4. The distances of data points above the curve, from the curve, are summed together. A higher weighting is applied to data points that lie further away from the curve.
5. The sum for the reference stimulus is deducted from the sum for each equalized stimulus, in addition to an “accept value”: a small degree of resonance was perceived as unobjectionable. Hence, when the equalized stimulus has notably higher peaks towards the middle of its spectrum than the reference, the measure has a positive value. Otherwise, the value is negative.

More detail can be found in Appendix A3. Fig. 9-7 shows an example of the curve fitted to the vocal 125Hz cut stimulus.

![Graph](image)

**Fig. 9-7:** The curve for the mid-range spectral peakiness metric is fitted to the vocal 125Hz cut stimulus LTAS (increased peakiness). It fits closely below the fundamental frequency, then floats above the LTAS for frequencies slightly higher than the fundamental frequency, in order to ignore any peaks in this area. In the upper mid-range, the curve lies below all peaks. Therefore, these peaks contribute to the mid-range spectral peakiness metric. At very high frequencies, the curve fits closely to the LTAS again.

The parameters of the metric (e.g. the degree of fit along various points of the curve and the accept value) were optimized such that most stimuli in the two groups were correctly identified (group 1 = positive value, group 2 = negative value). All group 1 and group 2 stimuli were successfully differentiated as a result, with the exception of the 2kHz cut on guitar, where a drop in clarity was not detected.

Additionally, it was attempted to also correctly identify stimuli where centroid and rating moved in opposite directions by less significant amounts. Hence, stimuli where the HC was
raised, while clarity did not change significantly (or changed just slightly negatively) were investigated. It was not always possible to detect these stimuli with the metric but it has previously been established that the HC needs to be raised above a certain limit before clarity increases (about 1.5 harmonics, chapter 6), and this could explain the opposing movement of centroid and clarity in most of these cases. Stimuli where the harmonic centroid was lowered and clarity did not change significantly did not show an increase in mid-range spectral peakiness. In this way, the model was able to provide information that was accurate enough to explain the observations in expt 5. Table 9-1 shows the direction of clarity change and harmonic centroid for all stimuli, as well as predicted change in mid-range spectral peakiness.

<table>
<thead>
<tr>
<th>Piano</th>
<th>Guitar</th>
<th>Bowed Violin</th>
<th>Plucked Violin</th>
<th>Bowed Cello</th>
<th>Plucked Cello</th>
<th>Vocal</th>
<th>Bass</th>
</tr>
</thead>
<tbody>
<tr>
<td>62Hz boost</td>
<td>SN, SN, -</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, SN, -</td>
<td>N/A</td>
</tr>
<tr>
<td>62Hz cut</td>
<td>SP, SP, +</td>
<td>P, SP, +</td>
<td>O, P, +</td>
<td>SP, P, -</td>
<td>SP, SP, +</td>
<td>P, SP, +</td>
<td>SP, SP, +</td>
</tr>
<tr>
<td>125Hz boost</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>SN, SN, -</td>
<td>SN, P, -</td>
<td>SN, SN, -</td>
<td>SN, SN, -</td>
</tr>
<tr>
<td>125Hz cut</td>
<td>SP, SP, +</td>
<td>P, SP, -</td>
<td>P, P, +</td>
<td>P, P, -</td>
<td>P, SP, -</td>
<td>N, N, -</td>
<td>SN, SP, +</td>
</tr>
<tr>
<td>250Hz boost</td>
<td>SN, P, -</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>N, P, -</td>
<td>SN, SN, -</td>
<td>N, N, -</td>
</tr>
<tr>
<td>250Hz cut</td>
<td>SP, SP, +</td>
<td>P, SP, -</td>
<td>P, P, +</td>
<td>P, SP, -</td>
<td>N, N, -</td>
<td>N, N, -</td>
<td>SN, SP, +</td>
</tr>
<tr>
<td>500Hz boost</td>
<td>SN, P, -</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>N, P, -</td>
<td>SN, SN, -</td>
<td>SN, N, +</td>
</tr>
<tr>
<td>500Hz cut</td>
<td>SP, SP, +</td>
<td>P, SP, -</td>
<td>P, P, +</td>
<td>P, SP, -</td>
<td>N, SP, -</td>
<td>N, N, -</td>
<td>SN, SP, +</td>
</tr>
<tr>
<td>1kHz boost</td>
<td>O, N, -</td>
<td>N, SP, -</td>
<td>O, SP, +</td>
<td>P, P, -</td>
<td>N, SP, +</td>
<td>N, N, -</td>
<td>N, N, -</td>
</tr>
<tr>
<td>1kHz cut</td>
<td>N, SP, +</td>
<td>N, SP, -</td>
<td>O, P, +</td>
<td>N, P, -</td>
<td>P, P, +</td>
<td>SN, N, -</td>
<td></td>
</tr>
<tr>
<td>2kHz boost</td>
<td>SN, N, -</td>
<td>SN, SP, +</td>
<td>N, P, +</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>N, N, -</td>
<td>SN, SP, +</td>
</tr>
<tr>
<td>2kHz cut</td>
<td>SP, SP, +</td>
<td>N, SP, -</td>
<td>SN, P, +</td>
<td>P, P, +</td>
<td>P, SP, -</td>
<td>P, P, +</td>
<td>0, SP, +</td>
</tr>
<tr>
<td>4kHz boost</td>
<td>SP, SP, +</td>
<td>N, SP, -</td>
<td>SN, P, +</td>
<td>P, P, +</td>
<td>P, SP, -</td>
<td>P, P, +</td>
<td>0, SP, +</td>
</tr>
<tr>
<td>4kHz cut</td>
<td>SN, N, -</td>
<td>SN, SN, -</td>
<td>N, N, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td></td>
</tr>
<tr>
<td>8kHz boost</td>
<td>SP, SP, +</td>
<td>P, SP, -</td>
<td>O, P, +</td>
<td>SP, P, -</td>
<td>SP, SP, +</td>
<td>O, P, +</td>
<td>N, SP, +</td>
</tr>
<tr>
<td>8kHz cut</td>
<td>SN, SN, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>SN, N, -</td>
<td>N, N, -</td>
<td>SN, N, -</td>
<td>0, SN, -</td>
</tr>
</tbody>
</table>

Table 9-1: for all stimuli in experiments 2, 3 and 5, changes in clarity, harmonic centroid and the prediction by the mid-range spectral peakiness model are shown in that order. ‘SP’ and ‘SN’ mark significant positive and negative changes, i.e. confidence intervals were completely above or below 0 for clarity ratings, and the harmonic centroid was either lowered or raised by more than 1.5 harmonics. ‘N’ and ‘P’ indicate smaller changes, 0 indicates ‘no change in clarity’. Group 1 is marked in yellow and group 2 is marked in green (significant changes in both HC and clarity).
9.1.2 Incorporating mid-range spectral peakiness into the clarity model

In the last section, a predictor for mid-range spectral peakiness was presented. Most of the stimuli that are predicted to be peaky (mainly vocal stimuli) have clarity ratings with overlapping confidence intervals (chapter 7). Tailoring a model closely to data exhibiting non-significant differences would likely result in an over-fitted model that lacks robustness and generalizability. The aim here is therefore not to develop a universal model for sound clarity in mixes but rather to show that the concept of EQ-related clarity changes can be related to harmonic centroid changes and mid-range spectral peakiness.

The peaky vocal stimuli are all rated at around -15 (chapter 7). Due to this and the limitations of the predictor, the degree of change predicted in mid-range spectral peakiness currently does not correlate with the degree of clarity change. Therefore, it may be useful to implement a gate in the overall clarity model that sets clarity ratings for all peaky stimuli to a fixed, negative clarity value.

The extent to which mid-range spectral peakiness was larger or smaller than in the reference was ignored, instead only the overall direction of change (positive of negative) was considered.

It is proposed that a linear regression model is fitted to all data excluding the stimuli that are predicted to sound peaky, while a fixed value is used for the latter. The proposed model works as follows, where $c$ is predicted spectral clarity, $h$ is the shift in harmonic centroid, and $a$ and $b$ are calculated from the data:

\[
c = a + (b \cdot h)
\]  

\[ (9.1) \]
In order to test whether the model can predict clarity more reliably than the previously tested metrics, it is fitted separately to the stimuli from each experiment. In each experiment, listeners were presented with a familiarization page featuring all stimuli of that experiment, and ratings for the different programme items were undertaken within the same sitting. Hence, it is likely that listeners used the same scale each for piano and guitar stimuli, for string stimuli, and the vocal and guitar, while ratings between experiments were completely independent from each other. Therefore, it would be useful to test the new metric separately for the experiments.

The new metric will be compared to a second metric based solely on a HC linear regression model (equation 9.1), ignoring mid-range spectral peakiness. The aim is not to devise a generally applicable model of clarity, but rather, to test whether clarity can be measured more reliably when both mid-range spectral peakiness and HC are considered. Table 9-2 shows the values for coefficients a and b for the HC-only and combined models, for all three experiments. In the experiment in chapter 5 (piano and guitar), mid-range spectral peakiness never increases and therefore, the model is identical for both.

<table>
<thead>
<tr>
<th></th>
<th>Combined Model</th>
<th>HC alone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Piano/guitar</td>
<td>-7.748</td>
<td>4.986</td>
</tr>
<tr>
<td>Strings</td>
<td>-7.094</td>
<td>7.163</td>
</tr>
<tr>
<td>Vocal/bass</td>
<td>-8.790</td>
<td>2.595</td>
</tr>
</tbody>
</table>

Table 9-2: linear regression model coefficients for HC/ mid-range spectral peakiness and HC only models

In order to assess whether the combined model is better suited to the measurement of clarity than the HC-only model, correlation coefficients and root mean square errors were calculated for both, for all experiments (Table 9-3). As an additional measure, a Monte Carlo method was applied, as suggested by Pearce [2016]. This considers the fact that datasets with overlapping confidence intervals are not significantly different, and that therefore, the rank ordering of means belonging to the confidence intervals may not always be useful in assessing the fit of a given clarity predictor. Throughout this thesis, correlations between the medians (not means) of clarity ratings and clarity predictors were investigated, as the data were not quite normally distributed in experiments. However, the Monte Carlo method requires the use of means. For all other tests, median clarity ratings were used.

1000 simulated datasets of clarity ratings were generated, based on the original data set. For each stimulus in each dataset, clarity is defined as a random value that lies within the standard
deviation of the original data set. The value is determined by multiplying the standard deviation by a random number, chosen from a Gaussian distribution with a mean of 0 and a width of 1, and by adding the result to the original mean clarity rating. In this way, the rank order of clarity ratings will vary between datasets when the original confidence intervals overlap.

By calculating Spearman’s Rho for each new dataset, for both predictors, it is possible to compare predictor fit in a way that considers stimuli with very similar clarity ratings. As clarity ratings are shuffled around, one dataset will fit the predictor the least well, and will have the smallest Rho value. The best metric should have the greatest min Rho value, as this means that the worst possible fit between this metric and a possible arrangement of mean values around the confidence intervals is still better than that for other metrics. Most correlation coefficients between the models and median clarity ratings exceed the minimum acceptable value of 0.670 (chapter 6). The smallest value (0.631) is not much smaller than this.

<table>
<thead>
<tr>
<th>Piano/guitar</th>
<th>Spearman ϱ</th>
<th>Spearman p</th>
<th>Monte Carlo Min</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Model</td>
<td>0.775</td>
<td>&lt; 0.000</td>
<td>0.315</td>
<td>13.876</td>
</tr>
<tr>
<td>HC only</td>
<td>0.775</td>
<td>&lt; 0.000</td>
<td>0.315</td>
<td>13.876</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strings</th>
<th>Spearman ϱ</th>
<th>Spearman p</th>
<th>Monte Carlo Min</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Model</td>
<td>0.848</td>
<td>&lt; 0.000</td>
<td>0.476</td>
<td>7.258</td>
</tr>
<tr>
<td>HC only</td>
<td>0.818</td>
<td>&lt; 0.000</td>
<td>0.472</td>
<td>7.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bass and vocal</th>
<th>Spearman ϱ</th>
<th>Spearman p</th>
<th>Monte Carlo Min</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Model</td>
<td>0.631</td>
<td>&lt; 0.000</td>
<td>0.485</td>
<td>4.947</td>
</tr>
<tr>
<td>HC only</td>
<td>0.525</td>
<td>0.002</td>
<td>0.351</td>
<td>5.934</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average</th>
<th>Spearman ϱ</th>
<th>Spearman p</th>
<th>Monte Carlo Min</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Model</td>
<td>0.751</td>
<td>&lt; 0.000</td>
<td>0.425</td>
<td>8.694</td>
</tr>
<tr>
<td>HC only</td>
<td>0.706</td>
<td>&lt; 0.000</td>
<td>0.379</td>
<td>8.937</td>
</tr>
</tbody>
</table>

Table 9-3: correlation coefficients, worst case Monte Carlo Spearman’s Rho and root mean square error for the mid-range spectral peakiness/HC and HC only linear regression models

As can be seen in the table, the combined model performs considerably better than the HC only model for vocals and bass, with improvements in correlation coefficients of up to 0.175. The RMSE is also reduced by one data point. For the strings experiment, the RMSE is slightly worse for the combined model but by an insignificant amount. This may be due to the fact that few stimuli were predicted to sound peaky (in the bowed violin only). On average, the new model performs much better than the model based only on the harmonic centroid. The correlation between the combined model and clarity for experiments 3 and 5 is shown in Fig. 9-8 and Fig. 9-9.
Overall, it can be concluded that clarity can be measured more reliably when unnatural sounding resonances caused by equalisation are taken into consideration.
9.2 Conclusion

The harmonic centroid was previously established to be a suitable clarity predictor for string, guitar and piano stimuli but it failed to explain why clarity was almost always reduced for the vocal and bass stimuli in chapter 7. Through a verbal elicitation task, it was established that it would be useful to consider a metric for mid-range spectral peakiness in addition to the harmonic centroid. The purpose of the current chapter was to answer the research question: can the harmonic centroid based predictor of single sound clarity be improved by including a metric for potentially unpleasant peaks in the spectra?

The qualitative data collected in the previous adjustment task was investigated and long-term average spectra of stimuli with significantly increased harmonic centroids, but significantly reduced clarity were inspected and compared to those spectra of significantly clearer stimuli. It was found that mid-range spectral peakiness (caused by unnatural sounding resonances) manifested as enhanced sharp peaks situated towards the middle of stimulus long-term average spectra. An algorithm was devised to estimate mid-range spectral peakiness.

A new clarity model was devised, whereby clarity is set to a fixed, negative value when mid-range spectral peakiness (as indicated by the algorithm) has increased. A linear regression model was fit to the remaining data, using the harmonic centroid as predictor variable for clarity. The combined model correlated better with clarity than the harmonic centroid alone and the root mean square error was reduced. All in all, by measuring changes in the harmonic centroid, as well as mid-range spectral peakiness, clarity can be predicted well.
10 Single sound spectral clarity in mixes

The clarity and separation of sounds (the extent to which individual components can be heard in a mix) is a key perceptual attribute of high quality music mixes (chapter 2). The spectra of sounds play a particularly important role here. In previous experiments, the harmonic centroid (spectral centroid divided by a sound’s average fundamental) and the direction of change in mid-range spectral peakiness (the presence of potentially unpleasant-sounding peaks in the middle of the spectrum) were established to be useful predictors of spectral clarity.

It is possible that the metric established for isolated sounds still plays an important role in mixes but additional factors are likely to play a role, as target audibility is likely to decrease due to fusion and masking between target and backing track (the mix excluding the target). Not only spectral clarity, but also separation from other instruments could be important here. Spectral clarity of sounds in a mix is the same percept as the spectral clarity assessed in previous experiments. Separation of target sounds in mixes is the extent to which a target sound can be distinguished from other sounds in a mix (which may be possible even when the target has low spectral clarity). In order to establish exactly how spectral clarity in isolation differs from target spectral clarity and separation in a mix, it would be useful to place some of the stimuli previously tested in isolation in a mix and to equalize either the target, or the backing track. Listeners can then be asked to rate clarity and separation similarly to before.

The aim of this chapter is to test and improve the previous clarity metric for sounds in mixes. The following research question will be answered: to what extent can the previous single-sound spectral clarity model predict the degree of target spectral clarity and separation change resulting from applying simple EQ to sounds in a mix? Since clarity will only be affected by changes in EQ, “clarity” and “spectral clarity” are used synonymously in this chapter.

In section 10.1, the setup for a new listening test is presented. In section 10.2, the results are analysed. Section 10.3 is a discussion and section 10.4 is a conclusion.

10.1 The setup for the listening test

Eighteen listeners were asked to compare the clarity and separation of target sounds inside mix stimuli. By adjusting sliders, they expressed in which stimulus the target was clearer and in which stimulus the target was more separated and by how much. The stimuli were equalised versions of the same mix, featuring boosts and cuts in either the target sound or all instruments excluding that target sound (the ‘backing track’). The target in each of these versions was compared with the target in an unequalized reference stimulus. Like in previous experiments,
warmth, fullness and brightness were rated occasionally in order to reduce the noise in the clarity ratings. Separation from other instruments was rated in addition to clarity, since this may be related to clarity, as established in chapter 2. This relationship is tested in section 10.2.2. If clarity and separation differ, this might further reduce the statistical noise in the ratings. The choice of the previous vocal and bass programme items as target sounds is explained in section 10.1.1.

Each test page contained two stimuli (one being the unequalised reference mix) that could be triggered by pressing buttons labelled ‘A’ and ‘B’ as often as required after pressing ‘play’. A ‘stop’ button could be used to reset the stimuli to their beginning position. The target sounds in the two stimuli were compared in terms of the perceptual attribute given at the top of each page. Clarity, separation, fullness, warmth or brightness ratings were undertaken by positioning a slider, allowing listeners to express which stimulus was e.g. clearer and by how much. Fig. 10-1 shows an example of a test page.

The vocal and bass ratings were undertaken in two separate test halves (vocal first), lasting approximately fifteen minutes each. Four pages were added to the vocal half in order to include the additional perceptual attributes. The other attributes were still shown in the bass half, despite not being tested, as both previous and simultaneous exposure to rating other attributes can reduce the noise in clarity ratings (chapter 4). In this way, both test halves did not exceed 32 pages. The entire listening test was repeated for each subject in a separate session in order to test listener consistencies.
Each listener was given an instruction sheet before undertaking the listening test (Appendix, Fig. A-7). The test was preceded by a familiarization interface (Fig. 10-2), allowing each subject to audition all stimuli. Listeners were also given the opportunity to try out the test interface before starting the test. The test interface and incorporated stimulus pairs were presented at a comfortable listening level in a treated, quiet listening room, using a VRM box interface as headphone amp (the virtual reference monitoring was disabled) and Sennheiser HD 600 headphones, similarly to previous experiments. In this section, the setup used in the current listening test will be summarized. In section 10.1.1, stimulus generation is explained and in section 10.1.2, the listening panel is introduced.

![Image](image-url)

**Fig. 10-2:** the familiarization interface for the bass test half, featuring either all bass stimuli or all vocal stimuli

### 10.1.1 Stimuli

The following sections justify the use of bass and vocal programme items, explain the added equalisation and outline the production of the stimuli.

**Programme items**

As mentioned in the last section, the previously tested male vocal and electric bass programme items were used as target sounds, for the following reasons. Firstly, using the previously tested programme items makes it possible to test exactly how spectral clarity for sounds in isolation differs from spectral clarity in the mix. Secondly, it is useful to use stimuli that feature changes both in harmonic centroid and mid-range spectral peakiness (both increases and decreases), as in this way, all previous findings for spectral clarity for single sounds in isolation can be
Single sound spectral clarity in mixes

compared to spectral clarity and separation in mixes. Both vocal and bass stimuli featured
stimuli with variation in harmonic centroid and mid-range spectral peakiness. As it is more
difficult to establish the harmonic centroid of inharmonic sounds, the use of harmonic sounds is
helpful, which is the case for the vocal and bass. It is also possible that the choice of familiar
timbres would help subjects make clarity judgements. This may have been more difficult if
synthesized timbres were used, as shown in chapter 8.

For the backing track, a piece was composed using a mixture of acoustic and electronic
instruments and the target sounds. The piece was assembled by using Logic 9 Apple Loops and
additional software synthesiser programming. Sounds were arranged around the male vocal
and bass to create partial masking and fusion effects. The EQ treatments (explained below) led
to increases and decreases in fusion and masking in different frequency areas. No elements
were hard panned to create as much overlap as possible, as spatial separation can reduce
masking [Moore, 2012]. The low end of the bass was left unmasked. In this way, it was possible
to test whether boosts in this area may contribute to spectral clarity positively, even when this
meant a decrease in bass harmonic centroid. This possibility was considered as basses are often
used to add low end to mixes and therefore, bass spectral clarity may depend on the audibility
of this frequency area. Similarly, Bregman [2008] states that boosts around the fundamental
may strengthen the sense of pitch in a sound in mixes which may be related to spectral clarity
and differ from spectral clarity measured in isolation.

Added equalisation

Altogether, thirty-two mixes were created for the bass and twenty-eight for the vocal. For
consistency, a single 9dB boost or cut was added to one of the eight octave bands centred at
62Hz (bass only), 125Hz, 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz respectively, with a Q value
of 1.41 (-3dB bandwidth = 1 octave). For each stimulus, the EQ treatments were applied to the
vocal, bass or all sounds excluding either vocal or bass.

Production of the stimuli

The stimuli were exported as 44.1kHz and 24 bit wave files after being processed with Logic Pro
9’s parametric equaliser and loudness-matched by four audio professionals: here, the equalised
targets and backing tracks were loudness matched to their respective unequalised references
before assembling the mix. If the sounds had been loudness matched inside the mix, it is
possible that subjects would have adjusted them to be as clear and separated as in the
reference, as it is likely that clarity and separation are related to loudness. If the loudness of the
overall mix was matched to the references, it is likely that the target sounds therein would have
varied in loudness, again influencing spectral clarity.
10.1.2 Type of listeners

Like in previous experiments, experienced listeners were used: the chosen listening panel comprised 18 male and female undergraduate students and postgraduate researchers of the University of Surrey's Institute of Sound Recording. The participants were aged between nineteen and twenty-seven years. All subjects were experienced in listening tests and in verbalising sensations of timbre. Additionally, the undergraduate students had extensive technical ear training, which included the blind identification of EQ changes. None of the participants reported having any hearing damage. Listeners sat the test individually.

10.2 Presentation of the results

In the following, the results of the listening test will be presented. Section 10.2.1 shows that the prerequisites for using parametric methods of statistics are not quite fulfilled. In section 10.2.2, clarity and separation data are tested for correlation. In section 10.2.3, the previous combined harmonic centroid and mid-range spectral peakiness metric is tested on the stimuli where the target was equalized, ignoring the influence of the backing track. In section 10.2.4, the metric is calculated for the audible, unmasked parts of the target sounds, this time incorporating the rest of the mix. In section 10.2.5, overall target-to-backing-track ratio is investigated. In section 10.2.6, the relationship between target and backing track spectra is investigated by analysing box plots and line plots. This leads to a target peak audibility metric, which is presented in sections 10.2.7 (peak audibility). In section 10.2.8, the influence of overall mix spectral centroids on target spectral clarity is assessed. In sections 10.2.9 and 10.2.10, peak audibility is assessed further.

10.2.1 Prerequisites for using parametric methods

Before deciding whether parametric statistical methods could be used, the conditions specified in previous experiment reports were tested (interval data, independence, normal distribution and homoscedasticity). The data gathered in the experiment were interval data. Like before, the fact that listeners most likely compared ratings to each other means that independence could not be guaranteed.

In order to assess whether the data were normally distributed, histograms were analysed for all stimulus pairs. For datasets smaller than 2000 elements, it is suggested that the Shapiro-Wilk test is used, otherwise, the Kolmogorov-Smirnov test should be used [Field, 2013]. The current dataset comprises 2160 values each for clarity and separation, hence the Kolmogorov-Smirnov test bears greater significance. It was concluded that the data were mostly normally distributed, but with some exceptions: for clarity, 16 of the 60 significance values were below 0.05 in the
10 Single sound spectral clarity in mixes

Kolmogorov-Smirnov test. Some of the plots indicated a roughly normal distribution of the data, while others did not (Fig. 10-3).

![Histogram for StimPair = 2.00](image1)

![Histogram for StimPair = 19.00](image2)

Fig. 10-3: two examples of the histograms used to assess normality of the clarity data. The clarity ratings are approximately normally distributed for stimulus pair 4 but slightly less for stimulus pair 19.

As a next step, it was assessed whether any listeners were considerably less consistent than the rest, and may therefore have contributed to the uneven distribution of the data. Each listener undertook each rating twice. The notched box plot in Fig. 10-4 shows the distribution of differences between the ratings for all stimulus pairs and for each listener. There is no clear cut-off point between consistent and inconsistent listeners, as consistencies are quite varied overall. Hence, it would be arbitrary to remove these listeners. All data are therefore retained for subsequent analysis.
10 Single sound spectral clarity in mixes

All in all, the data were not quite normally distributed with some exceptions. Levene’s test shows that the variance of the data is not homogenous, with a significance of less than 0.001. This was calculated separately for vocal and bass stimulus pairs because the Levene’s test cannot be calculated for more than fifty groups in SPSS. As the prerequisites for using parametric statistical methods are not fulfilled, mostly non-parametric methods are used.

10.2.2 Comparing clarity and separation

As mentioned in section 10.2.2, separation data were collected for each stimulus pair. Clarity and separation overall were defined as the extent to which individual components can be heard in a mix (chapter 2) and it is possible that both attributes correlate. On the other hand, clarity may for example relate more to timbral factors, while separation may relate solely to masking, as informal conversations with the listeners reveal. In order to find out if they are the same thing, all ratings were compared in a scatter plot (Fig. 10-5). This clearly indicates that clarity and separation differ. Similarly, the paired t-test shows that clarity and separation are significantly different: $t(2159) = -14.701, p < 0.001$.

As previous listening tests have focused on clarity only, and as separation appears to be different from this, separation will be ignored in the following analysis. As the two attributes do not correlate, it is unlikely that separation can provide more information about clarity. Once the factors influencing clarity have been defined, separation can be investigated separately in future studies.

![Fig. 10-4: distribution of absolute differences between the two ratings for all stimulus pairs, plotted individually for each listener.](image-url)
10.2.3 Target harmonic centroids and mid-range spectral peakiness

A combination between mid-range spectral peakiness and the harmonic centroid was useful in predicting spectral clarity for single sounds in isolation (chapter 9). It is possible that this is still the case when sounds are placed in mixes, although it is likely to be only partly effective since it ignores the fusion and masking effects caused by the presence of the backing track.

As a first step, it would be useful to test the previous metric on the clarity data gathered in the current experiment. As the masking and fusion caused by the backing track is ignored for now, the method can only applied to the half of the stimuli where the target was equalized. At first, the harmonic centroid of the target and spectral clarity will be tested for correlation. Subsequently, a combination between the harmonic centroid and mid-range spectral peakiness will be related to spectral clarity.

Target harmonic centroids are plotted against spectral clarity in Fig. 10-6. The Spearman correlation coefficient is 0.344 with a p-value of 0.06. The degree of correlation was low when vocal and bass were tested in isolation (0.567) but in the presence of the backing track, the correlation between target harmonic centroids and spectral clarity is even smaller. Therefore, the harmonic centroid alone appears to be an unsuitable predictor of target spectral clarity in mixes.

Next, the harmonic centroid was combined with the mid-range spectral peakiness metric described in chapter 9. In order to test the validity of this metric, it was necessary to fit a linear regression model to the data in the same way as in chapter 9: in this way, a meaningful constant
Single sound spectral clarity in mixes

value could be assigned to all predicted spectral clarity values corresponding to stimuli with increased mid-range spectral peakiness. The correlation between the combined predictor and target spectral clarity was calculated, again ignoring the influence of the backing track. The correlation coefficient is 0.437 ($p = 0.02$), which is notably higher than for the metric that only considers the harmonic centroid. The relationship is shown in Fig. 10-7.

Overall, it appears that a metric that considers both mid-range spectral peakiness and harmonic centroid works better than the harmonic centroid alone for the spectral clarity of target sounds in mixes. This is in line with the findings for spectral clarity for sounds in isolation (chapter 9), although the correlation there was better (0.724). Therefore, it seems as though the metric might be a useful starting point but could be improved further by taking masking effects caused by the backing track into consideration.

Fig. 10-6: harmonic centroid changes plotted against clarity ratings for stimuli where the target was equalized (masking and fusion caused by the backing track are ignored).
Single sound spectral clarity in mixes

Fig. 10-7: clarity prediction based on harmonic centroid and mid-range spectral peakiness changes plotted against clarity ratings (for stimuli where the target was equalized, masking and fusion caused by the backing track are ignored). Clarity = -4.918 + 1.095*HC, peaky stimuli set to -4.918.

In order to assess the influence of the backing track on perceived target spectral clarity, the clarity ratings of each target played in isolation (chapter 7) were compared to the same targets placed in the mix (Fig. 10-8—Fig. 10-11) and it becomes evident that in many cases, spectral clarity works differently in isolation than in the mix. In the next section, the harmonic centroid and target mid-range spectral peakiness metric is calculated for the audible, unmasked parts of the target.
Fig. 10-8: the influence of bass boosts on clarity, measured across 8 octave bands

Fig. 10-9: the influence of bass cuts on clarity, measured across 8 octave bands
10 Single sound spectral clarity in mixes

![Graph showing the influence of vocal boosts on clarity, measured across 7 octave bands](image1)

**Fig. 10-10:** The influence of vocal boosts on clarity, measured across 7 octave bands

![Graph showing the influence of vocal cuts on clarity, measured across 7 octave bands](image2)

**Fig. 10-11:** The influence of vocal cuts on clarity, measured across 7 octave bands

10.2.4 Audible target harmonic centroids and mid-range spectral peakiness

The backing track that the target sounds were placed in was specifically designed to cause masking, in order to test the influence of masking on target spectral clarity. As a result, some EQ changes were less audible than others and the effect of the EQ in each stimulus sounded
different than when the targets were played in isolation. It would be useful to extract those parts of the target sounds that were still audible in the presence of the backing track (unmasked) and to calculate the previous mid-range spectral peakiness and harmonic centroid metric for these unmasked parts only. It is possible that the extent to which the audible target areas become less peaky and have more HF and less LF content correlates better with spectral clarity than the metric that ignores the backing track. An important advantage of this new metric would be the fact that it can also be calculated for those stimuli where the backing track was equalized, as the changing spectrum of the backing track led to differences in the audibility of different target frequency areas, and hence changes in “audible harmonic centroids” and “audible mid-range spectral peakiness”. First, a method for extracting the unmasked parts of the spectrum is presented. Next, the impact of differences in audible harmonic centroids on spectral clarity is investigated and mid-range spectral peakiness is considered.

Extracting unmasked areas of the target sound
The audible areas of the target in each stimulus were established by calculating the target-to-backing-track ratio in 100 ERB-spaced frequency bands, out of which those with a target-to-backing-track ratio smaller than 1 were removed from the target. This approach was based on the recommendations of Wang [2005]. The harmonic centroids and mid-range spectral peakiness of the remaining areas in the target were calculated. This is explained in more detail below.

Step 1: targets and backing tracks were processed with an auditory model

For each stimulus pair, both the target sound and the backing track were processed with an auditory model. Firstly, A-weighting was applied in order to cater for the auditory system’s uneven sensitivity to different frequency regions.

The amount of energy in 100 ERB-spaced bands between 0kHz and 22.050kHz was established for each sample, using a Gammatone filterbank. In order to allow for a more reliable calculation of the target-to-backing-track ratio, the Hilbert transform was used in order to extract the overall signal envelopes in each ERB band. The Hilbert transform was implemented as follows:

- The output of each gammatone filter was transformed to the frequency domain via FFT. Half of the signal information in the frequency domain was removed by applying a step function to the FFT: all samples up to the Nyquist frequency were multiplied by 2 and the samples above this bin were multiplied by 0. In this way, the original signal amplitude could be maintained.

- The signal was transformed back into the time domain, leading to the analytic signal.
- The complex numbers in each sample of the analytic signal were treated as vectors with values for the real and imaginary parts. The length of the vector in each sample was used for a new signal, resulting in the overall signal envelope.

In order to take temporal masking into consideration, time smoothing was applied to both target and interferer Hilbert envelopes: a smoothing function was created and convolved with the target and interferer signals in each ERB band. The smoothing function was defined as

\[
s(n) = \begin{cases} 
    e^{-0.05 \frac{n}{f_s}} & t < 0 \\
    1 & t = 0 \\
    e^{0.01 \frac{n}{f_s}} & t > 0 
\end{cases} \quad (10.1)
\]

where \( t = \frac{n}{f_s} \) and \( t \) is a vector ranging from -0.05s, to 0.2s, in samples \( n \) at 44.1kHz (\( f_s \) is the sampling frequency). Subsequently, the resulting vector \( s \) was divided by the sum of all its elements in order to maintain the same original overall energy in the signal \( (s_{norm}) \). After convolving the target \( x_t(c,n) \) and interferer signals \( x_i(c,n) \) with the smoothing function in each frequency band \( c \),

\[
y_t(c,n) = s_{norm}(n) \ast x_t(c,n) \quad (10.2)
\]

\[
y_i(c,n) = s_{norm}(n) \ast x_i(c,n) \quad (10.3)
\]

the result was truncated to be as long as the original signal (4 s).

**Step 2: target to interferer ratio mask**

Next, a time-frequency target-to-backing-track ratio mask \( m_r(c,n) \) was created, containing the ratio between the target and backing track in each of the 100 ERB-spaced frequency bands and each sample, based on the recommendations of Francombe et al. [2013]. The ratio was defined as

\[
m_r(c,n) = 20 \cdot \log_{10} \left( \frac{y_t(c,n)}{y_i(c,n)} \right) \quad (10.4)
\]

for each sample and frequency band. The mask was then replaced by a binary mask \( m_b(c,n) \) as follows.

\[
m_b(c,n) = \begin{cases} 
    1 & m_r(c,n) \geq 0 \\
    0 & \text{otherwise}
\end{cases} \quad (10.5)
\]
10 Single sound spectral clarity in mixes

Step 3: removing the masked parts of the target for subsequent centroid calculation

The output of the auditory model for the target $\mathbf{y}_t(c, n)$ was multiplied element-wise with the binary mask $\mathbf{m}_b(c, n)$, removing all masked parts.

$$\mathbf{a}(c, n) = \mathbf{m}_b(c, n) \cdot \mathbf{y}_t(c, n)$$  \hfill (10.6)

In the resulting matrix $\mathbf{a}(c, n)$, all ERB spaced values were summed up for each sample $(n)$, resulting in a new target signal vector $\mathbf{a}(n)$.

$$\mathbf{a}(n) = \sum_c \mathbf{a}(c, n)$$  \hfill (10.7)

This was used to calculate the audible harmonic centroid and mid-range spectral peakiness like in previous experiments.

Audible harmonic centroids

The impact of differences in audible harmonic centroids, induced by equalisation, is shown in Fig. 10-12 and Fig. 10-13, separately for stimuli where the backing track; or target; was equalized. Generally, raising the audible harmonic centroid or the harmonic centroid did not increase spectral clarity for the targets. Depending on whether the target or backing track were equalised, the results appear to differ slightly. For stimuli in which the backing track was equalised, there seems to be a slight negative correlation between AHCs and spectral clarity (mainly for the vocal, $\rho = -0.508, p < 0.001$). However, this correlation is not visible in the stimuli where the backing track was equalized ($\rho = 0.328, p = 0.08$). There is no overall correlation as a result ($\rho = -0.0423, p = 0.75$). Therefore, this metric does not seem useful for predicting the spectral clarity for target sounds in mixes.
In order to assess whether the inclusion of the mid-range spectral peakiness metric might improve the above correlation, mid-range spectral peakiness was calculated for the audible
parts of all stimuli. In about half of the stimuli, mid-range spectral peakiness and clarity moved in the same direction and in the other half, they moved in opposite directions. The combination between mid-range spectral peakiness and harmonic centroids also did not seem to predict clarity well: when both were increased, the direction of clarity change did not follow a predictable pattern and the same was the case when both were decreased, or when both moved in opposite directions. In previous experiments, stimuli with increased mid-range spectral peakiness were always rated less clear than the reference. However, the direction of clarity change of the “audibly peaky” stimuli in the current experiment did not follow a predictable pattern. At the same time, many stimuli with increased harmonic centroid but decreased clarity were not recognized as peaky.

It is possible that simply removing masked parts of the target is not a fair representation of how the target is perceived. Humans have the ability to hear missing fundamentals and at the same time, listeners are familiar with bass and vocal sounds and therefore most likely still ‘heard’ parts of the sound that were not audible (schema based integration and Gestalt principle of closure [Bregman, 2008]). For example, the 500Hz octave band in the vocal may have been masked in some stimuli but not the reference. When playing both in turns, the listeners may have still ‘heard’ the missing frequency area. Therefore, the audible harmonic centroid is unlikely to be a ‘perceived harmonic centroid’. This was confirmed after listening to the unmasked parts of the audio in isolation, as extracted by the model described above, which sounded unnatural and different than in the presence of the backing track. Additional metrics that explore the relationship between target and backing track spectra are presented in the following.

10.2.5 Overall target-to-backing-track ratio

Clarity and separation was previously related to the extent to which sounds can be heard in a mix (chapter 2). As the audibility of sounds is directly linked to the overall target-to-interferer ratio, it would make sense to assess whether this influenced how clear target sounds were rated to be. In Fig. 10-14, target spectral clarity is plotted against the target-to-backing-track ratio, coloured separately for vocals and basses, as well as for stimuli where the target was equalised and stimuli where the backing track was equalised. All equalised versions of the target and the unequalized reference had been loudness matched before inserting them in the mix. Only due to the differences in masking, the mean target-to-backing-track ratio differed between stimuli. The mean target-to-backing-track ratio, averaged over all frequency bands, was calculated using the model introduced in section 10.2.4.
No consistent effect of the overall target-to-backing-track ratio on target spectral clarity can be observed. When the vocal overall target-to-backing-track ratio is increased, this appears to make the vocal clearer but only when the vocal is equalised. Bass targets with an increased overall target-to-backing-track ratio seem less clear when the backing track is equalised but no correlation between overall target-to-backing-track ratio and target spectral clarity change can be observed when the bass is equalised. As there is no clear trend visible across programme items and EQ treatments, it is concluded that overall target-to-backing-track ratio is a less suitable predictor of target spectral clarity than other factors presented in this chapter.

**Fig. 10-14:** Overall target-to-backing track-ratio, plotted against target spectral clarity

Overall mix loudness, which may also be linked to target audibility, also did not correlate with target spectral clarity either, as shown in Fig. 10-15.
10 Single sound spectral clarity in mixes

Fig. 10-15: the impact of changes in overall mix loudness on target clarity

It may be useful to test further ways to meter the relationship between target and backing track and target spectral clarity. In order to do this, it would be useful to investigate the contribution of octave bands to clarity in both the target and backing track.

10.2.6 The relationship between target and backing track spectra

In this section, the impact of boosts and cuts across the octave bands, as well as target harmonics, on target spectral clarity is analysed. In order to assess whether any octave bands contribute consistently positively or negatively to target spectral clarity, the ratings for boosts in each target are compared with the ratings for corresponding cuts in the backing track and, conversely, the ratings for cuts in the targets are compared with the ratings for corresponding boosts in the backing track (Fig. 10-16, Fig. 10-17, Fig. 10-18 and Fig. 10-19). There appears to be a relationship between the degrees of clarity change resulting from these boosts and cuts, and target-to-backing ratios of the corresponding targets in the reference mixes (Fig. 10-20 and Fig. 10-21). This is explored in detail below but, in general, it appears that if the audibility of frequency bands that are already high in amplitude is increased further, clarity increases, and that target spectral clarity is also increased when the less audible frequency bands become even less audible.

For the basses, the 2kHz octave band is the only band with a consistently positive contribution to target spectral clarity. In the audible bass LTAS, it can be seen that for the bass, a high target-to-backing-track ratio exists in the 2kHz band already; hence changes in audibility in this band may make a significant contribution to the audibility of the bass overall, and therefore to target
spectral clarity. At the same time, this band lies near the frequency area the auditory system is most sensitive to [Moore, 2012], and a boost here may allow the listener to follow the sound over time [Bregman, 2008]. On the other hand, all target stimuli were loudness matched before being inserted into the mix, hence a boost in this area in the target did not necessarily mean an increase in target loudness. The audible, unmasked spectra of all stimuli were inspected informally and an increase in energy in the band between 3kHz and 5kHz in the equalised version did not seem to correlate with an increase in target spectral clarity. Therefore, it is likely that the first interpretation is correct, that is that this band was already high in amplitude meant that making it even more audible lead to an increase in target spectral clarity.

Possibly for the same reason, boosting the fundamental of the bass (the band centred at 62Hz) also increases target spectral clarity: this band was also already high in amplitude before EQ. This may also be related to the fact that strengthening the fundamental of sounds may strengthen the perception of the sound’s pitch [Bregman, 2008]. Cutting the amplitude of the bass fundamental, or boosting or cutting the backing track in that area have no significant impact on target spectral clarity, however. The latter may be related to the fact that the backing track does not have much energy in this area. It is also possible that while on the one hand the fundamental becomes less audible in both cases, the band around 2kHz becomes more audible due to loudness matching, leading to a trade-off in target spectral clarity. Cutting the 125Hz and 250Hz bands in the backing track, as well as the bass, increases target spectral clarity, and conversely for cuts. This may be related to the influence of overall mix spectral centroids (overall mix clarity) on reported target clarity (section 10.2.8). Alternatively, these cuts may lead to release from upwards masking of the important 2kHz area.
For the vocal, target cuts and corresponding backing track boosts correlate with target spectral clarity in the same way as each other, except at 250Hz (cutting this band in both backing track and target increases target spectral clarity) and 8kHz (boosting this band in both backing track increases target spectral clarity, while both target boosts and cuts decrease clarity). This may,
10 Single sound spectral clarity in mixes

again, relate to overall mix clarity. At the same time, the original target-to-backing track-ratio is low in the 250Hz band. Apart from the 4kHz band, boosting the vocal always reduces target spectral clarity. The fact that most backing track boosts and cuts correlate with target spectral clarity in the same way as each other (for all octave bands with the exception of 4kHz, the boxes for backing track boosts and cuts move in opposite directions), is likely related to the relationship between target and backing track spectra again.

Fig. 10-18: the influence of vocal boosts and backing track cuts on clarity, measured across 8 octave bands

Fig. 10-19: the influence of vocal cuts and backing track boosts on clarity, measured across 8 octave bands
10 Single sound spectral clarity in mixes

Fig. 10-20: target-to-backing-track ratio between the vocal and backing track in the unequalised reference.

Fig. 10-21: target-to-backing-track ratio between the bass and backing track in the unequalised reference.

All in all, it appears that the relationship between target and backing track spectra influences target spectral clarity, i.e. the higher the amplitude of a frequency area is in the unequalised target, the more positively the amplitude ratio between target and backing track in that area contributes to target spectral clarity in the equalized version. The more that peaks in the target protrude through the backing track, the clearer the target is perceived (peak audibility). On the other hand, the more that areas further away from peaks protrude through the backing track, the more target spectral clarity is reduced. It is possible that this is due to the fact that increased peak audibility makes it easier for the timbral target characteristics to become apparent. A target peak audibility will therefore be devised.
10 Single sound spectral clarity in mixes

10.2.7 Peak audibility

In the previous section, it was shown that the audibility of peaks in the target appears to be particularly important, i.e. the higher the amplitude of a frequency area is in the unequalised target, the more positively the amplitude ratio between target and backing track in that area contributes to target spectral clarity in the equalized version. The more the peaks in the target protrude through the backing track, the clearer the target is perceived (peak audibility). This observation was used in order to devise a peak audibility metric as follows:

1. The audible spectra of the targets were calculated using the method described in section 10.2.4.

2. The LTAS of the audible spectrum in each stimulus was calculated in dBFS with 1/6 octave band smoothing. Smoothing was used because the overall shape of the spectrum is more important than the fine detail. The choice of 1/6 octave band smoothing ensured that the unnecessary detail was removed while still maintaining the overall shape of the spectrum.

3. The LTAS was interpolated on a logarithmic scale (so that e.g. each octave band contained the same number of points).

4. The amplitude value for each frequency bin in the reference was deducted from that in the equalized version (all amplitude values were in dBFS, and therefore negative, meaning that if the LTAS curve was higher for the equalized version in a given frequency area, this difference was positive):

\[ d(n) = e(n) - r(n), \quad (10.8) \]

where \( e(n) \) is the vector of \( n \) amplitude values for all frequency bins in the equalized stimulus and \( r(n) \) is the corresponding vector for the reference. The resulting vector of differences is \( d(n) \).

5. The resulting vector of differences \( d(n) \) was multiplied element-wise with a weighting function \( w(n) \), arriving at a vector \( a(n) \) of differences in peak audibility:

\[ a(n) = d(n) \cdot w(n), \quad (10.9) \]

\( w(n) \) is calculated as follows.

\[ w_{prep}(n) = \frac{(r(n) - \min(r(n)))}{\max(r(n)) - \min(r(n))} \quad (10.10) \]

\[ w(n) = \begin{cases} 0 & w_{prep}(n) < 0.7 \\ w_{prep}(n) & \text{otherwise} \end{cases} \quad (10.11) \]
6. All $n$ elements in the weighted vector of differences were summed, leading to the peak audibility metric $p$. Where the result was positive, this meant that the target peaks protruded through the backing track more than in the reference.

$$p = \sum_{i=1}^{n} a(n)$$

(10.12)

No correlation between the degree of change in target audibility and the degree of clarity change could be detected, therefore it was decided to use peak audibility as a binary metric, similarly to mid-range spectral peakiness. This was combined with the target harmonic centroid and mid-range spectral peakiness metric that proved effective in section 10.2.3.

It was decided to add 5 to the predicted clarity rating based on the harmonic centroid and mid-range spectral peakiness metric when peak audibility increased and to deduct 5 when it decreased. The small arbitrary value 5 was chosen to test whether a small improvement in the metric could be observed, i.e. whether medians would be arranged more linearly in this way. For those stimuli where the backing track was equalised, the resulting clarity prediction was always 5 or -5 because the target harmonic centroid and mid-range spectral peakiness was not altered.

The resulting model prediction is plotted against the recorded clarity values in Fig. 10-22. The Spearman correlation coefficient is 0.403 (p-value < 0.001). For the stimuli where the backing track was equalized, it is 0.618 (p-value < 0.001) and for stimuli where the target was equalized it is only 0.254 (0.176), therefore, this metric is a worse predictor for target clarity for the stimuli where the target was equalized than a metric that only considers target harmonic centroid and mid-range spectral peakiness. For the mix stimuli, the metric is rather crude, as predicted clarity can only take on 5 or -5. It was decided that the metric needed further improvement.
So far, all metrics using the masking model presented have been insufficient clarity predictors. This may have been because the model was not precise enough: for example, upward masking was not included and A-weighting is only a rough approximation of the auditory system’s uneven sensitivity to different frequency regions. Further temporal factors may need to be considered: it is possible that the target-to-backing-track ratio at the onsets of notes plays a greater role, as more neurons fire at the onsets of sounds [Roads, 1996].

It is also possible that the masking model was indeed sufficient but that simply removing masked parts of the target is not a fair representation of how the target is perceived, as discussed in section 10.2.4. It is possible that listeners perceive the whole mix more holistically and that target clarity is somehow connected to overall mix clarity. For instance, it is possible that the overall harmonic centroid of the mix (or some other type of centroid) may have influenced target spectral clarity, in the same way as was the case for single sounds previously. Listeners may have perceived overall mix spectral clarity changing and this may have impacted on the target spectral clarity ratings. This may have been either a result of dumping bias or the spectral clarity of sounds in mixes may not be separate from overall mix clarity. For example, when low frequencies in the mix are cut (frequencies below the mix's centroid), this could result in a release from upwards masking of higher frequencies in the target [Moore, 2012], making the target appear clearer. The influence of overall mix spectral centroids on clarity is tested in the next section.
10 Single sound spectral clarity in mixes

10.2.8 Overall mix spectral centroids

The harmonic and Mel centroids were the best predictors of spectral clarity in previous experiments. Therefore, these centroids were also calculated for the overall mix and correlations with target spectral clarity were calculated and plotted. The overall mix fundamental was defined as the average fundamental of the bass, the lowest frequency element in the mix. As all stimuli have the same overall average fundamental, the correlation between harmonic centroids or spectral centroids and target spectral clarity is the same.

Raising the overall mix harmonic centroid seems to be a useful predictor of target spectral clarity, mainly when the backing track is equalised (Fig. 10-23, Spearman, 0.619, p<0.05). When the target is equalised, no correlation between target spectral clarity and overall mix harmonic centroids can be observed. This is likely the case because the harmonic centroid was not moved far enough in the latter case: in previous experiments, an increase in spectral clarity only took place when the harmonic centroid was raised by more than 1.5–2 harmonics. As can be seen in Fig. 10-23, all ratings belonging to target EQ stimuli are clustered in the middle: it can be seen that the harmonic centroid was never raised or lowered by more than 3 harmonics. This may have been too little to affect reported target spectral clarity. The target spectral clarity ratings belonging to the backing track EQ stimuli feature harmonic centroid changes of up to 15 harmonics. The correlation between the overall harmonic centroid change of the full mix and the target spectral clarity change of all stimuli is 0.456 (p < 0.001).

Fig. 10-23: changes in harmonic centroids, plotted against target spectral clarity (resulting from target and backing track equalization)
In previous listening tests, the spectral centroid was also calculated in Mels, ERBs and Cents. Mel spectral centroids correlated more strongly with spectral clarity than linearly calculated centroids (Hz); hence the Mel spectral centroid of the overall mixes is included here, too. For each stimulus pair, the difference between the original and processed spectral centroid calculated in Mels (for the overall mix) is plotted against target spectral clarity (Fig. 10-24). Raising the Mel centroid of the overall mix once again seems to raise target spectral clarity (overall: Spearman, 0.463, p-value < 0.001), but only when the backing track is equalized (Spearman 0.543, p-value < 0.001 for the backing track EQ). When the target is equalized, no correlation between overall mix Mel centroids and target spectral clarity is visible, as before. The correlation coefficient for target spectral clarity and the overall mix Mel centroid resulting from backing track equalisation is more than 0.1 smaller than that for the harmonic centroid; hence, the overall mix harmonic centroid appears to be the most useful predictor for target spectral clarity.

In section 10.2.3, it was established that the interaction between target and backing track spectra appears to be important. However, the audible version of the previous harmonic centroid and mid-range spectral peakiness metric did not predict target spectral clarity well. The original metric, ignoring the backing track, is not applicable to stimuli where the backing track was equalized. Therefore, it is necessary to test further metrics that consider the interaction between target spectrum and backing track spectrum. By combining such metrics with the overall mix harmonic centroid, a useful predictor for all stimuli may be established.
10 Single sound spectral clarity in mixes

Fig. 10-24: changes in Mel centroids, plotted against target spectral clarity

10.2.9 Peak audibility and overall harmonic centroid

In section 10.2.6, it was shown that the audibility of peaks in the target spectrum appears to be particularly important, i.e. the higher the amplitude of a frequency area is in the unequalised target, the more positively the amplitude ratio between target and backing track in that area contributes to target spectral clarity in the equalized version. The more the peaks in the target protrude through the backing track, the clearer the target is perceived (peak audibility). A metric for peak audibility was presented in section 10.2.7. This will now be tested in combination with the overall mix harmonic centroid, the most useful metric for the overall stimulus set presented so far. The metrics were combined as follows:

1. The overall mix harmonic centroid was calculated (as in section 10.2.8). Next, the best fitting linear regression coefficients between this and target spectral clarity were established, in the format $c(x) = a + b \cdot h(x)$, where $c$ is clarity and $h$ is harmonic centroid change for a stimulus $x$. The approach was the same as in chapter 9 for target harmonic centroids and mid-range spectral peakiness (see equation 9.1).

2. The overall mix mid-range spectral peakiness was calculated. This was included because the peakiness of the targets did not seem to correlate with target spectral clarity, while another metric, calculated over the entire mix, did (harmonic centroid). No increases in mid-range spectral peakiness were found in any of the stimuli.
3. The peak audibility of the targets was calculated, as described in section 10.2.7, and again the predicted clarity change was incremented or decremented by 5 according to whether the peak audibility had increased or decreased.

The usefulness of this metric was compared to that of the overall mix harmonic centroids alone. As can be seen in Table 10-1, the overall correlation has increased by 0.1 and the root mean square error is reduced when peak audibility is considered in addition to the overall harmonic centroid. It was discussed in section 10.2.8 that the overall mix harmonic centroid was not moved far enough for those stimuli where the target was equalized, leading to a lower degree of correlation. When the half of the stimuli is analyzed where the backing track was equalized, the correlation coefficient is much higher, at 0.771. The combined overall harmonic centroid, mid-range spectral peakiness and peak audibility metric is plotted against clarity in Fig. 10-25. This metric is combined with target harmonic centroids and mid-range spectral peakiness in the next section.

<table>
<thead>
<tr>
<th>Metric Description</th>
<th>Spearman's rho</th>
<th>p-value</th>
<th>RMSE</th>
<th>Monte Carlo Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall mix HC</td>
<td>0.456</td>
<td>&lt; 0.001</td>
<td>6.799</td>
<td>0.012</td>
</tr>
<tr>
<td>Overall mix HC and peak audibility</td>
<td><strong>0.545</strong></td>
<td>&lt; 0.001</td>
<td><strong>6.694</strong></td>
<td>0.146</td>
</tr>
<tr>
<td>Overall mix HC and peak audibility (where the backing track was equalized)</td>
<td><strong>0.771</strong></td>
<td>&lt; 0.001</td>
<td><strong>5.619</strong></td>
<td></td>
</tr>
<tr>
<td>Overall mix HC and peak audibility (where the target was equalized)</td>
<td>0.104</td>
<td>0.6</td>
<td>7.769</td>
<td></td>
</tr>
</tbody>
</table>

Table 10-1: Spearman correlation coefficients, p-values, RMSEs and Monte Carlo simulation minimum Rho values for overall mix harmonic centroid and peak audibility metrics and median clarity ratings (no changes were measured in mid-range spectral peakiness)
10 Single sound spectral clarity in mixes

In the current section, all these metrics are combined. The target harmonic centroid and mid-range spectral peakiness metric \( u(x) \) (equal to 0 for all stimuli in which the backing track was equalized) and the overall mix harmonic centroid \( h(x) \) are combined in an overall linear regression model \( c(x) \). Subsequently, peak audibility is incorporated by adding 5 to the predicted clarity for stimuli \( x \) where the peak audibility change \( p(x) \) was positive (increase compared to reference) and by subtracting 5 for the stimuli where peak audibility change \( p(x) \) was negative (decrease compared to reference), as follows:

\[
c(x) = \begin{cases} 
4.055 + 0.670 \cdot u(x) + 0.704 \cdot h(x), & \text{if } p(x) > 0 \\
-6.055 + 0.670 \cdot u(x) + 0.704 \cdot h(x), & \text{if } p(x) < 0
\end{cases} \tag{10.13}
\]

\( p(x) \) was never 0. The model is plotted against the recorded clarity data in Fig. 10-26. The resulting Spearman correlation coefficient is greater than that for all other models tested in the

**Fig. 10-25:** The combined overall harmonic centroid, mid-range spectral peakiness and peak audibility metric, plotted against clarity

10.2.10 Target and overall mix harmonic centroid and mid-range spectral peakiness, combined with peak audibility

The most useful metric for predicting target spectral clarity in the overall stimulus set so far was a combination between overall mix harmonic centroid and peak audibility in the target. For those stimuli where the target was equalized, a combination of target harmonic centroid and mid-range spectral peakiness was more useful than a metric based on harmonic centroid only.
current chapter (0.568). All Spearman coefficients, RMSEs and Monte Carlo coefficients for linear regression models presented in this chapter are summarized in Table 10-2, unless no correlation was found.

![Graph showing extent to which the target is clearer than in the reference](image)

**Fig. 10-26: model prediction, plotted against clarity**

<table>
<thead>
<tr>
<th>Target and overall mix harmonic centroid and mid-range spectral peakiness, combined with peak audibility</th>
<th>Spearman’s rho</th>
<th>p-value</th>
<th>RMSE</th>
<th>Monte Carlo Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.568</td>
<td>&lt; 0.001</td>
<td>6.444</td>
<td>0.256</td>
<td></td>
</tr>
</tbody>
</table>

| Overall mix HC and peak audibility                                                                 | 0.545         | < 0.001 | 6.694 | 0.146 |

| Overall mix Mel centroid                                                                                 | 0.463         | < 0.001 |      |      |

| Overall mix HC                                                                                           | 0.456         | < 0.001 | 6.799 | 0.012 |

| Target HC and mid-range spectral peakiness (for stimuli where the target was EQ’d)                     | 0.437         | 0.02    | 9.556 | -0.181 |

| Peak audibility and target HC/mid-range spectral peakiness                                              | 0.403         | < 0.001 |      |      |

| Target HC (for stimuli where the target was EQ’d)                                                       | 0.344         | 0.06    | 10.141 | -0.274 |

Table 10-2: Spearman correlation coefficients, p-values, RMSEs and Monte Carlo simulation minimum Rho values for all models tested in the current chapter that exhibited correlation with median clarity ratings.
10.3 Discussion

Overall, the most suitable predictor for the single sound spectral clarity of targets in mixes was the combination of overall mix harmonic centroid, target harmonic centroids and mid-range spectral peakiness and the audibility of peaks in the target. Implications for further research outside the scope of this project are summarized below.

Firstly, it would be useful to repeat some of the above analyses with a more complex masking model, considering additional factors such as upwards masking and temporal variation. Secondly, it is likely that parts of the target sound are still ‘heard’ despite being masked which might have a big impact on perceived spectral clarity. This should be investigated further. Thirdly, it is possible that target spectral clarity in mixes is interpreted slightly differently from the way it is interpreted for isolated sounds. Whether the correlation between overall mix harmonic centroids and target spectral clarity is a result of dumping bias or whether target clarity does depend on mix clarity would need to be researched further to establish this.

In the bass, target spectral clarity appears to increase if either the low end or the string plucking sound becomes more audible. It is possible that some listeners interpreted “bass clarity” as the actual presence of low end in the bass. The loudness of the string plucking articulation sound also seemed to be very important in the bass. Due to the fact that both areas contribute to target spectral clarity, boosts and cuts in the same frequency area do not always move in opposing directions: for example, boosting the low end in the bass increases target spectral clarity. Cutting the same area does not decrease target spectral clarity, however, as due to the loudness matching, the pluck sound increases in audibility, leading to a clarity tradeoff. It would be useful to investigate the possibility that while separation relates to the ease with which a sound can be audibly distinguished from other sounds, clarity might relate (at least in part) to the ease with which each component part of a sound can be audibly distinguished from each other component part of that sound and those of other sounds.

10.4 Conclusion

In previous listening tests, a combination between the harmonic centroid and mid-range spectral peakiness was established to be a useful predictor of spectral clarity for harmonic, monophonic sounds in isolation. Sounds in mixes were considered in the current chapter. The aim of the current chapter was to test and improve the previous clarity metric for sounds in mixes and answer the research question: to what extent can the previous single-sound spectral clarity model predict the degree of target spectral clarity and separation change resulting from applying simple EQ to sounds in a mix? This question can now be answered.
Eighteen test subjects compared bass and vocal target sounds in equalised mix stimuli to the same target sounds in unequalised reference mix stimuli. Stimuli featured one boost or cut each in one of the octave bands centred at 62Hz (bass only), 125Hz, 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz in either the target sound or backing track. The programme item was a simple pop mix. Ratings were undertaken using a paired comparison test. For the first time in this thesis, sounds were evaluated within a mix, rather than in isolation, and so separation data were also collected. Clarity and separation were established to be significantly different; hence separation data were ignored in order to continue the focus of this thesis on clarity.

A combination of mid-range spectral peakiness and the harmonic centroid was useful previously in predicting spectral clarity in isolation (chapter 9). In mixes, the metric was not quite as useful as for isolated sounds but a positive correlation could still be observed and the combined metric correlated better with clarity than the target harmonic centroid alone. The metric was not suitable for those stimuli where the backing track was equalised, however, as the harmonic centroid and mid-range spectral peakiness of the target did not change here.

By using a masking model, harmonic centroid and mid-range spectral peakiness changes in the audible parts of the target sounds were calculated. These did not correlate with clarity, however. The overall target-to-backing-track ratio and overall mix loudness also did not correlate with clarity. It is likely that complex fusion phenomena need to be considered to understand this further.

From inspecting box plots, as well as the long-term average spectra of the unmasked parts of the sounds, the relationship between target and backing track spectra appeared to be important for target spectral clarity. Following this, a model measuring the audibility of peaks in the target was devised. A combination between the peak audibility and a combination of target harmonic centroid and mid-range spectral peakiness did not predict clarity well, however.

It was suggested that clarity in mixes may be perceived more holistically. Overall mix spectral centroids (harmonic and Mel) correlated positively with target spectral clarity. It is possible that overall mix spectral clarity may have influenced target spectral clarity ratings, either due to dumping bias or because the mix is perceived holistically. Alternatively, a release from upwards masking of higher frequencies in the target may have occurred when lower areas in the backing track were cut. Whether this phenomenon is a result of dumping bias or whether target clarity does depend on mix clarity would need to be researched further. Peak audibility, in combination with the overall mix harmonic centroid and timbral mid-range spectral peakiness yielded results that were more useful than all other metrics tested up to that point.
Lastly, target harmonic centroids and mid-range spectral peakiness were combined with overall mix harmonic centroids and mid-range spectral peakiness (which did not change), as well as peak audibility. The resulting model was able to predict clarity better than all other metrics tested in this chapter. Overall, it appears useful to include the harmonic centroid of both the target and overall mix, target mid-range spectral peakiness and peak audibility in an overall clarity model for targets in mixes.
11 Conclusion

Mixing music is the process of combining tracks of recorded audio to an overall piece. This is a complicated process and, hence, automatic mixing or metering tools would be useful. The aim of the current research project was to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes. Findings could contribute to the development of marketable products such as a piece of software able to judge the overall sound quality of a mix, automatic mixers or sonically improved music production software.

A grounded theory approach was employed, as follows: first, the relevant high-level descriptive mix quality criteria and the lower-level perceptual attributes that relate to them were established through a literature search. One particularly important parameter (clarity and separation) was chosen to act as the focus for the remainder of the project. By drawing on academic literature, the acoustic factors that this may depend on were established. These factors were used as a guideline for experiment-based investigations of the impact of EQ on the spectral clarity part of the attribute, which ultimately lead to the development of a metric of this important aspect of mix quality.

A set of goals that structure this thesis have been presented in the introduction. In each chapter, one of these goals was reached, leading to the fulfilment of the overall aim. The findings from each chapter are summarized below, followed by a comparison with existing literature (section 11.1), implications for further research (section 11.2), main contributions to knowledge (section 11.3) and an overall summary (section 11.4).

Chapter 2 — High-level descriptive quality criteria and lower-level perceptual components of music mixes

Goal: compile a list of the high-level descriptive quality criteria and lower-level perceptual components that, according to published literature, contribute to the perceived quality of mixed music.

High-level descriptive quality criteria and lower-level perceptual components of music mixes were established through a search of the current literature. The high-level parameters are as follows.

- Clarity and separation — the extent to which individual components can be heard in a mix
11 Conclusion

- Balance — an even distribution of energy in the spatial and frequency domains
  o Horizontal or stereo balance is the extent to which within any given frequency range sound energy is distributed symmetrically and “evenly” between the left and right channels.
  o Depth is defined here as a sense of perspective in a mix, where sound sources can be placed at various distances from the listener and inside a fictional, reverberant space of a certain size and shape.
  o Tonal balance is the extent to which sound energy is distributed “evenly” across the frequency spectrum
- Impact and interest — the extent to which the mix grabs the listener’s attention
- Freedom from technical faults (e.g. unwanted recording artefacts)
- Context specific characteristics — the extent to which the mix fits current trends, fashions and norms, complements artistic purpose and supports the musical content

These high level perceptual attributes are influenced by a variety of low-level attributes, as outlined in table 2-1. Impact and interest may depend on clarity and separation, as well as balance. Context specific parameters are difficult to generalize. The perception of balance and the freedom from technical faults may also, to an extent, depend on the current fashion (e.g. distortion as stylistic device, ‘lopsided’ Beatles mixes). The extent to which instruments can be heard in a mix and timbral clarity are likely to depend on acoustic and psychoacoustic factors only (e.g. masking and loudness) and appear independent from the other high-level parameters. Therefore, clarity and separation were used as a starting point for further research.

Chapter 3 — Relating psychoacoustic findings to clarity and separation in music mixes

Goal: establish the acoustic and mix parameters that clarity and separation are likely to depend on, by consulting literature on acoustics and psychoacoustics.

The psychoacoustic factors determining clarity and separation were summarized by consulting literature on timbre, concert halls, auditory scene analysis, masking, loudness and speech intelligibility and grouped into four salient categories. The following three research questions were answered. Are there any characteristic parameters of clarity and separation that are presented across several areas of research? Is it likely that the factors influencing clarity and separation can be manipulated during the mix process? How do the lower-level parameters of clarity and separation presented in chapter 2 relate to the factors established in this chapter?
11 Conclusion

Although the ‘timbre’ and ‘clarity’ are not clearly defined, there seems to be a common agreement on which sounds are timbrally clear. The acoustic parameters that influence timbral clarity are related to the relative balance between HF and LF content, namely the number of harmonics, spectral centroid, relative level of low-order even harmonics, spacing of harmonics, and spectral slope.

In the context of concert halls, clarity depends on the composition, room acoustics, performers and the auditory acuteness of the listeners. It can be impaired by reverberation, echoes, noise, tonal distortion, sympathetic ringing tones, flutter echoes and non-uniformities of hearing conditions.

Cues for the grouping of sonic elements in the context of auditory scene analysis are similarities in their frequencies, pitches, amplitudes, locations, timbres and the centre frequencies of noise bands, the absence of sudden changes of these variables or long durations of silence between sounds, the sounds’ onset and offset synchrony, the regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation and parallel gliding of spectral components. The absence of these cues leads to the separation of sounds. Attention can also influence auditory scene analysis. In mixes, separation needs to exist between some sounds while others need to blend.

Sounds are likely to be masked when other, louder sounds are present. In the spectral domain, the likelihood of masking is increased when the masker has energy close to the target's frequency. Temporal factors (e.g. comodulation masking release, dip listening, profile analysis and the overshoot effect) can create a release from masking. Temporal factors are also important (i.e. forward and backward masking). Spatial cues like interaural time and level differences, spectral differences caused by differences in placement in the median plane and differing temporal or reverberation characteristics of sounds can counteract masking. Reverberation itself can also cause masking.

Increased loudness corresponds to an increased amount of energy between 3kHz and 5kHz, an increased intensity, temporal changes in intensity (longer durations, abrupt onsets) amplitude modulation between 30Hz and 100Hz, spatial factors (diotic sounds and diffuse stimuli) and partials that are in phase. Loudness adaptation, differences between listeners (e.g. fatigue and hearing loss) and the presence of visual stimuli also influence loudness.

Speech consists of harmonic tones featuring resonances (formants) and noises. The knowledge of the linguistic context, visual cues and audibility of relevant acoustic cues increase speech intelligibility. Acoustic cues exist in the dimensions of intensity, frequency and time, i.e. the manner of phoneme production, timing, intensity, voicing, frequency, noise filtering, rhythm and
Conclusion

intonation. The frequency area between 200Hz to 3kHz appears to be particularly important. The temporal fine structure information is important and usable over a wide range of frequency areas. Certain cues are redundant: speech with masked or missing parts can still be understood in many cases.

The research questions could now be answered. All parameters that were related to clarity and separation can be assigned to the categories spectrum, spatial factors, intensity and temporal changes. The spectrum appears to play a particularly important role for clarity and separation across all areas of literature. Sounds are separated and mask each other less if they have different spectra. Single sounds appear timbrally clearer and louder when they have more high mid frequency energy and certain important speech cues lie in this area, too.

The spectra of sounds can be manipulated during the mix process through tools such as equalization, distortion and exciters. Spatial factors can be manipulated through tools such as panning and reverberation. The intensity of sounds can be altered by adding or deducting gain and by using tools such as compressors or limiters. The changes of these factors over time can be altered through e.g. automation.

Most of the lower-level parameters of clarity and separation presented in chapter 2 could be related to the factors found in chapter 3 and these relationships were set out in table 3-3.

Therefore, it seems useful to assess the relationship between clarity and separation and spectral equalisation in music mixes through experimentation. In order to reduce the complexity, it was decided to focus on the relationship between single band EQ and the spectral clarity of single, isolated sounds.

Chapter 4 — The influence of dumping bias on spectral clarity ratings (pilot)

Goal: evaluate a suitable experimental setup for assessment of the relationship between spectral factors and clarity, and test for dumping bias.

It was suggested that the impact of boosts and cuts in different frequency regions on the timbral clarity of single sounds was assessed in a listening test. In order to prepare for such a listening test, a pilot study was conducted with the aim of testing the suitability of the setup and to consider the possibility of dumping bias. Dumping bias is an effect whereby listeners’ ratings of a specific attribute are affected by an inability to rate other changes that are perceived. The following research questions were answered. Does previous or simultaneous exposure to rating other attributes have an impact on ratings of spectral clarity? Are the chosen stimulus pairs useful and do they have any visible impact on the clarity ratings? Do the interface, reproduction equipment and facilities work properly and are they easy for test subjects to use?
11 Conclusion

Pairs of loudness-matched programme items (guitar and piano) featuring 9dB boosts and cuts in the 6 octave bands centred at 250Hz, 500Hz, 1kHz, 2kHz, 4kHz and 8kHz were presented to listeners. Listeners rated which stimulus in each pair they perceived as clearer and by how much.

The mean ratings of timbral clarity did not depend on the presence or absence of exposure to rating additional attributes (no dumping bias). However, both previous and simultaneous exposure to rating other attributes can reduce statistical noise in ratings of the attribute under investigation. Hence, it was suggested that listeners rate the same additional attributes in future tests.

Boosts and cuts as detailed above can impact the clarity ratings of the chosen stimulus pairs significantly, and the same stimuli will therefore be useful in the next listening test.

The interface, reproduction equipment and facilities worked as planned and listeners found the task easy. All in all, the setup was deemed suitable for the investigation of the relationship between spectra and single sound clarity.

Chapter 5 — Assessing the contribution of different octave bands to the single sound spectral clarity of piano and guitar stimuli

Goal: establish how boosts and cuts in different frequency regions influence single sound spectral clarity.

A listening test was carried out with the aim to answer the research question: how do boosts and cuts in different frequency regions influence single sound spectral clarity? Listeners rated the clarity of equalized stimuli against unfiltered references, using the setup described in the last section. Frequencies above 2kHz to 4kHz appeared to contribute to clarity positively and lower frequencies negatively. Some differences could be observed between the guitar and piano programme items.

It was suggested that impact of differences in the spectra of programme items on the relationship between EQ and clarity ratings is assessed in more detail in future listening tests.

Chapter 6 — Assessing the contribution of different octave bands to single sound clarity, depending on programme items

Goal: establish how changes in the spectral clarity of equalized programme items with differing fundamentals and original spectral centroids can be predicted.
The chapter 6 listening test aimed to answer the research question: how can changes in the spectral clarity of equalized programme items with differing fundamentals and original spectral centroids be predicted?

Using the same setup and EQ settings as in previous listening tests, listeners rated the clarity of plucked and bowed violin and cello stimuli. The four programme items were chosen as they differ in their fundamental frequencies and spectral centroids, while still having similar timbres.

Mel scaled and linear spectral centroids were found to correlate well with clarity. Boosting higher harmonics or cutting lower ones also increased clarity, indicating that the inclusion of the fundamental in an overall clarity metric would be useful. Following the Spearman correlation coefficient, a measure used to determine to what degree the relationship between data sets is monotonic, the harmonic centroid was established to be the most useful predictor for the impact of equalisation on single sound clarity. The harmonic centroid considers both the spectral centroid and fundamental frequency.

Chapter 7 — The single sound spectral clarity of vocal and bass stimuli

Goal: test the predictor of single sound spectral clarity on a new set of stimuli.

In order to test the ability of the harmonic centroid (and other previously tested predictors) to predict clarity for a larger group of programme items, a similar experimental setup was used as before, this time using bass and vocal programme items. The EQ treatments were the same as before (except that this time the octave bands centred at 62Hz and 125Hz were included in the boosts and cuts in order to cater for the lower fundamentals of the programme items).

It was suggested that for both programme items, clarity may work differently than for the previously tested stimuli, as the human voice is likely perceived differently from other instruments and as the low frequency content may contribute differently to bass clarity than that of other instruments (basses are often used to add low frequency content to mixes, while a reduction in LF content previously led to a clarity increase). The listening test aimed to answer the research question: does raising the harmonic, linear, Mel, ERB and Cent centroids, boosting higher harmonics or cutting lower ones increase single sound clarity for vocal and bass stimuli, when the same EQ settings are used as in the last experiment?

None of the EQ treatments made the vocal clearer in isolation and only two of the cuts increased bass clarity significantly. However, the further the centroids were raised, the more the negative impact on clarity diminished. It was concluded that the harmonic centroid can only predict clarity with limited accuracy and that it may be either unsuitable, or that further factors may need to be considered additionally.
Chapter 8 — Single sound clarity adjustment task

Goal: establish what additional factors are likely to be important for spectral clarity, what the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ are and how spectral clarity be defined in the context of EQ adjustments.

In order to investigate whether the harmonic centroid was an unsuitable predictor of single sound clarity, or whether additional factors needed to be investigated, and to understand clarity in the context of EQ treatments more fully, a combined EQ adjustment and verbal elicitation task was carried out, with the aim of answering the following three research questions. What are the favoured centre frequencies, gains and bandwidths for clarity-enhancing EQ? How can spectral clarity be defined in the context of EQ adjustments? What additional factors are likely to be important for spectral clarity?

The favoured centre frequencies spanned 1kHz to 7kHz. Preferred gains were around -7dB for cuts and 9dB for boosts, which is similar to previously used gains. Boosts were preferred to cuts. The Q settings differed considerably from those used in previous experiments. In general, much wider bandwidths were preferred. The preferred settings did not depend heavily on each other or on the stimuli.

The most frequently mentioned factors were those related to frequency content (22 instances). Many of the phrases within this category appear to be directly related to the harmonic centroid, namely bandwidth, brightness, frequency balance, the energy in specific frequency bands, warmth and thinness. It is therefore likely that the harmonic centroid is still useful in an overall clarity metric. Other important frequency related factors appear to be related to unpleasant lumps or 'holes' in the spectrum, i.e. cuts and boosts to specific frequencies, harmonic content, muddy and muffled sound and resonances. In addition, factors relating to amplitude, envelope, expectations, readability/intelligibility, recording artefacts, reverberation and hedonic responses were found.

Considering the factors that were mentioned most often, spectral clarity in this context could be defined as the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard; it seems that their audibility might be increased by high frequency boosts and low frequency cuts (since most sounds have less energy in the upper parts of their spectra), and avoidance of potentially distracting unnatural/unpleasant resonances and other artefacts.
Chapter 9 — An improved spectral clarity predictor

Goal: establish if the previous predictor of spectral clarity (for single sounds) can be improved by including a metric for potentially unpleasant peaks in the spectrum.

Spectral peakiness (caused by unnatural sounding resonances) was suggested by the previous experiment to be a useful, additional clarity factor. The aim of the current chapter was to use this information to improve the overall clarity metric and to answer the research question: can the harmonic centroid based predictor of single sound clarity be improved by including a metric for potentially unpleasant peaks in the spectra? Mid-range spectral peakiness was estimated by measuring the height of sharp peaks, as affected by spectral equalization situated towards the middle of stimulus long-term average spectra. Through a model, it was shown that by measuring changes in the harmonic centroid, as well as mid-range spectral peakiness, single sound spectral clarity can be predicted better than by using the harmonic centroid only (improvements in the Pearson correlation coefficient of up to 0.196 could be measured).

Chapter 10 — Single sound spectral clarity in mixes

Goal: test and improve the metric for sounds in mixes, as affected by spectral equalization.

As single sound clarity in isolation can be successfully measured with the metric introduced above, the purpose of the chapter 10 experiment was to investigate the clarity and separation of sounds in mixes. The aim was to answer the research question: to what extent can the previous single-sound spectral clarity model predict the degree of target spectral clarity and separation change resulting from applying simple EQ to sounds in a mix? Using the same setup as for the first four experiments, test subjects compared bass and vocal target sounds in equalised mix stimuli to the same target sounds in unequalised reference mix stimuli. Stimuli featuring one boost or cut each in either the target sound or the backing track were created for vocal and bass.

Clarity and separation were established to be significantly different; hence separation data were ignored in order to continue the focus of this thesis on clarity. Similarly to isolated sounds, the clarity of targets in mixes could be predicted better through a combination of target mid-range spectral peakiness and harmonic centroid than harmonic centroid alone. However, the correlation was worse than for isolated sounds. Furthermore, it was not suitable for those stimuli where the backing track was equalised, as the harmonic centroid and mid-range spectral peakiness of the target did not change here. Changes in the harmonic centroids and mid-range spectral peakiness of the audible, unmasked part of the targets did not correlate with clarity. The overall target-to-backing track ratio and overall mix loudness also did not correlate with
11 Conclusion

clarity. It is likely that complex fusion phenomena need to be considered to understand this further.

However, the relationship between target and backing track spectra did appear to be important for target spectral clarity. Following this, a model measuring the audibility of peaks in the target was devised. A combination between the peak audibility and a combination of target harmonic centroid and mid-range spectral peakiness did not predict clarity well, however.

It was suggested that clarity in mixes may be perceived more holistically. Overall mix spectral centroids (harmonic and Mel) correlated positively with target spectral clarity. Whether this phenomenon is a result of dumping bias or whether target clarity does depend on mix clarity would need to be researched further. Peak audibility, in combination with the overall mix harmonic centroid and mid-range spectral peakiness yielded results that were more useful than all other metrics tested up to that point.

Lastly, target harmonic centroids and mid-range spectral peakiness were combined with overall mix harmonic centroids and mid-range spectral peakiness, as well as peak audibility. The resulting model was able to predict clarity better than all other metrics tested in this chapter. Overall, it appears useful to include the harmonic centroid of both the target and overall mix in an overall clarity model for targets in mixes, as well as a peak audibility metric and target mid-range spectral peakiness. The previous predictor of single sound spectral clarity still appears somewhat useful for predicting target clarity changes in mixes but the presence of the backing track means that complex masking and fusion phenomena are likely to be important additionally.

11.1 Relating findings to the literature

In the current section, the findings of the experimental phase of the research project will be related to the literature consulted in chapter 3. It will be assessed which of the factors that have been shown to conclusively relate to clarity were noted in the literature, whether the relationship with clarity was found to be the same or different and whether additional factors were noted that have thus far not been studied in the knowledge of the authors.

The two factors contributing to the spectral clarity of single sounds established in this research project are the harmonic centroid and mid-range spectral peakiness (related to sharp peaks in the frequency spectrum). While the harmonic centroid was not directly related to spectral clarity in the literature consulted, timbral clarity was shown to depend on the relative amount of energy in a sound’s low and high frequency areas, as measured by the spectral centroid and spectral slope. Clarity was also related to brightness, which also correlates with the spectral
Conclusion

Authors came to conflicting conclusions as to the nature of the relationship between high and low frequency and timbral clarity: some related an increase in high frequencies and a decrease in low frequencies to an increase in timbral clarity, while others found the opposite to be true. It was argued that the disagreement may have been due to the different contexts in which clarity had been researched. The findings for single sound spectral clarity in the current research project were in agreement with the majority of studies undertaken, however, where high frequencies contributed to clarity positively and low frequencies negatively (increase in harmonic centroid). In chapter 10, low frequency boosts on the bass could increase its clarity in the presence of the mix in some cases. In this case, the harmonic centroid was lowered. It is likely that the increased prominence of the fundamental made it easier to hear it apart from the rest of the mix, even though its spectral clarity may have been lowered.

Mid-range spectral peakiness has thus far not explicitly been related to spectral clarity in the literature but it is possible that it can be related to some of the found factors. These are tonal distortion and sympathetic ringing tones in concert hall clarity and the evenness of the high frequency response in timbral clarity. Tonal distortion is likely to be similar to unpleasant resonances. Listening test subjects in chapter 8 mentioned ringing as a factor that can reduce clarity. The evenness of the high frequency response may also be related to high frequency resonances that may reduce clarity. However, mainly resonances towards the centre of the spectrum were established to be important for mid-range spectral peakiness, which none of the sources mentioned. Expectations and naturalness were also found to be related to clarity in chapter 8 but no literature was found that confirmed this.

For sounds in mixes, additional factors were found to be important. Adding a peak audibility measure proved useful here. Masking in mixes has been researched previously but the audibility of specific frequency areas in target sounds (i.e. the peaks) has not explicitly been related to the spectral clarity of target sounds in mixes. The overall target to backing track ratio did not predict clarity well in the current research project. None of the factors found in the literature were shown to not relate to clarity. Table 11-1 shows the extent to which all clarity factors found in the literature and in the experimental phase of the research project overlap.
11 Conclusion

<table>
<thead>
<tr>
<th>Only in the literature review</th>
<th>Both literature review and experimental phase</th>
<th>Experimental phase only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of harmonics</td>
<td>Masking</td>
<td>Masking of specific frequency areas/peak audibility</td>
</tr>
<tr>
<td>Even harmonics</td>
<td>Amplitude</td>
<td>Mid-range spectral peakiness</td>
</tr>
<tr>
<td>Spacing of harmonics</td>
<td>Loudness/intensity</td>
<td>Expectations</td>
</tr>
<tr>
<td>Evenness of HF response</td>
<td>Frequency content (bandwidth, brightness, amount of energy in specific frequency areas, frequency balance, harmonic content)</td>
<td>Naturalness</td>
</tr>
<tr>
<td>Decay time of the midrange frequency response</td>
<td>Other timbral attributes like ‘muddy’, ‘muffled’, ‘thinness’, ‘warmth’</td>
<td>Resonances</td>
</tr>
<tr>
<td>Specific ASA factors (those that can be related to mixes are the timbres of target and backing track, harmonic relations, binaural frequency matches, the suddenness of changes of amplitudes, spatial features, timbres and the center frequencies of noise bands, onset and offset synchrony, parallel amplitude modulation, parallel gliding of spectral components, the principle of common fate and the old plus new heuristic)</td>
<td>Readability/intelligibility</td>
<td>Ringing</td>
</tr>
<tr>
<td>The phases of overtones</td>
<td>Noise, hiss, distortion, artifacts</td>
<td>Doppler effect</td>
</tr>
<tr>
<td>Duration</td>
<td>Sibilance</td>
<td>Pleasantness/hedonic responses</td>
</tr>
<tr>
<td>Speech audibility cues</td>
<td>Spill</td>
<td></td>
</tr>
<tr>
<td>Temporal fine structure</td>
<td>Reverb</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Echoes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modulation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sound envelopes/attack/timing of onsets</td>
<td></td>
</tr>
</tbody>
</table>

Table 11-1 clarity factors found in the literature and in the experimental phase of the research project and their overlap.

11.2 Implications for further research

In the following, implications for further research are presented. In section 11.2.1, steps towards investigating single sound spectral clarity further are proposed. In section 11.2.2, research on the spectral clarity in mixes is proposed. In section 11.2.3, further potential research of additional parameters of clarity and separation is presented. Section 11.2.4 proposes assessing the interrelation between clarity factors. Lastly, in section, 11.2.5, it is proposed that the other parameters of high quality mixes are investigated.

11.2.1 Further investigation of single sound spectral clarity

The spectral parameters established here could be tested for a much larger group of stimuli, leading to a more generalizable model of spectral clarity. For instance, a more accurate measure of mid-range spectral peakiness could be devised. Currently, the metric employs a simple gate, only assessing whether mid-range spectral peakiness was increased or decreased. The extent to which this was the case was ignored, as this may have led to an over-fitted model most likely
only applicable to the stimuli it was trained on. At the same time, no listener data was available against which to directly assess the accuracy of predictions of mid-range spectral peakiness or peak audibility. If these models were to be improved, listener data should be gathered as a first step. By carrying out listening tests on much larger groups of stimuli, it might be possible to predict single sound spectral clarity more precisely. Another useful way to understand mid-range spectral peakiness more fully would be to formally establish which types of distortions of the spectral shape exactly sound unpleasant. The difficulty here is to establish which resonances are a natural part of the sound and which are unwanted. For example, the absence of a natural resonance may also lead to timbral unpleasantness or unnaturalness, and therefore a reduction in clarity.

It is possible that spectral clarity can reach a saturation point and that each sound can have a maximum or minimum clarity. For a generally applicable model of spectral single sound clarity, this should be investigated further. The harmonic centroid was established to be a useful predictor of single sound spectral clarity. It was shown that when the harmonic centroid is raised, the clarity increase eventually stops. Similarly, listeners stated in the adjustment task that in some cases, programme items were already clear, making the clarity adjustment task difficult. Interestingly, when the harmonic centroid and clarity were already high, listeners increased the harmonic centroid even further in many cases. This may make it difficult to measure absolute clarity: due to the fact that further harmonic centroid increases do not impact on clarity significantly when the sound is already clear, several stimuli with the same clarity rating may differ in their harmonic centroids. The additional parameters of clarity found in the literature and chapter 8 could also be investigated further, i.e. the relationship between clarity and amplitudes, envelope related factors, and the loudness of articulation sounds, reverb, artefacts and distortion.

11.2.2 Further investigation of spectral clarity in mixes

It would be useful to reassess spectral target clarity in mixes using CASA models. The clarity of sounds appeared to depend on fusion phenomena in a complex way. The gathered mix experimental data could also be investigated more fully with a more complex masking model, considering additional factors such as upwards masking and temporal variation. It is likely that parts of the target sound are still ‘heard’ despite being masked which has a big impact on perceived clarity.

It is possible that clarity in mixes is interpreted slightly differently from the way it is defined for isolated sounds. For example, in the bass, clarity appeared to increase either if the low end or the pluck sound became more audible. In order to investigate this further, it would be useful to
11 Conclusion

carry out a verbal elicitation task similar to chapter 8, where listeners are asked to adjust the clarity of targets in mixes and to provide further definitions. Clarity was shown to differ from separation; hence separation was not assessed in as much detail. It would be useful to understand separation fully also. Lastly, it would be interesting to assess whether equalisation, placed on the master bus, can make an entire mix appear clearer.

11.2.3 Investigating additional parameters of clarity and separation

Clarity and separation could be altered outside the context of spectral parameters. Spatial, temporal and intensity related factors should also be investigated.

11.2.4 Interrelation between clarity factors

As clarity is influenced by a multitude of factors, it is not one-dimensional like e.g. loudness or pitch. In chapter 8, definitions for clarity were elicited from listeners. The definitions were split into several categories: amplitude, envelope, expectations, frequency content, readability/intelligibility, recording artefacts, reverb and hedonic responses. Spectral factors were established to be the most important and were the focus of this research project. It is possible that e.g. spectral clarity can be high, while spatial clarity is low, indicating that clarity may be a multidimensional attribute.

Once the nature of the relationship between each of the factors and clarity, depending on programme items, has been established, it is also necessary to assess how the factors relate to each other. For example, an increase in one factor may lead to a decrease in another and the resulting impact on clarity may be a weighted sum of the factors. Lastly, the clarity of time varying signals needs to be measured: it is possible that if a given factor changes over time, clarity depends on the average position of that factor, or the position of that factor at specific times, e.g. the onsets of notes.

11.2.5 Other parameters of high quality mixes

In order to measure mix quality successfully, it would be necessary to measure all high-level parameters of mixes, that is, not only clarity and separation, but also balance, impact and interest and the freedom from technical faults. Some context-specific parameters could be measured through comparison to a reference, e.g. mixes of a similar fashion or style.
11 Conclusion

11.3 Key contributions to knowledge

The aim of the current research project was to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes. In the current section, key contributions to knowledge will be summarized.

The high-level parameters of music mixes are clarity and separation, spatial and tonal balance, impact and interest, freedom from technical faults and context-specific parameters. Clarity and separation were chosen as a starting point for further research. All parameters that clarity and separation depend on can be assigned to the categories spectrum, spatial factors, intensity and temporal changes. The spectrum appears to play a particularly important role for clarity and separation. For example, sounds are separated and mask each other less if they have different spectra. Single sounds appear timbrally clearer and louder when they have more high mid frequency energy and certain important speech cues lie in this area, too.

In devising a suitable experimental test setup, it was established that both previous and simultaneous exposure to rating other attributes (in this case warmth, fullness and brightness) can reduce statistical noise in ratings of the attribute under investigation (spectral clarity). For piano and guitar stimuli, frequencies above 2kHz to 4kHz contribute to clarity positively and lower frequencies negatively. Some differences could be observed between programme items. Mel scaled and linear spectral centroids were found to correlate well with clarity for string, piano and guitar stimuli. Boosting higher harmonics or cutting lower ones also increased clarity, indicating that the inclusion of the fundamental frequency in an overall clarity metric might be useful. The harmonic centroid (a metric that considers both the spectral centroid and fundamental frequency) was established to be the most useful predictor for the impact of equalisation on single sound spectral clarity.

For a more comprehensive spectral clarity predictor, additional potentially relevant factors were identified. These include factors related to unpleasant lumps or ‘holes’ in the spectrum, muddy and muffled sound, readability/intelligibility, recording artefacts and reverberation. Considering these factors, spectral clarity in this context could be defined as the extent to which the spectral shape of a sound allows all the important components of its natural timbre to be heard; their audibility might be increased by high frequency boosts and low frequency cuts (since most sounds have less energy in the upper parts of their spectra), and avoidance of potentially distracting unnatural/unpleasant resonances and other artefacts. The favoured centre frequencies for clarity enhancing EQ span 1kHz to 7kHz. Preferred gains lie around -7dB for cuts and 9dB for boosts. Boosts are preferred to cuts. For Q settings, wide bandwidths are preferred.
11 Conclusion

By measuring changes in the harmonic centroid, as well as mid-range spectral peakiness, single sound spectral clarity could be predicted better than by using the harmonic centroid only (improvements in the Pearson correlation coefficient of up to 0.196 could be measured). These two factors were sufficient to model clarity changes with a Spearman correlation of 0.631 (bass and vocal stimuli) to 0.848 (string stimuli).

For sounds in mixes, it appears useful to include the harmonic centroid of both the target and overall mix in an overall clarity model for targets in mixes, as well as a peak audibility metric and target mid-range spectral peakiness. The previous predictor of single sound spectral clarity still appears somewhat useful for predicting target clarity changes in mixes but the presence of the backing track means that complex masking and fusion phenomena are likely to be important additionally.

11.4 Summary

The overall goal of the research project was to work towards measuring the perceived quality of music mixes by establishing predictors for one important perceptual attribute of high-quality mixes.

The spectral clarity of single sounds is important to overall mix quality. Two factors contributing to the spectral clarity of single sounds were established. These are the harmonic centroid and mid-range spectral peakiness (related to sharp peaks in the frequency spectrum). For sounds modified by simple spectral filtering these two factors are sufficient to model clarity changes with a Spearman correlation of 0.631 (bass and vocal stimuli) to 0.848 (string stimuli). For sounds in a mix, however, other factors become important. Adding a peak audibility measure proved useful but further work and a more complex model is needed.
Appendices

The listening test instructions for all experiments (A1), the factor classification for the adjustment task (A2, chapter 8) and the computational model for the mid-range spectral peakiness metric (A3, chapter 9) are presented, followed by a list of publications related to the research project (A4).

A1 Listener instructions

In the following, images of the listener instructions are presented. In Fig. A-1 and Fig. A-2, the instructions for the experiment in chapter 4 are presented (pilot study used to evaluate a suitable setup and to test for dumping bias). Fig. A-3 shows the instructions for the experiment in chapter 5 (spectral piano and guitar clarity), Fig. A-4 shows the instructions for the experiment in chapter 6 (spectral string clarity) and Fig. A-5 shows the instructions for the experiment in chapter 7 (spectral vocal and bass clarity). Fig. A-6 shows the instructions for the experiment in chapter 8 (single sound clarity adjustment task) and Fig. A-7 shows the instructions for the experiment in chapter 10 (clarity in mixes). All test instructions were accompanied by images of the test pages.
Appendices

Listening test: instructions

Thank you for taking part in this listening test.

1. The test comprises 11 GUI pages in total: a stimulus familiarisation page followed by 10 paired stimulus comparison pages.
2. On the stimulus familiarisation page (figure 1), select stimuli by clicking the buttons featuring a musical note, and use the play, pause and stop buttons to audition the selected stimuli. Please familiarize yourself with all the stimuli on this page.
3. Once you feel familiar with the stimuli, please tell the experimenter.
4. You will then be asked to familiarise yourself with the paired stimulus comparison procedure. This is as follows:
   a. On each comparison page (figure 2), select stimuli by clicking on the buttons labelled ‘A’ and ‘B’ (or by pressing ‘1’ and ‘2’ on the keyboard), and use the play, pause and stop buttons to audition the selected stimuli.
   b. Use the slider to indicate how the two stimuli compare in terms of clarity. If A sounds clearer than B then move the slider to the left, and set the distance of the slider from the middle to indicate the extent to which A sounds clearer than B. If B sounds clearer than A then move the slider to the right, and set the distance of the slider from the middle to indicate the extent to which B sounds clearer than A. If the stimuli sound equally clear then position the slider in the middle (you will have to move it to one side and back again, or at least click on it, to be able to go to the next page).
   c. You can adjust the slider as many times as you like until you are happy with its position. When you are happy with it, move on to the next page using the ‘> |’ button.
5. Complete further pages until you feel used to the interface and procedure and ready to begin the test; then please tell the experimenter.
6. You will then be asked to use the same GUI and procedure to rate the relative clarities of stimulus pairs on 10 test pages. Once you have completed all 10 pages, please tell the experimenter.

I’d be grateful if you’d avoid discussing this test with others please, until the end of the month. Should you have any questions, do not hesitate to ask!

Thank you.

Fig. A-1: listener instructions for the clarity only test half in the pilot study (chapter 4).
Listening test: instructions

Thank you for taking part in this listening test.

1. The test comprises 11 GUI pages in total: a stimulus familiarisation page followed by 10 paired stimulus comparison pages.
2. On the stimulus familiarization page (figure 1), select stimuli by clicking the buttons featuring a musical note, and use the play, pause and stop buttons to audition the stimuli. Please familiarize yourself with all the stimuli on this page.
3. Once you feel familiar with the stimuli, please tell the experimenter.
4. You will then be asked to familiarise yourself with the paired stimulus comparison procedure. This is as follows:
   a. On each comparison page (figure 2), select stimuli by clicking on the buttons labelled ‘A’ and ‘B’ (or by pressing ‘1’ and ‘2’ on the keyboard), and use the play, pause and stop buttons to audition the selected stimuli.
   b. Use the four sliders to indicate how the two stimuli compare in terms of warmth, clarity, brightness and fullness as follows.
   c. If A sounds warmer than B then move the slider to the left, and set the distance of the slider from the middle to indicate the extent to which A sounds warmer than B. If B sounds warmer than A then move the slider to the right, and set the distance of the slider from the middle to indicate the extent to which B sounds warmer than A. If the stimuli sound equally warm then position the slider in the middle (you will have to move it to one side and back again, or at least click on it, to be able to go to the next page).
   d. Repeat this procedure for clarity, brightness and fullness using the corresponding sliders.
   e. You can adjust the sliders in any order and as many times as you like until you are happy with all the slider positions. When you are happy with them, move on to the next page using the ‘>’ button.
5. Complete further pages until you feel used to the interface and procedure and ready to begin the test; then please tell the experimenter.
6. You will then be asked to use the same GUI and procedure to rate the relative warmths, clarities, brightnesses and fullnesses of stimulus pairs on 10 test pages. Once you have completed all 10 pages, please tell the experimenter.

I’d be grateful if you’d avoid discussing this test with others please, until the end of the month. Should you have any questions, do not hesitate to ask!

Thank you.

Fig. A-2: listener instructions for the clarity, brightness, warmth and fullness test half in the pilot study (chapter 4)
Appendices

**Listening test: instructions**

Thank you for taking part in this listening test.

1. The test comprises 58 GUI pages in total: a stimulus familiarisation page followed by 57 paired stimulus comparison pages.

2. On the stimulus familiarization page (figure 1), select stimuli by clicking the buttons featuring a musical note, and use the play, pause and stop buttons to audition the stimuli. Please familiarize yourself with all the stimuli on this page.

3. Once you feel familiar with the stimuli, please tell the experimenter.

4. You will then be asked to familiarise yourself with the paired stimulus comparison procedure. This is as follows:
   a. On each comparison page (figure 2), select stimuli by clicking on the buttons labelled 'A' and 'B' (or by pressing '1' and '2' on the keyboard), and use the play, pause and stop buttons to audition the selected stimuli.
   b. Use the slider to indicate how the two stimuli compare in terms of either warmth, clarity, brightness or fullness, as explained in c. The attribute to be rated is given at the top of each page (and may change on each page), while the other attributes are crossed out.
   c. If A sounds e.g. warmer than B then move the slider to the left, and set the distance of the slider from the middle to indicate the extent to which A sounds warmer than B. If B sounds warmer than A then move the slider to the right, and set the distance of the slider from the middle to indicate the extent to which B sounds warmer than A. If the stimuli sound equally warm then position the slider in the middle (you will have to move it to one side and back again, or at least click on it, to be able to go to the next page).
   d. You can adjust the slider in any order and as many times as you like until you are happy with all the slider positions. When you are happy with them, move on to the next page using the '>' button.

5. Complete further pages until you feel used to the interface and procedure and ready to begin the test; then please tell the experimenter.

6. You will then be asked to use the same GUI and procedure to rate the relative warmths, clarities, brightnesses or fullnesses of stimulus pairs on 57 test pages. Once you have completed all 57 pages, please tell the experimenter.

I’d be grateful if you’d avoid discussing this test with others please, until the end of the month. Should you have any questions, do not hesitate to ask!

Thank you.

Fig. A-3: listening test instructions for the guitar and piano clarity rating task (chapter 5)
Listening test: instructions

Thank you for taking part in this listening test.

1. The test comprises 50 GUI pages in total: a stimulus familiarisation page followed by 57 paired stimulus comparison pages.

2. On the stimulus familiarisation page (figure 1), select stimuli by clicking the buttons featuring a musical note, and use the play, pause and stop buttons to audition the stimuli. Please familiarize yourself with all the stimuli on this page.

3. Once you feel familiar with the stimuli, please tell the experimenter.

4. You will then be asked to familiarise yourself with the paired stimulus comparison procedure. This is as follows:
   a. On each comparison page (figure 2), select stimuli by clicking on the buttons labelled 'A' and 'B' (or by pressing '1' and '2' on the keyboard), and use the play, pause and stop buttons to audition the selected stimuli.
   b. Use the slider to indicate how the two stimuli compare in terms of either warmth, clarity, brightness or fullness, as explained in c. The attribute to be rated is given at the top of each page (and may change on each page), while the other attributes are crossed out.
   c. If A sounds e.g. warmer than B then move the slider to the left, and set the distance of the slider from the middle to indicate the extent to which A sounds warmer than B. If B sounds warmer than A then move the slider to the right, and set the distance of the slider from the middle to indicate the extent to which B sounds warmer than A. If the stimuli sound equally warm then position the slider in the middle (you will have to move it to one side and back again, or at least click on it, to be able to go to the next page).
   d. You can adjust the slider in any order and as many times as you like until you are happy with all the slider positions. When you are happy with them, move on to the next page using the '>>|' button.

5. Complete further pages until you feel used to the interface and procedure and ready to begin the test; then please tell the experimenter.

6. You will then be asked to use the same GUI and procedure to rate the relative warmths, clarities, brightnesses or fullnesses of stimulus pairs on 57 test pages. Once you have completed all 57 pages, please tell the experimenter.

I’d be grateful if you'd avoid discussing this test with others please, until the end of the month. Should you have any questions, do not hesitate to ask!

Thank you.

Fig. A-4: listening test instructions for the plucked and bowed violin and cello clarity rating task (chapter 6)
Fig. A-5: Listening test instructions for the bass and vocal clarity rating task (chapter 7)
Appendices

Fig. A-6: listening test instructions for the EQ adjustment task (chapter 8)
Appendices

Listening test: instructions

Thank you for taking part in this listening test.

1. The test comprises 66 GUI pages in total; these are presented in two test halves with a break in the middle. Each half comprises a stimulus familiarisation page followed by 32 paired stimulus comparison pages.

2. On the stimulus familiarisation page (figure 1), select stimuli by clicking the buttons featuring a musical note, and use the play, pause and stop buttons to audition the stimuli. Please familiarize yourself with the range of stimuli on this page.

3. Once you feel familiar with the stimuli, please tell the experimenter.

4. You will then be asked to familiarise yourself with the paired stimulus comparison procedure. This is as follows:
   a. On each comparison page (figure 2), select stimuli by clicking on the buttons labelled ‘A’ and ‘B’ (or by pressing ‘1’ and ‘2’ on the keyboard), and use the play, pause and stop buttons to audition the selected stimuli.
   b. Use the slider to indicate how the vocal in mix compares between the two stimuli in terms of either warmth, clarity, separation, brightness or fullness, as explained in c. The attribute(s) to be rated are given at the top of each page (and may change on each page), while the other attributes are crossed out. Do not rate the warmth, fullness, separation, brightness or clarity of the overall mix, but ONLY of the vocal!
   c. If the vocal in A sounds e.g. warmer than in B then move the slider to the left, and set the distance of the slider from the middle to indicate the extent to which A sounds warmer than B. If B sounds warmer than A then move the slider to the right, and set the distance of the slider from the middle to indicate the extent to which B sounds warmer than A. If the vocal sounds equally warm in both stimuli, then position the slider in the middle (you will have to move it to one side and back again, or at least click on it, to be able to go to the next page).
   d. You can adjust the sliders in any order and as many times as you like until you are happy with all the slider positions. When you are happy with them, move on to the next page using the ‘next’ button.

5. Complete further pages until you feel used to the interface and procedure and ready to begin the test; then please tell the experimenter.

6. You will then be asked to use the same GUI and procedure to rate the relative warmth, clarities, separations, brightnesses or fullnesses of stimulus pairs on 32 test pages. Once you have completed all 32 pages, please tell the experimenter.

7. After a short break, the exact same procedure is repeated, except that the warmth, fullness, brightness, clarity and separation of the bass are rated this time.

I'd be grateful if you'd avoid discussing this test with others please, until the end of the month. Should you have any questions, do not hesitate to ask!

Thank you.

Fig. A-7: listening test instructions for the rating task for spectral target clarity and separation in mixes (chapter 10)
A2 Factor classification

Table A-1 shows the factor classification carried out by the external researcher for experiment 9.

<table>
<thead>
<tr>
<th>Category</th>
<th>Factor</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors related to amplitude</td>
<td>Loudness</td>
<td>The overall loudness of the recording</td>
</tr>
<tr>
<td>Factors related to envelope</td>
<td>Amplitude envelope</td>
<td>The overall temporal envelope of the sound</td>
</tr>
<tr>
<td>Factors related to envelope</td>
<td>Attack of transients</td>
<td>Definition of the attack portion of transient sounds</td>
</tr>
<tr>
<td>Factors related to expectations</td>
<td>Corresponds with expectation</td>
<td>How well the sound corresponds to an internal reference/expectation</td>
</tr>
<tr>
<td>Factors related to expectations</td>
<td>Naturalness</td>
<td>How natural the reproduction sounds</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Bandwidth</td>
<td>Range between the highest and lowest frequency</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Brightness</td>
<td>Perceptual attribute related to the balance of frequency content</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Cuts and boosts to specific frequencies</td>
<td>Changes to a narrow range of frequencies</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Frequency balance</td>
<td>Relative balance of different frequency regions</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Harmonic content</td>
<td>Balance of harmonics</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>HMF &amp; HF content</td>
<td>Relative balance of high-mid and high frequency regions to other frequency regions</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>LF content</td>
<td>Relative balance of low frequency regions to other frequency regions</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>LMF content</td>
<td>Relative balance of low-mid frequency regions to other frequency regions</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Muddy and muffled</td>
<td>Perceptual attribute related to the balance of frequency content</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Resonances</td>
<td>Presence of strong resonances</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Thin</td>
<td>Perceptual attribute related to the balance of frequency content</td>
</tr>
<tr>
<td>Factors related to frequency content</td>
<td>Warmth</td>
<td>Perceptual attribute related to the balance of frequency content</td>
</tr>
<tr>
<td>Factors related to Readability_intelligibility</td>
<td>Intelligibility</td>
<td>The ability to focus on different details of the sound</td>
</tr>
<tr>
<td>Factors related to Readability_intelligibility</td>
<td>Readability</td>
<td>The ability to focus on different details of the sound</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Artefacts</td>
<td>The presence of unwanted sounds in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Distortion</td>
<td>The presence of distortion in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Doppler</td>
<td>The presence of Doppler type effects in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Noise and hiss</td>
<td>The presence of unwanted noise/hiss in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Performance artefacts</td>
<td>The prominence of performance artefacts in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Ringing</td>
<td>The presence of ringing in the sound</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Sibilance</td>
<td>The prominence of sibilants in the recording</td>
</tr>
<tr>
<td>Factors related to recording artefacts</td>
<td>Spill</td>
<td>Spill from other sources</td>
</tr>
<tr>
<td>Factors related to reverb</td>
<td>Reverb</td>
<td>The balance between direct and reverberant sound</td>
</tr>
<tr>
<td>Factors related to reverb</td>
<td>Room artefacts</td>
<td>How natural the reverberation sounds</td>
</tr>
<tr>
<td>Hedonic responses</td>
<td>Pleasantness</td>
<td>How pleasant the sound is</td>
</tr>
<tr>
<td>Hedonic responses</td>
<td>Quality</td>
<td>Overall quality of the sound</td>
</tr>
<tr>
<td>No change</td>
<td>Nothing/no change</td>
<td>No suggested changes</td>
</tr>
</tbody>
</table>

Table A-1: factor classification
Appendices

A3 The mid-range spectral peakiness model introduced in chapter 9

In the following, the mid-range spectral peakiness model introduced in chapter 9 is presented in more detail. The metric predicts whether a given equalized stimulus is more or less peaky than the reference stimulus.

A3.1 LTAS

The LTAS of the equalized stimulus and reference are calculated in dBFS with 1/6 octave band smoothing. Smoothing is used because the overall shape of the spectrum is more important than the fine detail and 1/6 octave band smoothing provides a useful balance between removing unnecessary fine detail and still keeping the overall shape of the spectrum.

A3.2 Interpolation

Each LTAS is interpolated to comprise equidistant points on a logarithmic scale (so that e.g. each octave band contains the same number of points).

A3.3 Curve fitting

A curve is fitted to each LTAS as follows. The amplitude values for the curve and LTAS bins corresponding to the lowest frequency are equal. The curve is fitted to each LTAS value in turn from the second to lowest to the highest frequency bin afterwards, using the following formula.

\[ r(n) = (c(n) \cdot f_{inc}(n) + r(n-1) \cdot (1 - f_{inc}(n))) \cdot \frac{1 + d(n)}{2} \]

\[ + (c(n) \cdot f_{dec}(n) + r(n-1) \cdot (1 - f_{dec}(n))) \cdot \frac{1 - d(n)}{2} \]

where \( n \) is the current frequency bin that the curve is fitted for, \( c(n) \) is the current LTAS value and \( r(n) \) is the resulting curve value. \( r(n) \) is calculated using the function \( f_{inc} \) when the current LTAS value is greater than the previous curve value \( r(n-1) \), or the function \( f_{dec} \) when the current LTAS value is smaller than the previous curve value. A mixture of both is used if the difference between the values is smaller than 1. \( f_{inc} \) and \( f_{dec} \) determine how similar the currently fitted
curve value is either to the previous curve value (loose fit) or the current LTAS value (tight fit).

The difference between the current LTAS value and the previous curve value is determined by the variable \(d(n)\), calculated as a value between -1 and 1, as follows:

\[
d(n) = s(n) \cdot t(n)
\]

where

\[
s(n) = \begin{cases} 
1 & c(n) > r(n - 1) \\
0 & c(n) = r(n - 1) \\
-1 & c(n) < r(n - 1)
\end{cases}
\]

\[
t(n) = \begin{cases} 
|c(n) - r(n - 1)| & |c(n) - r(n - 1)| < 1 \\
1 & |c(n) - r(n - 1)| > 1
\end{cases}
\]

\(d\) lies between 0 and 1 when the current LTAS value is greater than the previous curve value and between 0 and -1 when the opposite is the case.

\(f_{inc}(n)\) is the function that determines how tightly the curve is fitted to the current LTAS value \(c(n)\) when the LTAS increases in amplitude compared to the previous curve value \(r(n - 1)\). \(f_{inc}(n)\) is calculated as follows:

\[
f_{inc}(n) = \begin{cases} 
1 & , \text{for } n \leq F_0 \\
x^{12} & , \text{for } n > F_0
\end{cases}
\]

\(F_0\) is the signal’s fundamental frequency. \(i_m\) is the index of the mid point between \(F_0\) and the highest frequency in the LTAS. \(n\) is the current frequency bin, \(x\) is a vector of integers:

\[
x = [-m \ldots 0 \ldots m],
\]

where \(m\) is half the LTAS vector length (from the fundamental frequency upwards) and

\[
x(n) = x(n - 1) + 1.
\]

The power of 12 was established to be most useful for distinguishing groups 1 and 2 in chapter 10. Following the definition of \(f_{inc}\), the fit is tight up to the fundamental (\(f_{inc}\) is 1 up to this point). After this, the fit becomes gradually looser towards the centre of the spectrum. Lastly, the fit becomes tighter again.
In this way, peaks in the centre of the spectrum are most likely to lie above the curve and count towards mid-range spectral peakiness.

\( f_{dec} \) is calculated as follows:

\[
f_{dec}(n) = \begin{cases} 
1 & \text{for } n \leq F_0 \\
\frac{y^2 \cdot 0.5}{(i_n)^2} & \text{for } n > F_0 \text{ and if } \frac{y^2 \cdot 0.5}{(i_n)^2} < 0.5 \\
0.5 & \text{for } n > F_0 \text{ and if } \frac{y^2 \cdot 0.5}{(i_n)^2} \geq 0.5
\end{cases}
\]  

(A.6)

\( F_0 \) is the signal’s fundamental frequency. \( i_n \) is the number of points spanning \( F_0 \) to the highest frequency in the LTAS. \( n \) is the current frequency bin, \( y \) is a vector of integers:

\[
y = [0 \ldots v],
\]

(A.7)

where \( v \) is the LTAS vector length and

\[
y(n) = y(n - 1) + 1.
\]

(A.8)

Similarly to \( f_{inc} \), \( f_{dec} \) has a tight fit up to the fundamental of the sound (all values up to this point are set to 0.5). After that, the fit is straight away very loose (0) and then becomes gradually tighter towards higher frequencies again.

Again, the chosen values (0.5 and the power of 2) were established to be most useful for distinguishing groups 1 and 2 in chapter 10.

A3.4 Calculation of mid-range spectral peakiness

The distances of data points above the curve \( u(n) \), from the curve, are summed together, leading to a mid-range spectral peakiness value. A higher weighting is applied to data points that lie further away from the curve, using the equation

\[
u(n) = \left( \frac{c(n) - r(n)}{h} \right)^4, \text{ for } c(n) > r(n)
\]

(A.9)

The dynamic range \( h \) is defined as the absolute difference between the smallest and greatest LTAS amplitude values. The chosen exponent of 4 was found to distinguish groups 1 and 2 in chapter 10 with the highest success rate.

The added mid-range spectral peakiness resulting from spectral equalisation is calculated as the difference between the mid-range spectral peakiness of the equalized stimulus and the mid-range spectral peakiness of the reference. An
“accept value” (0.016) is deducted from the result, as a small degree of resonance increase was perceived as unobjectionable. When the equalized stimulus has notably higher peaks towards the middle of its spectrum than the reference, the measure has a positive value. Otherwise, the value is negative.

A4 Publications

The data resulting from this research project are available at the following repositories:

- Timbral Clarity Dumping Bias Dataset (Ch.4) doi: 10.5281/zenodo.21341
- Single Sound Clarity (Strings) Dataset (Ch.6) doi: 10.5281/zenodo.30599
- Single Sound Spectral Clarity Dataset (Ch.5, Ch.7-10) doi: 10.5281/zenodo.192342

In the following, publications related to the current research project are presented. For each publication, the chapter it relates to is shown.

A4.1 Peer reviewed conference papers

Work documented in chapter 6 has also been published in:


Work documented in chapter 4 has also been published in:


A4.2 Research dissemination activities

Work documented in chapter 6 has also been published in:

Appendices


A 4.3 Internal University seminars and events

Work documented in chapters 2—8 has also been published in:


Work documented in chapters 2—6 has also been published in:


Work documented in chapters 2—5 has also been published in:

- Hermes, K., Brookes, T., Hummersone, C., “Towards measuring and modelling the perceived quality of music mixes”, PhD confirmation report, Surrey University, Guildford, UK, October 2014. This presentation relates to the findings in chapters 2—5.

Work documented in chapters 2—4 has also been published in:

- Hermes, K., Brookes, T., Hummersone, C., “The perceived quality of music mixes: literature review and first experiments”, Seminar day presentation, Surrey University, Guildford, UK, May 2014. This presentation relates to the findings in chapters 2—4.

Work documented in chapters 1 and 2 has also been published in:

References


References


Clark, R., 2011, Mixing, recording, and producing techniques of the pros, 2nd ed. Course Technology. Cengage Learning, Boston, USA.


References


Francombe, J., Mason, R., Dewhirst, M., Bech, S., 2013, 'Modelling listener distraction resulting from audio-on-audio interference.' In: Meetings on Acoustics 19, Montreal, Canada.


References


Helmholtz, H. von, 1863, Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. F. Vieweg, Braunschweig, Germany.


References


References


References


Reed, D., 2000, ‘A perceptual assistant to do sound equalization.’ In: 5th International Conference on Intelligent User Interfaces, IUI ’00. New York, USA.

References


Ronen, Y., 2015, 'Vocal Clarity in the Mix: Techniques to Improve the Intelligibility of Vocals.' In: Audio Engineering Society 139th Convention. New York, USA.


References


Toole, F.E., 1985, 'Subjective measurements of loudspeaker sound quality and listener performance.' *Journal of the Audio Engineering Society*, vol. 33, 1/2: 2–32.


299
References


