



## **University of Huddersfield Repository**

Fenton, Steven Michael

Audio Dynamics - Towards a Perceptual Model of Punch

### **Original Citation**

Fenton, Steven Michael (2017) Audio Dynamics - Towards a Perceptual Model of Punch. Doctoral thesis, University of Huddersfield.

This version is available at <http://eprints.hud.ac.uk/id/eprint/32629/>

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: [E.mailbox@hud.ac.uk](mailto:E.mailbox@hud.ac.uk).

<http://eprints.hud.ac.uk/>

# **AUDIO DYNAMICS – TOWARDS A PERCEPTUAL MODEL OF PUNCH**

**STEVEN MICHAEL FENTON**

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Huddersfield  
School of Computing and Engineering,  
Queensgate,  
Huddersfield,  
HD1 3DH.

March 2017

## Copyright statement

- i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Huddersfield the right to use such copyright for any administrative, promotional, educational and/or teaching purposes.
- ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the University Library. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.
- iii. The ownership of any patents, designs, trademarks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions

## Abstract

This thesis discusses research conducted towards the development of an objective model that predicts punch in musical signals. Punch is a term often used by engineers and producers when describing a particular perceptual sensation found in produced music. Music is often characterised by listeners as being punchier yet the term is subjective, in terms of its meaning and the subsequent auditory effect on the listener. An objective model of punch would therefore prove useful for both music classification purposes and as a possible further metric that could be employed in music production and mastering metering tools.

The literature reviewed within this body of work encompasses both subjective and objective audio evaluation methods in addition to low-level signal extraction and measurement techniques. The review concludes that whilst there has been a great deal of work in the area of semantic description and audio quality measurement, low-level analysis with respect to the perception of punch remains largely unexplored.

The project was completed in a number of phases each designed to investigate the perceptual effects resulting from manipulation of test stimuli. The rationale behind this testing was to establish the key low-level descriptors relating to the punch attribute with the aim of producing a final objective and perceptually based model. The listening tests in each phase were conducted according to the ITU-R BS 1534-1 recommendation.

In producing an objective model for the prediction of punch, listener perception to the attribute shows a strong correlation to the signal onset times, octave frequency band, signal duration and dynamic range. The punch measure obtained using the model is named PM95, where 95 indicates the upper percentile used in the measurement.

Secondary measures were also obtained as a result of the iterative approach adopted. These are Inter-Band-Ratio (IBR), Transient to Steady-state Ratio (TSR) and Transient to Steady-state Ratio+Residual (TSR+R). These measures are useful in quantifying overall audio quality with respect to its dynamic range across frequency bands in addition to being a more reliable metric for defining the overall compression being applied to a piece of music. In addition, the latter two measures proposed may be useful in highlighting perceptual masking artefacts.

The completed perceptual punch model was validated using the scores obtained from a large scale and independently conducted forced pairwise comparison test using expert listeners and a wide range of musical stimuli. From the results obtained, the PM95 measure showed a ‘very strong’ positive correlation with listener punch perception. Both  $r$  and  $\rho$  coefficients (0.849 and 0.833) being significant at the 0.01 level (2-tailed). The PM95M measure, which is the PM95 measure divided by the mean value of punch frames also correlated very well with the perceptual punch scale having both  $r$  and  $\rho$  coefficients (0.707 and -0.750) being significant at the 0.05 level (2-tailed).

A real-time implementation of the punch model (and other measures proposed in this thesis) could be utilised as extensions to the metrics currently being used in Music Information Retrieval.

## Table of Contents

<b>Chapter 1 Introduction .....</b>	<b>12</b>
<b>1.1 Background to the research.....</b>	<b>12</b>
<b>1.2 Scope and aims of the thesis.....</b>	<b>13</b>
<b>1.3 Key objectives.....</b>	<b>14</b>
<b>1.4 Structure of the thesis document.....</b>	<b>17</b>
<b>1.5 Novelty of this work.....</b>	<b>19</b>
<b>Chapter 2 Background review and terminologies .....</b>	<b>20</b>
<b>2.1 Research outline.....</b>	<b>21</b>
<b>2.2 Subjective assessment of audio attributes .....</b>	<b>22</b>
<b>2.3 Objective assessment of audio attributes.....</b>	<b>24</b>
2.3.1 ITU-Recommendation BS.1387-1 (1998) .....	26
<b>2.4 Audio semantics .....</b>	<b>30</b>
<b>2.5 The punch attribute .....</b>	<b>32</b>
<b>2.6 Music information retrieval.....</b>	<b>35</b>
2.6.1 Basic low-level feature examples .....	36
2.6.2 Additional Low-Level Descriptors (LLDs) .....	39
<b>2.7 Loudness .....</b>	<b>44</b>
2.7.1 ITU-Recommendation BS.1770-4 (2015) .....	49
<b>2.8 Audio dynamics.....</b>	<b>54</b>
<b>2.9 Compression, dynamics and audio quality.....</b>	<b>56</b>
<b>2.10 Audio dynamics measurement methods .....</b>	<b>58</b>
<b>2.11 Signal transients.....</b>	<b>61</b>
<b>2.12 Summary .....</b>	<b>63</b>
<b>Chapter 3 Objective measurement of music quality using multiband dynamic range analysis.....</b>	<b>66</b>
<b>3.1 Test methodology .....</b>	<b>67</b>
3.1.1 Biasing .....	69
3.1.2 Stimuli.....	69
3.1.3 Test subjects.....	71
<b>3.2 Discussion .....</b>	<b>72</b>

<b>3.3</b>	<b>Dynamic Range Analysis.....</b>	<b>74</b>
3.3.1	Wideband dynamic range measurement .....	75
3.3.2	Multiband dynamic range .....	76
3.3.3	Inter-Band Ratio (IBR) .....	78
<b>3.4</b>	<b>Conclusions.....</b>	<b>79</b>
<b>Chapter 4 Inter-Band Ratio and music quality perception. ....</b>		<b>80</b>
<b>4.1</b>	<b>Test Methodology.....</b>	<b>80</b>
4.1.1	Subjective listening test .....	82
4.1.2	Stimuli.....	82
<b>4.2</b>	<b>Results .....</b>	<b>83</b>
4.2.1	Listening test results .....	83
4.2.2	Objective results.....	85
<b>4.3</b>	<b>Discussion of results.....</b>	<b>87</b>
<b>4.4</b>	<b>Conclusions.....</b>	<b>90</b>
<b>Chapter 5 Profiling of punch and clarity using Inter-Band Ratio .....</b>		<b>91</b>
<b>5.1</b>	<b>Clarity and punch in music production.....</b>	<b>91</b>
<b>5.2</b>	<b>Temporal Inter-Band Ratio .....</b>	<b>94</b>
<b>5.3</b>	<b>Method of testing .....</b>	<b>95</b>
5.3.1	Subjective testing .....	95
5.3.2	Objective testing .....	96
5.3.3	Stimuli.....	97
<b>5.4</b>	<b>Results .....</b>	<b>104</b>
<b>5.5</b>	<b>Discussion of results.....</b>	<b>109</b>
<b>5.6</b>	<b>IBR statistical output.....</b>	<b>113</b>
<b>5.7</b>	<b>Conclusions.....</b>	<b>116</b>
<b>Chapter 6 Elicitation and grading of punch in music .....</b>		<b>117</b>
<b>6.1</b>	<b>Method of testing .....</b>	<b>117</b>
6.1.1	Elicitation exercise.....	118
6.1.2	Subjective testing .....	120
6.1.3	Objective measurement.....	121
<b>6.2</b>	<b>Subjective listening test results.....</b>	<b>121</b>
6.2.1	Verbal punch descriptors .....	122
6.2.2	Statistical analysis.....	122
<b>6.3</b>	<b>Objective measurement results .....</b>	<b>125</b>
<b>6.4</b>	<b>Discussion of results.....</b>	<b>128</b>

6.5	Conclusions.....	134
<b>Chapter 7 Hybrid multiresolution analysis of punch in music.....</b>		<b>136</b>
7.1	Sines, transients and residuals.....	137
7.2	Source separation.....	138
7.3	Fast onset detection method.....	138
7.4	Implemented analysis model.....	139
7.4.1	Multiresolution analysis.....	141
7.4.2	Separation of components.....	143
7.5	Analysis parameters .....	146
7.6	Results and discussion .....	149
7.7	Conclusion .....	154
<b>Chapter 8 Towards a perceptual model of punch .....</b>		<b>155</b>
8.1	Noise burst listening test .....	155
8.2	Loudness normalisation .....	156
8.3	Noise burst test results.....	160
8.4	Test analysis and model parameters.....	162
8.5	Punch model implementation .....	166
8.6	Model output .....	167
8.7	Statistical output .....	172
8.8	Conclusions.....	176
<b>Chapter 9 Validation of the punch model .....</b>		<b>177</b>
9.1	Overview of the objective measures.....	177
9.2	Experimental design and listening conditions.....	178
9.3	Stimuli.....	179
9.4	Subjective test results .....	181
9.5	Rank score and between sample significance testing .....	182
9.6	Bradley-Terry-Luce model .....	183
9.7	Model output correlation analysis.....	186
9.8	Model validation conclusion .....	191
<b>Chapter 10 Conclusions and future work.....</b>		<b>192</b>
10.1	Main research findings.....	192
10.2	Further work.....	195

## List of Figures

FIGURE 1 - MEASUREMENT MODEL DIFFERENCES (INTRUSIVE – TOP, NON-INTRUSIVE – BOTTOM) .....	25
FIGURE 2 - BS.1770 LOUDNESS MODEL OUTLINE.....	49
FIGURE 3 – LOUDNESS K-WEIGHTED FILTER.....	51
FIGURE 4 – AUDIO DYNAMICS .....	55
FIGURE 5 - AUDIO TRANSIENT.....	62
FIGURE 6 - MODIFIED MUSHRA INTERFACE .....	68
FIGURE 7 - MEAN SUBJECT SCORES VS. MAXIMISATION LEVEL .....	72
FIGURE 8 - COMBINED MSS VS. MAXIMISATION LEVEL .....	73
FIGURE 9 - MEAN DYNAMIC RANGE VS. MAXIMISATION LEVEL .....	75
FIGURE 10 - Ex1 MDR VS. MAXIMISATION LEVEL.....	77
FIGURE 11 - Ex2 MDR VS. MAXIMISATION LEVEL.....	77
FIGURE 12 - Ex3 MDR VS. MAXIMISATION LEVEL.....	77
FIGURE 13 – INTER-BAND-RATIO VS. MAXIMISATION LEVEL .....	79
FIGURE 14 - MEAN SUBJECT SCORES .....	83
FIGURE 15 - INTER-BAND RATIO MEASUREMENTS.....	86
FIGURE 16 - SUBJECTIVE & OBJECTIVE RANK ORDER COMPARISON .....	87
FIGURE 17 - NORMALISED SUBJECTIVE & OBJECTIVE SCORES VS. EXCERPT .....	88
FIGURE 18 - SUGABABES INTRODUCTION TIME DOMAIN .....	98
FIGURE 19 - SUGABABES VERSE WITH DRUMS TIME DOMAIN .....	99
FIGURE 20 - SUGABABES BREAKDOWN TIME DOMAIN .....	100
FIGURE 21 - NICKELBACK INTRO TIME DOMAIN .....	101
FIGURE 22 - NICKELBACK BREAKDOWN TIME DOMAIN .....	102
FIGURE 23 - NICKELBACK VERSE WITH DRUMS TIME DOMAIN .....	103
FIGURE 24 - NICKELBACK INTRO - IBR VS. SUBJECTIVE.....	104
FIGURE 25 - NICKELBACK VERSE WITH DRUMS - IBR VS. SUBJECTIVE.....	105
FIGURE 26 - NICKELBACK BREAKDOWN - IBR VS. SUBJECTIVE.....	105
FIGURE 27 (A)/27(B) SUGABABES INTRO - IBR VS. SUBJECTIVE .....	106
FIGURE 28(A)/28(B) SUGABABES VERSE WITH DRUMS - IBR VS. SUBJECTIVE .....	107
FIGURE 29(A)/29(B) - SUGABABES BREAKDOWN – IBR VS. SUBJECTIVE(WITH AND WITHOUT THRESHOLD) .....	108
FIGURE 30 - NICKELBACK INTRO IBR (400MS AND 3S WINDOW SIZES) .....	111
FIGURE 31 - IBR HISTOGRAM - NICKELBACK INTRO VS. SUGABABES INTRO.....	113
FIGURE 32 - IBR PERCENTILE NICKELBACK INTRO VS. SUGABABES INTRO.....	114
FIGURE 33 - IBR PERCENTILE NICKELBACK VS. SUGABABES.....	115
FIGURE 34 - TEST INTERFACE WAVE SHAPER .....	118
FIGURE 35 - EXAMPLE WAVESHAPE SETTING.....	119
FIGURE 36 - MODIFIED MUSHRA INTERFACE .....	120

FIGURE 37 - SOURCE 1 (INSTANTANEOUS ATTACK) - MPS VS. FILE .....	121
FIGURE 38 - SOURCE 2 - MPS VS. FILE .....	122
FIGURE 39 -SPECTRAL CENTROID OF SOURCE 1 KICK DRUMS VS. TIME.....	126
FIGURE 40 -SPECTRAL FLUX OF SOURCE 1 KICK DRUMS VS. TIME .....	127
FIGURE 41 -SPECTRAL SKEWNESS OF SOURCE 1 KICK DRUMS VS. TIME.....	127
FIGURE 42 -SPECTRAL SPREAD OF SOURCE 1 KICK DRUMS VS. TIME .....	128
FIGURE 43 - SOURCE 1 - RANK SCORED .....	129
FIGURE 44 - SOURCE 2 - RANK SCORED .....	129
FIGURE 45 -SIGNAL INTENSITY VS. TIME .....	132
FIGURE 46 - ANALYSIS MODEL.....	139
FIGURE 47 - FILTERBANK OF CASCADED QMF FILTERS.....	140
FIGURE 48 - MULTIREOLUTION STFT OF 'ANIMAL' WAV. ....	142
FIGURE 49(A) TRANSIENT (B) STEADY STATE MEDIAN FILTERING .....	145
FIGURE 50(A) TSR AND 50(B) TSR+R VS. TIME .....	149
FIGURE 51 - TRANSIENT INTENSITY SUMMATION VS. TIME.....	150
FIGURE 52 - SPECTRAL CENTROID (A) TRANSIENT (B) STEADY-STATE AND (C) OVERALL.....	151
FIGURE 53 -SPECTRAL WEIGHTING FILTER (AND INVERSE AS DOTTED ).....	156
FIGURE 54 - OCTAVE BAND FILTER RESPONSES .....	157
FIGURE 55 - LOUDNESS COMPENSATOR .....	158
FIGURE 56 - TEST INTERFACE - PUNCH PERCEPTION TEST.....	159
FIGURE 57 - MEAN PUNCH SCORES VS. OCTAVE BAND PER ONSET.....	160
FIGURE 58 - PUNCH SCORE 0MS ONSET NOISE BURST .....	160
FIGURE 59 - PUNCH SCORE 5MS ONSET NOISE BURST.....	161
FIGURE 60 - PUNCH SCORE 10MS ONSET NOISE BURST .....	161
FIGURE 61 - PUNCH SCORE 20MS ONSET NOISE BURST .....	161
FIGURE 62 - PUNCH SCORE 60MS ONSET NOISE BURST .....	161
FIGURE 63 - MODEL VS. SUBJECTIVE RESULTS .....	165
FIGURE 64 - PUNCH MODEL DIAGRAM .....	166
FIGURE 65 - MEASUREMENT OF NOISE BURSTS (PROGRESSIVE OCTAVE BANDS) 100MS BLOCK SIZE .....	167
FIGURE 66 - MEASUREMENT OF NOISE BURSTS (WEIGHTED PROGRESSIVE OCTAVE BANDS) 100MS BLOCK SIZE .....	168
FIGURE 67 - MEASUREMENT OF NOISE BURSTS USING STANDARD 400MS MOMENTARY LOUDNESS MODEL.....	168
FIGURE 68 - MEASUREMENT OF FULL SCALE & FULL SPECTRUM NOISE BURST USING THE PUNCH MODEL .....	169
FIGURE 69 - MEASUREMENT OF BILLY JEAN SAMPLE, USING STANDARD 400MS MOMENTARY LOUDNESS MODEL .....	170
FIGURE 70 - MEASUREMENT OF BILLY JEAN SAMPLE, USING 100MS PUNCH MODEL (WEIGHTED) .....	170
FIGURE 71 - MEASUREMENT OF BILLY JEAN SAMPLE, USING 100MS PUNCH MODEL (WEIGHTED WITH ONSETS) .....	171
FIGURE 72 - HISTOGRAM OF PUNCH SCORES DETECTED IN BILLY JEAN SAMPLE.....	172
FIGURE 73 - PERCENTILE PLOT OF PUNCH SCORES DETECTED IN BILLY JEAN SAMPLE. ....	173
FIGURE 74 - HISTOGRAM PLOT OF PUNCH SCORES DETECTED IN RAGE AGAINST THE MACHINE SAMPLE.....	174

FIGURE 75 - PERCENTILE PLOT OF PUNCH SCORES DETECTED IN RAGE AGAINST THE MACHINE SAMPLE. ....	174
FIGURE 76 - PERCENTILE PLOT OF PUNCH SCORES DETECTED IN THE AMBIENT SAMPLE. ....	175
FIGURE 77 - PERCENTILE PLOT OF PUNCH SCORES DETECTED IN AMBIENT SAMPLE.....	175
FIGURE 78 - BTL MODEL OUTPUT BASED ON PAIRWISE COMPARISON DATA.....	184
FIGURE 79 - BTL MODEL OUTPUT BASED ON 'SIGNIFICANT' PAIRWISE COMPARISON DATA. ....	184
FIGURE 80 - CORRELATION MATRIX OF ALL TESTED MEASURES.....	188

## List of Tables

TABLE 1 - PEAQ MODEL OUTPUT VARIABLES - BASIC .....	27
TABLE 2 - PEAQ MODEL OUTPUT VARIABLES - ADVANCED .....	28
TABLE 3 - 2 WAY ANOVA TEST .....	74
TABLE 4 - THREE BAND FILTER CORNER FREQUENCIES .....	76
TABLE 5 - RANK SCORE ORDER OF EXCERPTS .....	83
TABLE 6 - ANOVA TABLE .....	85
TABLE 7 - RANK SCORE ORDER BASED ON IBR .....	86
TABLE 8 - DYNAMIC RANGE AND IBR MEASURES .....	88
TABLE 9 - PEARSON CORRELATION PER EXCERPT .....	109
TABLE 10 - PEARSON CORRELATION OF WAVE-SHAPER PARAMETERS TO PUNCH SCORE .....	123
TABLE 11 - SPECTRAL CENTROID (1024-POINT FFT) .....	130
TABLE 12 - INTENSITY RATIO OF SOURCE 1 REFERENCE .....	131
TABLE 13 - INTENSITY RATIO OF SOURCE 1 REFERENCE USING WAVE-SHAPER BANDS .....	131
TABLE 14 - RHYTHM STRENGTH .....	132
TABLE 15 - MEASUREMENT FREQUENCY BANDS .....	141
TABLE 16 - MEASUREMENTS PERFORMED .....	148
TABLE 17 - SUMMARY FOR PUNCH SCORE MODEL .....	163
TABLE 18 - ANOVA SUMMARY FOR PUNCH SCORE MODEL .....	163
TABLE 19 - COEFFICIENTS FOR PUNCH SCORE MODEL .....	163
TABLE 20 - PUNCH MODEL NAMING AND DESCRIPTION .....	173
TABLE 21 - STIMULI USED IN THE MODEL VALIDATION TEST .....	180
TABLE 22 - FORCED-PAIRWISE TEST SCORES .....	181
TABLE 23 - CORRELATIONS OF THE TESTED MEASURES WITH THE PERCEPTUAL PUNCH SCALE .....	187

## **Dedications and Acknowledgements**

There are a number of people I'd like to thank and acknowledge for their help, support and encouragement and inspiration during the research process and creation of this thesis.

Thank you to my supervisors Dr Jonathan Wakefield and Dr Hyunkook Lee for guidance throughout what at times has been a difficult process. I have also benefitted from critical comments and the many discussions that ensued, which have improved the overall quality of this thesis.

I am grateful to the University of Huddersfield for supporting this thesis and also for providing financial support in order to present my work at numerous AES conferences. I hope this continues into the future as my research path grows.

I thank Esben Skovenborg for the discussions regarding both loudness and dynamics perception, in addition to experimental support.

Thank you to all the listeners who participated in my listening tests and also AES members, with whom I discussed ideas with at conferences throughout the world. Also, a special thanks to Dr Bruno Fazenda, for starting me off on this great journey and for the lively and thought provoking discussions thereafter.

Thanks go to my family (Helen, Kai, Charlie, Meg & Monty) for the support and love, which has been so critical at home during this long process. Finally, my Mum and Dad and sister, to whom this work is dedicated. xx

## Glossary of Abbreviations

<p><b>AHD</b> – Audio Harmonic Descriptor</p> <p><b>ANOVA</b> – Analysis Of Variance</p> <p><b>BAQ</b> – Basic Audio Quality</p> <p><b>C50</b> – Early To Late Sound Ratio</p> <p><b>CCR</b> – Comparison Category Rating</p> <p><b>CF</b> – Crest Factor</p> <p><b>CQS</b> – Continuous Quality Scale</p> <p><b>dB</b> – Decibel</p> <p><b>DRC</b> – Dynamic Range Compression</p> <p><b>EDT</b> -Early Decay Time</p> <p><b>ERB</b> – Equivalent Rectangular Bands</p> <p><b>FFT</b> – Fast Fourroer Transform</p> <p><b>FS</b> – Digital Full Scale</p> <p><b>HLSD</b> – High Level Sample Density</p> <p><b>IBR</b> – Inter-Band Ratio</p> <p><b>IBR_diff</b> – Inter-band Ratio Difference</p> <p><b>LAT</b> – Log Attack Time</p> <p><b>LDR</b> – Loudness Dynamic Range</p> <p><b>LF, MF &amp; HF</b> – Low, Mid and High Frequency respectively</p> <p><b>LFE</b> – Low Frequency Enhancement</p> <p><b>LKFS</b> – Loudness K-weighted Full Scale</p> <p><b>LLD</b> – Low Level Descriptor</p> <p><b>LMIBR</b> – Low-Mid Inter-Band Ratio</p> <p><b>LRA</b> – Loudness Range</p> <p><b>LU</b> – Loudness Units</p> <p><b>LUFS</b> – Loudness Unit Full Scale</p> <p><b>MDR</b> – Multiband Dynamic Range</p> <p><b>MIR</b> – Music Information Retrieval</p> <p><b>MOS</b> – Mean Output Score</p> <p><b>MOV</b> – Model Output Variable</p>	<p><b>MPEG</b> – Moving Picture Expert Group</p> <p><b>MPS</b> – Mean Punch Score</p> <p><b>MSS</b> – Mean Subject Score</p> <p><b>MUSHRA</b> – Multiple Stimulus Hidden Reference and Anchor</p> <p><b>ODG</b> – Objective Difference Grade</p> <p><b>PEAQ</b> – Perceptual Evaluation of Audio Quality</p> <p><b>PLR</b> – Peak To Loudness Ratio</p> <p><b>PM95</b> – Punch Measure (95<sup>th</sup> Percentile)</p> <p><b>PM95M</b> – Punch Measure (95<sup>th</sup> Percentile / Mean)</p> <p><b>PMF</b> – Probability Mass Function</p> <p><b>QMF</b> – Quadrature Mirror Filter</p> <p><b>RLB</b> – Revised Low Frequency B Weighting Filter</p> <p><b>RMS</b> – Root Mean Square</p> <p><b>RT60</b> – Reverb Time (to decay to -60dB)</p> <p><b>SDG</b> – Subjective Difference Grade</p> <p><b>SNR</b> – Signal To Noise Ratio</p> <p><b>STFT</b> – Short Time Fourier Transform</p> <p><b>THD</b> – Total Harmonic Distortion</p> <p><b>TR, SS &amp; R</b> – Transient, Steady State and Residual respectively.</p> <p><b>TSR</b> – Transient to Steady State Ratio</p> <p><b>TSR+R</b> – Transient to Steady State Ratio + Residual</p> <p><b>TWC</b> – Temporal Weighting Coefficient</p> <p><b>VIF</b> – Variance Inflation Factors</p> <p><b>WDR</b> – Wideband Dynamic Range</p>
--	---

## **Chapter 1 Introduction**

In music production, sound-engineers and producers employ techniques to deliberately colour or enhance the completed piece in order to achieve what is deemed release quality material. What one deems as release quality is largely subjective and due to the proliferation of self-publication and accessibility through download, there now exists a vast amount of musical data that is largely uncatergorised both in terms of its overall quality and underlying audio and musical attributes.

Whilst it is possible to include metadata in the form of short data fields to categorise musical data, for example, artist name, composer, year of release and in some cases, more subjective elements such as mood and genre the information is very limited with respect to its scope in either defining overall quality or contributing to its measurement. Further to that, the inclusion of metadata is very labour intensive and often just not done.

### **1.1 Background to the research**

Recently there has been a large amount of interest in the field of Music Information Retrieval (MIR) in a bid to enable large scale indexing, organising and navigation of digital music. In tandem with this work, many of the low-level descriptors utilised in the MIR process are being examined to establish if any of the descriptors show a correlation with overall mix quality (De Man et al. 2014).

Determining overall mix quality of a piece of music is a complex process; some might argue that it could be deemed entirely subjective. However, taking a very simple case where a piece of audio is directly compared against a known ‘good’ reference, it may be easier to establish whether the stimulus under test is good or not. The attributes that are compared are important in this case, thus enabling such things as genre, mood and artist to be completely ignored. An example of this could be the objective comparison of an attribute such as ‘brightness’, for many years being equated to the ‘spectral centroid’

objective measure (Grey & Gordon, 1978). This is a very simplistic but valid model, if you were looking for pieces of music that exhibited the same brightness as the reference.

In order to establish a more accurate predictor of overall mix quality, a larger number of low-level attributes would need to be combined and/or new attributes found. In addition, these attributes need to be rigorously tested to see if perceptual correlation can be found. Due to this complexity, a ‘single’ objective metric that categorises mix quality is somewhat of a ‘holy grail’ within the music industry but continued research to establish new low-level descriptors may allow a higher resolution of music categorization to be established.

## **1.2 Scope and aims of the thesis**

As outlined in the previous section, additional low-level descriptors for use in MIR and audio quality measurement are desirable. As the title of this thesis infers, this work presents research that aims to establish a new descriptor for use in music categorisation and quality measurement, that of punch. A number of novel objective measures are proposed and explored in this work that indicate a correlation with subjective scores obtained from listening tests in terms of both mix quality and punch perception.

Through conducting the literature review it was apparent that whilst the punch semantic descriptor is freely used, no formal perceptual based studies of punch exist. This study will aim to provide both a formal definition of punch and also seek to identify measurable attributes that will inform an objective model of punch perception. The model proposed can be used as an additional search metric in MIR, a predictor in objective metering tools and as an additional variable in the prediction of audio quality.

It is the author’s hypothesis, and one that will be proven within this thesis, that punch can be described as a short period of significant change in power in a piece of music or performance. The magnitude of change is associated with and proportional to the signal dynamics that are present and thus, productions that do not possess any transient or dynamic attribute cannot possess punch. The onset of the transient present across octave

bands affects the listener perception of punch, with the lowest octave attributing the most punch as the onset is decreased and vice-versa. Punch is therefore related to transient change and the energy density (summation across frequency bands) occurring at a particular moment in time and duration.

Further to the above hypothesis, dynamic change in particular frequency bands may contribute to the perception of punch perceived by the listener and the overall average loudness level inherently affects this at that time (Moore, 2004). Thus, by mapping the perception of the punch attribute to objectively measured key attributes of the signal, one can produce a metric that could be utilised in music production and classification.

Analysis of existing measurement models found in the literature review revealed that they are not suitable for predicting the punch attribute. Indeed, at the time of writing there is not a complete model for predicting overall mix quality although research in the area is very active.

### **1.3 Key objectives**

Based upon the previously stated hypothesis, the key objectives of each experiment are now outlined. The subsequent description explains the steps involved in achieving each objective.

- *To investigate the effects of dynamic range reduction with respect to listener perception.*

Signal dynamics are inherently related to the dynamic range of the material being measured. An important aim of this project was to establish the effects of dynamic range reduction from a quality point of view and also to elicit parameters relating to signal dynamics that could be isolated with respect to the punch attribute.

Existing methods of dynamic range measurement and the effects of dynamic range reduction were critically analysed. Additional measures were explored which incorporate multiband filtering and dynamic range correlation calculations of the audio under test.

- *To investigate salient effects of frequency change on both perceived clarity and punch within a recording.*

This study continued work outlined in the previous key objective and measures were explored based on those reviews.

From a quality measurement perspective, this work investigated the effect of transient change in relation to both clarity and punch perceived by the listener. This is an important step in establishing the key frequency bands of interest, what effects the magnitude of change in signal dynamic may have and the resulting perceptual effects upon overall quality, punch and clarity.

- *To elicit parameters that could be useful in producing an objective model of punch.*

It is hypothesised that punch can be described as a short period of significant change in power in a piece of music or performance and that dynamic change in particular frequency bands may contribute to the perception of punch perceived by the listener. Consequently, the mapping of objectively measured attributes of the signal and their correlation with perceived punch was explored.

In order to achieve these goals, further literature was reviewed and subjective listening test data collected. This work investigated the manipulation of temporal and frequency based parameters of stimuli whilst monitoring the effects upon perceived punch by expert listeners. From this work, salient parameters associated with the punch attribute are identified which may prove useful in the development of the objective model.

- *To investigate methods that will enable salient features to be extracted.*

A musical signal is a complex collection of tonal and non-tonal components. In order to create a perceptually motivated objective model, methods of salient feature extraction were investigated. The aim of this work was to establish both a reliable and relatively low overhead method of signal decomposition and to investigate resulting features and parameters that are measureable following this decomposition.

- *To propose and validate an objective model for the prediction of the punch attribute.*

The final goal of this thesis is to present a possible model for the objective measurement of punch within a musical signal. The steps in achieving this were iterative and formed from the body of work presented.

This thesis concludes with a perceptually motivated objective model for the prediction of the punch attribute in a two channel audio file. The model was validated using a number of varied stimuli and subjective listening tests. The results of this study can be useful in establishing a more rigorous ‘overall mix quality measure’ in addition to being utilised in real-time implementation as a useful mix or mastering metering tool. Measures with reference to dynamic range are also proposed, some of which isolate and take into account specific features of the audio that are related to the perception of punch, namely the transient and steady state components. In order to achieve that goal, literature within the fields of MIR, subjective and objective assessment and signal separation were reviewed. All of these subject areas are interlinked and thus this body of research will serve as a further reference in these areas.

## **1.4 Structure of the thesis document**

This thesis describes the research undertaken to develop objective measurement techniques that relate to perceived punch in audio signals.

Chapter 2 aims to introduce the reader to the wider context of this research by outlining key terminologies and related subject areas. Audio semantics, MIR, Loudness, Signal Transients and Punch are all defined in this chapter. Since subjective and objective assessment of audio is an important part of this thesis, this area is also covered.

In addition to defining the meaning of audio dynamics within this chapter, a review of current methodologies relating to both dynamic range and ‘signal dynamics’ is included along with a review of current low-level descriptors and their relevance to this thesis.

Chapter 3 contains a discussion of the first experiment that was conducted to establish the correlation between perceived audio quality and dynamic range. This pilot test employed a controlled listening test whereby stimuli that had been manipulated with respect to overall dynamic range were presented to listeners. The results of the experiment are discussed and a new measure is proposed that incorporates multiband processing.

Chapter 4 represents a wider study into the multiband approach first introduced in Chapter 3. A controlled listening test is employed using a greater number of audio stimuli and a greater number of expert listeners. The experiment is described and the results are discussed. Finally, the limitations of the experiment are considered.

Chapter 5 outlines an experiment that utilised the measure described in the earlier experiments to profile audio stimuli with reference to listener perception of punch and clarity. Temporal measurements were collected and correlations are discussed with respect to data collected in a controlled listening test. It was concluded that the transient content

and dynamic range de-correlation between frequency bands could relate to higher subjective scores being given by the listeners when asked to judge punch and clarity.

Chapter 6 continues the work outlined in the previous chapter. A novel reverse elicitation test was employed to establish which components of an audio signal contribute greatly to the perception of punch. Expert listeners were asked to create audio samples that they perceived as having punch using a multi-band wave shaping process. The listeners then graded the generated punchy audio samples in a controlled listening test. Statistical analysis identified correlations between Mean Subject Scores and the parameters that created the punchy audio samples suggesting that an algorithm could be developed to objectively evaluate punch in produced music.

Chapter 7 describes an experiment to investigate the elements of the audio deemed to be of relevance in the preceding experiments. Signal separation was explored which enabled the transient components of the stimuli to be isolated and therefore measured independently. The experiment is described and the results are discussed along with some new measurement proposals. Finally, the limitations of the experiment are considered.

Chapter 8 defines a model for the objective measurement of punch. It details the elicitation of the model parameters through a controlled noise burst listening test. The tests are described and the results are discussed. A model is then proposed and differing output statistics are considered.

Chapter 9 contains the details of a controlled listening test that evaluates the punch model proposed in the preceding chapter. The model output correlation to subjective listener scores is evaluated alongside other existing measurement models which include the Inter-Band-Ratio measure (IBR) proposed in earlier chapters. The results of this experiment are discussed along with further work and limitations.

The main conclusions of this work are summarised in Chapter 10 and future work is identified based on this work.

## **1.5 Novelty of this work**

The extensive literature review undertaken throughout the body of this work indicated that there is no formal definition of the punch attribute. Along with the proposal of four novel objective measurements, this work formalises the definition of punch in terms of low-level features extracted from the audio under test.

The punch model (PM95) presented offers the ability to measure a perceptual parameter that was previously only able to be described subjectively by listeners. It shows a very strong correlation to the perceptual attribute.

The Inter-Band-Ratio (IBR) measure presented shows a stronger correlation to audio quality affected by dynamic range compression than existing dynamic range measures. In addition, the statistical output of this measure is shown have a moderate correlation to the perception of punch as graded by a panel of expert listeners.

Transient to Steady-state Ratio (TSR) and Transient to Steady-state Ratio+Residual (TSR+R) measurements are also presented. These measurements indicate the perceptual dynamics and masking within a piece of audio.

The use of signal separation within the TSR, TSR+R and PM95 measures is, to the knowledge of the author, a method that has not been employed elsewhere in an audio measurement context. This process enables the extraction and measurement of individual components of interest within the audio signal.

## **Chapter 2 Background review and terminologies**

The scope and aims of this thesis were defined in Section 1.2 which included a definition of the main hypothesis. The focus of the work is on the development of an objective model for predicting the punch attribute of a musical signal. The aim of this chapter is to therefore familiarise the reader with the wider context of the research, with respect to both its applicability and terminologies.

This thesis is related to the perceptual audio measurement of music, therefore, a definition of music followed by an outline of subjective and objective assessment schemes are provided in Sections 2.1, 2.2 and 2.3 respectively. Punch is an audio semantic so to familiarise the reader with both, audio semantics and the punch attribute are outlined in Sections 2.4 and 2.5. Audio semantics are used extensively in MIR therefore Section 2.6 defines MIR and defines some low level descriptors that can be used for information retrieval. Loudness, being an important metric in music characterisation and production, is included in its own Section 2.7.

Audio dynamics, in the context of this thesis, is defined in Section 2.8 and this leads into Section 2.9 which explores the link between dynamics, compression and the perception of quality. Differing methods have been proposed to measure the dynamic range within a piece of music and these are defined in Section 2.10. The hypothesis of this thesis declares that the onset will play a key role in the perception of the punch attribute, therefore a definition of signal transients is outlined in Section 2.11

## **2.1 Research outline**

In the context of this thesis, music is a musical performance or programmed sequence captured by a recording process and stored on a medium for later listening and enjoyment. Ever since the very first recordings were made we have strived to improve the quality of the recording and playback process (Feaster, 2008).

Over the decades, recording technology has improved (in particular in the digital domain) to such an extent that the signal path from capture to recording could be argued to be virtually transparent in terms of colouration of the original signal source. Of course, there are slight differences due to microphone responses, the performance of the pre-amplifier and the signal conversion, if applicable. These differences are either compensated for or exploited by the audio engineer in the production stages, to form what is termed ‘produced music’.

In general, a completed piece of music will be the sum (i.e. mix) of the product of a number of discrete processes and audio stems resulting in a stereo or multichannel audio file. The stems are formed by the recorded sound sources and the complexity of the individual sources is what we hear and describe as timbre. Timbre is what distinguishes them as different types of voices or musical instruments.

Therefore, music is a complex signal, carrying lots of information to the listener. It is made up of many different harmonic components of varying phases and magnitudes, in addition to both correlated and uncorrelated components such as noise. Informational elements of music include lyrics, pitch, rhythm and the sonic qualities of timbre and texture.

Overall quality can therefore not be attributed to a single metric or measurement; in fact, many of the metrics relating to audio system performance measurement are not applicable to the measurement of ‘overall music sound quality’. Total Harmonic Distortion (THD), for example, is referenced to a pure tone at the system input. Whilst it is possible to perform

conventional ‘system’ measurements on produced music to describe aspects of the audio under test, these do not correlate very well with the overall perception of quality by the listener (Boley et al. 2010). It is therefore important to identify new measures that are linked perceptually with how we hear music. This could lead to the possibility of an overall music sound quality score being established.

## **2.2 Subjective assessment of audio attributes**

The aim of this section is to outline current methodologies used in the assessment of audio attributes. A literature review of subjective audio assessment methodologies was necessary to establish the appropriate and best practise method of evaluation of the punch attribute described in this thesis. Objective assessment methods and existing measurements were also reviewed to establish whether attributes pertaining to punch had been investigated or relevant measures existed. This review also forms the basis for the development of new methods and measurement strategies proposed in this thesis.

Formal listening tests are regarded as the most reliable method for audio quality assessment and a number of methodologies have been established (Bech & Zacharov. 2006). The proliferation of such tests have, in the most part, been in response to a need to evaluate the quality of low bit rate CODECS (Stoll & Kozamernik, 2000; Marston & Mason, 1994) used in voice over internet, streaming technologies and the MP3 format for music distribution.

Audio is often perceptually encoded, such as in the MP3 format. Consequently, typical objective measures such as Signal-to-Noise ratio (SNR) may give varying and wide ranging results depending on and throughout the signal. However, there may in all these cases be no noise audible due to masking. The solution to this problem is therefore to ask listeners to evaluate the audio through listening tests (Bech & Zacharov, 2006).

Three major recommendations with regard to the subjective assessment of audio quality have been established. These are standardised as:

- *ITU-R (1994) BS.1116, developed primarily to evaluate small impairments in audio quality,*
- *ITU-R (2003) BS.1534-I, commonly referred to as MUSHRA and developed to evaluate intermediate impairments in audio quality and*
- *ITU-T (1990) P.800, primarily used to evaluate narrowband speech quality.*

Generally, these testing and measurement techniques are employed to establish audio quality in audio systems (such as CODECs) under test with respect to an original ‘untreated’ reference signal. The resulting index is named the subjective difference grade (SDG) which attempts to categorize the subjective audio quality.

Subjective listening tests are also of primary importance in establishing user preference for a particular auditory sensation. These grades or mean subject scores (MSS) can then be used to formulate objective models through analysis. For example, listeners could be asked to score a number of stimuli with reference to a particular sensation they experience such as ‘warmth’. The mean subject scores are collated and correlation testing with low-level attributes extracted from the stimuli can be performed to attempt to establish a link to this sensation.

Conducting listening tests for evaluating audio quality can be very time consuming and careful experimental design is paramount in the successful extraction of useable data. The testing is usually context dependent and it is often the case that listeners can differ in their ratings. These differences can correlate with both hearing threshold levels and age and therefore careful screening of subjects is important.

Generally, listeners with near normal hearing thresholds show the smallest individual variations and the closest agreement with each other (Toole, 1985). However, these types of test can be subject to errors through various forms of biasing (Zielinski & Rumsey, 2008)

and careful selection of experimental procedure, use of appropriate scales and use of anchor points amongst other things should be considered.

Subjective tests are by nature, based on listener or subject opinion. That said, they form an important role in establishing scales of response or user preference to a variety of auditory sensations and in turn the data can be used in the calibration of objective measurement models.

An example of subjective testing would be to ask a number of listeners to score stimuli with reference to a particular sensation they experience such as ‘warmth’. The Mean Subject Scores (MSS) are collated and correlation testing with respect to low-level attributes extracted from the stimuli can attempt to establish any particular link to this sensation. Subjective listening tests have been utilised to establish user responses in all of the experiments outlined in this thesis. The associated test methodologies are explained in each chapter as appropriate.

### **2.3 Objective assessment of audio attributes**

Objective assessment of audio can offer an alternative solution to subjective testing. For example, where a repeatable and automatic system of measurement and classification is required.

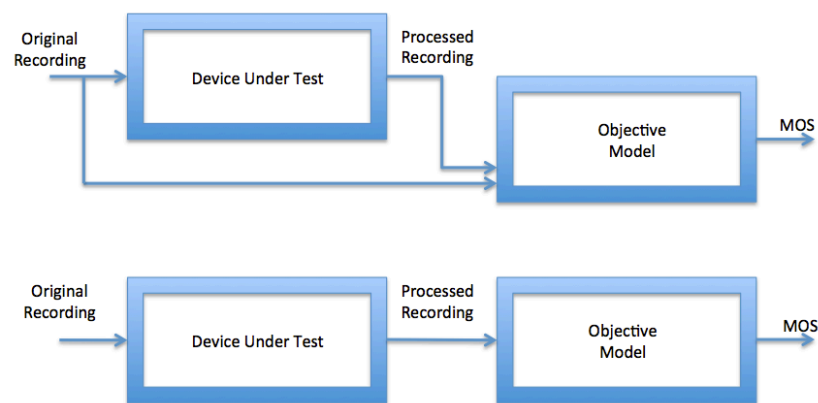
It can be loosely categorised by its application. For example, where automatic quality measurement is a requirement (CODEC or broadcast testing) or to allow automatic classification to take place e.g. MIR. Often a number of feature measurements are combined to form an overall measure or score to quantify the stimuli under test.

The key difference between objective measurement and its subjective counterpart is that, should the measure be repeated using the same set of stimuli, the same output would always result. In general, objective models or measures are based upon rule bases or scales that have been derived through extensive subjective testing.

It is important to this research to establish if there are any current models or feature measurements that may predict the perception of punch within stimuli and perhaps offer avenues of research that may bear fruit with respect to this attribute.

Objective audio tests can be categorised in two ways, ‘single ended’ (non-intrusive) and ‘double ended’ (intrusive) as described in ITU-T P.563 (2004). Outlines of these models are shown in Figure 1 - . The former category of test does not require a reference signal in order to output a quality score, however the model will need to incorporate an algorithmic rule base of some kind in order to offer valid results, either through heuristically derived means or by cumulative score methods.

Double-ended approaches might typically be employed if one were to be measuring the degradation of a particular audio process, such as the measurement of CODEC quality. The test stimuli can be directly compared to its unprocessed version and differences can easily be highlighted. Use of a single ended approach in this case would give somewhat meaningless results unless the rule base of the model were calibrated to that of the known reference.



**Figure 1 - Measurement Model Differences (Intrusive – Top, Non-intrusive – Bottom)**

Non-intrusive objective measurement models are difficult to develop due to the large number of different stimuli that need to be coped with. Their internal coefficients, that are

used to predict the model output variables (MOVs), need to be carefully calibrated with respect to the low-level feature set that is being extracted. This is usually done through lengthy subjective testing and evaluation with the goal of the model being able to predict an overall quality score based on its input stimuli alone, without influence or bias. The Mean Output Score (MOS) is representative of this and combines a number of MOVs in its derivation.

Whichever objective assessment method is utilised, a regression process is necessary to establish the accuracy of the objective model. Ideally, the output of the model should have a high correlation with the corresponding subjective experiments.

In order to address the need for automatic quality measurement of audio, a number of objective measures have been proposed. These attempt to predict the Basic Audio Quality (BAQ) from extracted features of the audio under test. Many of the techniques have been standardized as ITU-R BS.1387-1 (1998), otherwise known as PEAQ (Perceptual Evaluation of Audio Quality). The basic concept of making objective measurements within this recommendation is intrusively based methods (i.e. double ended).

### ***2.3.1 ITU-Recommendation BS.1387-1 (1998)***

ITU-R. BS.1387-1 (1998), otherwise known as PEAQ (Perceptual Evaluation of Audio Quality), is an example of an objective model that evaluates the audio quality differences caused by the presence of noise and/or distortions. Although intrusive, it is included in this thesis due to it utilising an artificial auditory system (ear model) either through the use of a filter bank or Fast Fourier Transform (FFT) and to highlight some of the low-level features it utilises. PEAQ combines a number of different model variables (MOVs) in order to compute the objective difference grade (ODG). The ODG indicates the BAQ of the stimuli on a continuous scale from -4 (very annoying) to 0 (imperceptible). In addition, the model outputs a Distortion Index (DI). The DI is a quality indicator like the ODG except for its higher sensitivity towards very low signal qualities.

The basic version of PEAQ combines 11 of the MOV's to calculate the ODG whilst the advanced version combines a further 5. These are shown in Table 1 and Table 2 respectively. The MOVs shown with subscript 'A' are based upon the filterbank model whilst those shown with subscript 'B' are based on the FFT model.

MOV	Description
WinModDiff1B	Windowed modulation difference. All modulation difference calculations relate to roughness and the temporal envelopes are derived from the auditory filter outputs.
AvgModDiff1B	Average modulation difference
AvgModDiff2B	Averaged modulation difference with emphasis on introduced modulations and modulation changes where the reference contains little or no modulations
RmsNoiseLoudB	RMS value of the perceived noise loudness.
BandwidthRefB	Bandwidth of the reference signal
BandwidthTestB	Bandwidth of the output signal of the device under test
TotalNMRB	Logarithm of the averaged total noise to mask ratio.
RelDistFramesB	Relative fraction of frames for which at least one frequency band contains a Significant noise component
MFPDB	Maximum of the probability of detection after low pass filtering
ADBB	Average distorted block, the logarithm of the ratio of the total distortion to the total number of severely distorted frames
EHSB	Harmonic structure of the error over time

**Table 1 - PEAQ Model Output Variables - Basic**

MOV	Description
RmsNoiseLoudAsym <sub>A</sub>	Loudness of the distortion
RmsModDiff <sub>A</sub>	Changes in modulation (related to roughness)
AvgLinDist <sub>A</sub>	Linear distortions (frequency response etc.)
Segmental NMR <sub>B</sub>	Noise-to-mask ratio
EHS <sub>B</sub>	Harmonic structure of the error over time

**Table 2 - PEAQ Model Output Variables - Advanced**

The inputs for the MOV calculations are derived from either an FFT or a filter bank output and are:

- *The excitation patterns for both test and reference signal.*
- *The spectrally adapted excitation patterns for both test and Reference Signal.*
- *The specific loudness patterns for both test and Reference Signal.*
- *The modulation patterns for both test and Reference Signal.*
- *The error signal calculated as the spectral difference between test and Reference Signal (only for the FFT-based ear model).*

A detailed breakdown of the model is presented by Thiede et al (2000) with each MOV calculation, model outline and validation explained. Looking at this model, it is apparent that its primary function is to quantify the overall perceptual noise in the form of the ODG output. The perceptual noise is primarily focused on that created by a device under test (DUT), for example an audio codec. The model is after all, based upon auditory perceptual thresholds relating to audio signal compression algorithms, mp3 for instance. Therefore, its use as a ‘measurement of perceived audio quality’ is limited somewhat to that application.

The use of both FFT based and filterbank ear models allows modelling of the human auditory response thus making the model perceptually based. Objective measurements

based on the human auditory system have existed since 1979 (Schroeder et al., 1979) and more recently, albeit in a much simpler implementation in the current loudness model specified in ITU-R BS.1770-4 (2015) as detailed in Section 2.7 of this thesis. The choice of the use of either the FFT and/or filterbank approach in modelling the auditory response is largely down to the need for real time processing.

The FFT size adopted in the PEAQ model utilises a 2048 window size with Hann windowing. This results in 1024 spectral bins and a temporal resolution of 21ms (at 48 kHz). A frame analysis (STFT) approach is adopted with an overlap of 50%.

This thesis discusses the development of non-intrusive models.

## 2.4 Audio semantics

Semantic terms are often used to describe music, whether it's a small section of music being listened to during production or listening to a released song in the car whilst driving down the road. Audio semantics attempt to categorise or give meaning to an audio stream.

As listeners, we make conscious and unconscious interpretations of what we hear based on the recognition of individual instruments and voices within a complex mix. Harmonic change, chord progression, emotional inference and melody are also often distinguishable to the listener. Factors such as individual listening taste, mood or hearing differences will have an impact on how two listeners may describe a piece of music or the underlying perceptual effects of that music.

Completed recordings may be deemed as '*clear*', '*defined*', '*punchy*' or '*highly polished*' by some listeners, on the other hand they may be referred to as '*woolly*', '*distorted*', '*poorly balanced*' or '*muddy*' by a different set of listeners.

These semantic descriptors are of course 'subjective'. However, they are frequently used and recognized within the audio industry and for the vast majority of engineers these descriptors are used to categorise the production of a piece of music even though many of them have no clear and defined perceptual relevance.

Artists also use these terms to describe their requirements during the recording, mixing and mastering stages. Obviously, if both parties hear different things but describe them using the same semantic term, this may cause issues. Similarly, but perhaps not as critical to the production process, both parties may hear the same perceptual effect but describe it with a different semantic term.

The main issue in both cases is that the semantic descriptor would not allow for a consistent qualitative measure to be established unless it had been rigorously tested with respect to its perceptual effect.

Therefore, the need to research and link semantic terms to underlying feature sets is paramount in aiding the development of both successful automatic transcription and classification tools along with perceptually motivated metering and analysis systems.

Automatic feature extraction would enable higher-level musical semantic representation of the underlying musical content (Tzanetakis, 2012). Both the features extracted, and in part the resulting semantic terms may form an objective descriptor of the music under test. The task of the researcher is to determine the correlation between musical features or characteristics and the semantic term that is being explored. Subjective listening tests are often employed to do this. Using too many parameters not related to semantic term can cause the situation where relevant parameters are masked under the noise of other less relevant ones. Conversely, use of too few parameters may cause a lower level of correlation.

The SAFE Project (Stables et al., 2015) is an example where higher-level parameters, such as plug-in control parameters are retrievable based on mappings to semantic descriptors. It represents a sort of top down approach to enable more intuitive control of the low-level parameters themselves. Users of the SAFE based plugins can save their parameter sets along with a semantic of their choice, e.g. *'punchy kick'*. This metadata is saved anonymously to a server along with a time-series matrix of audio features (Bullock, 2007), a static parameter space vector and a selection of optional metadata tags, such as age, location and production experience. The latter optional metadata tags were deemed to be statistically significant factors in the variance of semantic terminology between different user groups (Stables et al., 2015). Their work is motivated by the lack of statistically defined transferable semantic terms.

A semantically motivated gestural compressor (Wilson et al., 2015) is an example whereby audio semantics can greatly reduce the overall complexity of an audio processor whilst at

the same time offering the same timbral transformations without the need for extensive training.

Punch is the semantic word used to describe the perceptual attribute that forms the focus of this body of work. A detailed description will be provided in the following section.

## **2.5 The punch attribute**

The previous section covering audio semantics described the use of a wide vocabulary of terms to describe musical attributes with them often having perceptual relevance; for example, *warm*, *bright*, *soft* or *heavy*. As ‘punch’ can be classed a semantic term, it is important to outline the meaning within the context of this thesis. The following section will cover this.

Work to establish verbal descriptions and dimensions for some of these perceptual attributes has been extensively explored in previously published papers (Grey, 1977; Stepanek, 2006; Lakatos, 2000). Early work by Freed (1990) and others, focusing on the perception of mallet hardness and noted that whilst the musical importance of the attack portion of a signal is well known, most studies have focused on steady state sounds. Freed concluded that the mean spectral centroid (see Equation 1) is a strong predictor for the mallet hardness. This is one such study whereby an objective measure is linked to a listener perceptual attribute.

A common term often used by engineers and producers when describing a particular perceptual sensation found in produced music is called ‘punch’. Music is often characterised by listeners as being punchier yet the term is entirely subjective, in terms of both its meaning and subsequent auditory effect on the listener. Music of differing genre, tempo and playback level may all be perceived as having a different level of the punch attribute.

Indeed, punch could be described as a result of a different process depending on application, for example, a speaker manufacturer may utilise phase alignment in order to achieve a punchier output whilst a vocalist may add more vocal dynamics give a punchier performance.

If a mix engineer needs to achieve a level of punch required by an artist or client can this be done easily without a known reference? A mastering engineer may want to achieve an equal level of perceived punch between two songs without affecting any other perceptual attributes, creating additional nuisance artefacts or annoyance.

Pederson & Zacharov (2015) describe the semantic term punch as “Specifies whether the strokes on drums and bass are reproduced with clout, almost as if you can feel the blow”. Their study which seeks to define a sound wheel (i.e. lexicon) for the characterization of sound quality in loudspeakers, headphones, or other sound reproduction systems, places the semantic term ‘punch’ within the dynamics category.

Goodwin & Avendano (2004) refer to punch as a legitimate perceptual attribute and stated that a sound designer may design an attribute that would control low-level parameters that would in turn, for example, control a perceptual modification algorithm. They state that “a punch attribute might be established in terms of a range of sensitivity parameters for a transient detector and a range of intensity parameters for the intensity modifier.” The level of punch is, in this case, mapped to the perceptual dimension set by the sound designer, which in turn might not match that expected by the listener.

Zaunschirm et al. (2012) also refer to ‘punchiness’ as a perceptual attribute of a mix and their study conducted experiments measuring perceived transient suppression, increased punch and effects on quality. Whilst the authors of this paper collected subject scores which indicated differing listener preferences for the ‘punch’ attribute, the attribute itself was undefined within this study, therefore what the listeners were basing their scores on was unclear.

During the literature review, very few references on measuring the perceptual attribute ‘punch’ and indeed its definition were found other than the texts outlined above. This is surprising given that, as stated earlier, music is often characterised by listeners as being punchy or not and it is a term that is often used in audio testing.

Although the works cited above don’t measure punch objectively, they do imply that the perception of punch is altered by the modification of the transient. Zaunschirm et al. (2012) state that although the perception of punch was greater in all modified cases than the hidden references, there was also no significant difference between the use of different transient detection models.

As stated in Section 1.2, it is the author’s hypothesis that punch can be described as a short period of significant change in power in a piece of music or performance. The magnitude of change is associated with and proportional to the signal dynamics that are present and thus, in essence, productions that do not possess any transient or dynamic attribute cannot possess punch. The onset of the transient present across octave bands affects the listener perception of punch, with the lowest octave attributing the most punch as the onset is decreased and vice-versa. Punch is therefore related to transient change and the energy density (summation across frequency bands) occurring at a moment in time and duration.

Further to the above hypothesis, dynamic change in particular frequency bands may contribute to the perception of punch perceived by the listener and the overall average loudness level inherently affects this at that time (Moore, 2004). Thus, by mapping the perception of the punch attribute to objectively measured key attributes of the signal, one can produce a metric that could be utilised in music production and classification.

Signal dynamics are interlinked with signal loudness and therefore the latter will have a bearing on the perception of punch provided signal dynamics are sufficient. The sensation of punch may additionally have a bearing on the perception of overall signal quality. From a hierarchical viewpoint, punch would therefore form a lower level feature of overall signal quality. In terms of signal dynamics, the presence of signal transients is paramount.

## 2.6 Music information retrieval

Since the rapid expansion and ever growing availability of musical downloads, music retrieval and classification has become a very interesting and challenging topic (Casey et al., 2008). MIR can be useful in the automatic classification of such things as genre and also more recently in the use of quality measurement of audio.

MIR involves the analysis of low-level features, or characteristics of a piece of audio. If one considers genre classification, this can be thought of as a top layer semantic whereby its use is for categorizing and labeling selections of music (Scaringella et al., 2006). This is achieved through the analysis and grouping of low-level features, or characteristics. Once these characteristics are extracted, different classification approaches can be used to train the classifiers. The assumption is that the same genres of music may exhibit some similarity between the groups of characteristics. On a very basic level, loudness may be used to characterise groups of music, in this case it can be considered a high-level feature.

Transformations of features such as mean and standard deviation are often used to create further features. Distributions of features over longer window sizes are often employed where shorter analysis frames are not long enough to allow any meaningful data to be expressed. Longer window sizes are also able to signify temporal changes during the audio under test.

Audio features, once extracted are in a sense being used to ‘describe’ a particular perceptual sensation, thus the utilisation of the term Low-Level-Descriptor (LLD).

A large number of LLDs have been defined and can be referred to in ISO/IEC MPEG-7(2001), Tzanetakis and Cook (1999) and Peeters (2004). C libraries and Matlab scripts are available which offer integrated sets of functions dedicated to the extraction of associated musical low-level features. MIRToolbox is one such library for use within the MATLAB environment (Lartillot & Toivainen, 2007).

Some of the musical features extracted have shown a degree of perceptual relevance to listener opinion. In order to familiarise the reader with their relevance herein, some basic low-level features and their relative descriptors are now reviewed.

### 2.6.1 Basic low-level feature examples

*Spectral Centroid* can be described as the balancing point or centre of mass of the frequency spectrum. Perceptually it has a strong connection with the brightness of a sound, conversely it can be used to indicate relative dullness (Grey & Gordon, 1978). It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights

It can be calculated using

$$\text{Spectral Centroid} = \frac{\sum_{k=0}^{N/2-1} f(n,k)X(n,k)}{\sum_{k=0}^{N/2-1} X(n,k)} \quad (1)$$

where  $X(n, k)$  represents the magnitude of the  $k$ th frequency bin of the Fourier transform of the  $n$ th frame.  $f(n, k)$  represents the centre frequency of that bin within the frame.

*Spectral rolloff* is defined as the  $N$ th percentile of the power spectral distribution, where  $N$  is usually 85% or 95%. The rolloff point is the frequency below which  $N^{\text{th}}$  percentile of the energy in the spectrum resides.

$$\text{Spectral Rolloff} = \sum_{k=0}^{f_c} X(n, k)^2 = 0.95 \sum_{k=0}^{N/2-1} X(n, k)^2 \quad (2)$$

where  $X(n, k)$  represents the magnitude of the  $k$ th frequency bin of the Fourier transform of the  $n$ th frame,  $f_c$  is the spectral roll-off frequency and  $N$  is the total number of bins relating to the size of the FFT.

This measure can be used to distinguish between voiced and unvoiced speech due to unvoiced speech having a higher proportion of energy contained in the high-frequency range of the spectrum than voiced. (Scheirer & Slaney, 1997)

*Spectral Flux* is a measure of the absolute difference in the frequency distribution of two successive time frames. More precisely, it is usually calculated as the Euclidean distance between the two spectra if phase deviation is taken into account. By normalising the spectra, the flux calculation isn't dependent upon overall power.

It is therefore a measure of the rate of local change in the spectrum. It can be utilised for the purposes of onset detection.

$$\text{Spectral Flux} = \sum_{k=0}^{\frac{N}{2}-1} |X(n, k)| - |X(n-1, k)| \quad (3)$$

where  $X(n, k)$  represents the magnitude of the  $k$ th frequency bin of the Fourier transform of the  $n$ th frame,  $N$  is the total number of bins relating to the size of the FFT.

*Spectral Skewness* is a measure of the skewness in the magnitude frequency spectrum. A symmetrical distribution of components would result in a skewness value of 0. A positive skew would see a long tail in magnitude components to the right of the spectrum. Conversely, a negative skew would see a tail to the left. The coefficient of skewness is the ratio of the skewness to the standard deviation raised to the third power. It can be used to indicate if the spectrum is skewed towards a particular range of values.

$$\text{Spectral Skewness} = \frac{\mu_3}{\sigma^3} \quad (4)$$

$$\text{where } \mu_3 = \sum_{n=0}^{N/2} (x(n) - \bar{x})^3 f(n) \text{ and } \sigma = \sqrt{\frac{\sum_{n=0}^{N/2} (x(n) - \bar{x})^2 f(n)}{n}}$$

$x(n)$  and  $\bar{x}$  are the sample and sample mean respectively and  $f(n)$  is the centre frequency of the FFT bin  $n$ .  $N$  are the total number of bins relating to the size of the FFT.  $\sigma$  is the standard deviation.

*Spectral Spread* is a measure of the bandwidth of the spectrum. It also gives an indication of how distributed the spectrum is about its centroid value, hence the inclusion of the centroid in the measure. A high value of spectral spread would indicate a wide range of power distribution whilst a small value would indicate concentration about the centroid. Therefore, the measure could be utilised to distinguish between noise like and tonal sounds.

$$\text{Spectral Spread}(n) = \frac{\sum_{k=0}^{N/2-1} (f(k) - SC(n))^2 |X(n, k)|}{\sum_{k=0}^{N/2-1} |X(n, k)|} \quad (5)$$

where  $X(n, k)$  represents the magnitude of the  $k$ th frequency bin of the Fourier transform of the  $n$ th frame,  $N$  is the total number of bins relating to the size of the FFT.  $SC(n)$  represents the spectral centroid of the  $n$ th frame and  $f(k)$  is the centre frequency of the  $k$ th frequency bin.

*RMS* is a measure of the root-mean-square of the energy content within a frame.

$$RMS(n) = \sqrt{\frac{1}{N} \sum_{k=-N/2}^{N/2} X(n, k)^2} \quad (6)$$

It can be roughly equated to the loudness of that frame although a study Moore et al. (2003) suggested that loudness can be modified without any change in RMS occurring. This low-level feature is useful for performing segmentation since changes in loudness are important cues for new sound events. Intensity is approximated by the RMS level (Erling, 1996) and can be used for mood detection. RMS values can be extracted using a sub-band approach whereby each sub-band intensity is assessed individually or a sum of appropriate bands is made. In contrast, classification algorithms must be loudness invariant.

*Crest Factor* (CF), is a measure that incorporates the RMS measure and is often given either as a ratio value or expressed in dB (Hartmann, 1998). It is used to give an indication of dynamic range within the section of audio being measured.

$$CF = \frac{X_{peak}}{X_{rms}} \quad (7)$$

where  $X_{peak}$  is the peak amplitude detected in the sample frame and  $X_{rms}$  is the RMS level measured in the audio frame.

*Rhythm Strength* is a measure of the lowest sub-band energy content within a frame.

Drums and bass are typically the most important components used to represent rhythm within a musical piece. These components show their properties mainly in the lowest sub-bands. *Rhythm Strength* is a measure proposed by Lu et al. (2006) that utilises the lowest sub-band to extract an amplitude envelope (using a half Hamming window) and a canny estimator (to detect the rhythm difference curve). From this, the average strength of the instrument onset is calculated. An average strength is assumed due to the notion that the majority of the signal strength will be as a result of the lowest band. Whilst this may be the case, additional frequency bands of the signal must be considered if timbral aspects are of importance. Lu et al. (2006) do not state which lowest sub-band is utilised in their publication.

### 2.6.2 Additional Low-Level Descriptors (LLDs)

Audio stimuli exhibit both temporal and frequency domain characteristics. These characteristics can be measured in different ways and combined to offer an overall impression or quality score to the listener. They may also be used independently to describe an aspect of the stimuli under test, for example, loudness.

A very good example of the use of low-level descriptors in a qualitative measurement application is the *Sound Goodness System* (Oriol et al, 2015). This real-time system complements the tuner functionality by evaluating the sound quality of a music performer

in real-time. It consists of a software tool that computes a score of how well single notes are played with respect to a collection of reference sounds. This system employs both a combination of audio feature extraction and machine learning based on an annotated training database.

Lu et al. (2006) also demonstrated that mood could be predicted using a combination of underlying audio features, features investigated were signal intensity (in the context of this study this was effectively the signal loudness) and rhythmic features such as regularity, tempo and bass onset strength. They found that rhythmic features dominate on ‘exuberant’ groups of audio.

In terms of quality measurement, a number of lower level descriptors are often combined to form an overall qualitative score as in the Sound Goodness System. The two examples previously outlined indicate the possibility of utilizing low-level features for effective perceptual model development.

A large number of descriptors have been defined ISO/IEC (2001) MPEG-7, (Tzanetakis & Cook, 1999) and (Peeters, 2004) and their relevant semantics have, in some cases been investigated and defined. MIR (see Section 2.6) relies heavily in the extraction of these features. Whilst a number of the measures proposed in the MPEG-7 standard are not new, e.g. the Spectral Centroid, the standard contributes by collecting them together for common use in content analysis and subsequent description of underlying audio features.

Whilst outlining the context and terminology within this thesis, *Spectral Centroid*, *Spectral Rolloff*, *Spectral Skew*, *Spectral Spread*, *RMS*, *Rhythm Strength* and *Loudness* were defined, see Subsection 2.6.1. Part 4 of the ISO/IEC (2001) MPEG-7 standard outlines other low-level descriptors (LLD), some of which are identical to those already mentioned. Measures relevant to this thesis can be categorised as follows:

- *Basic: Instantaneous waveform and power values.*
- *Basic spectral: Log-frequency power spectrum and spectral features (for e.g. spectral centroid, spectral spread, spectral flatness).*
- *Signal parameters: Fundamental frequency and harmonicity of signals.*
- *Temporal Timbral: Log attack time and temporal centroid.*
- *Spectral Timbral: Spectral features in a linear frequency space.*

These LLDs are either given as a single value for a stimuli segment, as a series or evaluated statistically if taken as a group. The LLDs are now summarized along with a discussion as to their relevance to the perception of punch.

#### **2.6.2.1      *AudioWaveform Descriptor***

This descriptor is utilised for display purposes and is based on temporally sampled scalar values of the stimuli in order to represent the waveform envelope.

#### **2.6.2.2      *AudioPower Descriptor***

This is the temporally-smoothed instantaneous power of the stimuli. It can be used to give a summary overview of the signal under test. This metric although useful perhaps in onset detection, isn't feasible for use in the differentiation of power formed across the spectra of the stimuli. A more appropriate measure for this could be that detailed in 2.6.2.4

### **2.6.2.3      *Silence Descriptor***

Utilised to signify frames where silence has been detected, or no significant audio level is present. Silence or low-level signals could be relevant to the overall impression of audio dynamics. Without silence or low signal levels, RMS levels are higher and CF is reduced.

### **2.6.2.4      *AudioSpectrumEnvelope Descriptor***

The *AudioSpectrumEnvelope* descriptor contains the logarithmic spectrum of the audio stimuli. The log-frequency scaling of the output serves two purposes: a) It gives a compact view of the spectral components contained in the stimuli b) It mirrors the logarithmic response of the human ear. Since the envelope is a power spectrum, the sum of all the spectral coefficients is equal to the power in the windowed data.

It is possible therefore, through the use of envelope extraction over time, to monitor the literal power in the audio stimuli with respect to the components being summed and the time over which it is being measured. In the case where sub-band filtering of the audio is taking place, it is possible to extract both the overall *Signal Intensity*, which is the sum of all the FFT bins and *Sub-band Intensity Ratio*, which is the ratio of power within each sub-band. The latter being of use in determining where the majority of power lies within the frequency spectrum.

### **2.6.2.5      *AudioSpectrumCentroid Descriptor***

This descriptor is identical to Spectral Centroid detailed in Subsection 2.6.1.

As discussed, this measure gives an indication of the centre of gravity within the spectral envelope. As such it can be utilised to indicate the ‘Brightness’ of a sound. In addition, the ISO/IEC (2001) MPEG-7 standard proposes instrument timbre definitions and the centroid value plays a role in this, for example, in categorising percussive type sounds. As such, the centroid could be relevant where moments of punch are perceived.

#### **2.6.2.6     *AudioSpectrumSpread Descriptor***

This descriptor is identical to Spectral Spread detailed in Subsection 2.6.1.

This descriptor indicates how spread the spectrum is about the centroid. In the case where no spread is evident, indicated by a low spectrum spread value, the stimuli might be classified as pure-tonal. In the case where a large spread of spectra is detected, this would indicate a noise like stimuli. Transients within stimuli, particularly those that are percussive in nature should be revealed by a large spectrum spread value, therefore, it may have a fundamental link to punch perception.

#### **2.6.2.7     *AudioSpectrumFlatness Descriptor***

This measure reflects the flatness properties of the power spectrum. For a given stimuli frame, it consists of a series of values, each one expressing the deviation of the signal's power spectrum from a flat shape.

A flat spectrum shape can correspond to a noise or an impulse signal. Therefore, in a similar nature to the *AudioSpectrumSpread* descriptor, the measure may have use in indicating impulses or onset/offset points within an audio signal. A measure approaching 1 is usually indicative of a spectrum that is similar to that associated with white noise.

#### **2.6.2.8     *AudioFundamentalFrequency Descriptor***

As the name might suggest, this measure returns the fundamental frequency of the stimuli under test. It may be useful to detect a particular fundamental and observe its correlation with perceived punch. However, an audio mix is likely to be a combination of instruments and consequently many harmonics would be present in a mix. The measure would therefore be irrelevant when used for this purpose unless signal separation takes place.

### **2.6.2.9     *AudioHarmonicity Descriptor***

This measure gives an indication of whether or not the audio under test is harmonic in nature. It is a measure of the proportion of harmonic components in the power spectrum. An AHD equal to 1 would indicate a purely harmonic signal whilst noise like signals are shown as  $\leq 0.5$ .

### **2.6.2.10    *LogAttackTime Descriptor***

The log attack time (LAT) is defined as the time it takes to reach the maximum amplitude of a signal from a minimum threshold. It can be used to quantify the onset of instruments. Its use in a complex mix is somewhat limited unless some form of signal separation takes place initially. This could be a useful measure if combined with the power spectrum of the stimuli under test in order to determine power vs. time measurement, which could correlate with punch.

### **2.6.2.11    *TemporalCentroid Descriptor***

Unlike the LAT, the temporal centroid is a measure of the centre of gravity of the onset time itself. This can be seen as equivalent to the spectral centroid measure in that it is the time where the energy of the signal is most concentrated. It is possible for example to have two signals with the same attack time during their onset phase, however, the shape of that transient may be very different. This measure may be of use to determine if onset shape has a bearing on punch perception.

## **2.7    Loudness**

Loudness is a very good example of a perceptual attribute that has been researched and objective measures proposed which has resulted in a new standard being adopted within the broadcast industry and albeit, more slowly by the wider music production industry (Skovenborg & Nielsen, 2004). Although it could have been included as a subsection of

MIR after all, it can be thought of as a characteristic that can be extracted from audio and used for normalisation or categorisation, it has been given its own section as it can be considered as a high-level descriptor rather than a low-level one.

The loudness of a sound, from a perceptual viewpoint, is a measure of the effect of the energy content of the audio signal on the ear. It is also dependent on the frequency content of the audio signal itself. For example, the perceived loudness of a pure tone, say 100Hz at 40 decibels (dB) would be perceived to be quieter to a normal hearing person than a 1kHz tone at 40dB.

A definition of the loudness of tones has been constructed through extensive research in classical psychoacoustics, traditionally using stationary signals. Stationary signals are ones that can be described by their frequency spectrum, such as noises and tone complexes and are not changing with respect to pitch or volume for example. Fletcher & Munson (1933) proposed a set of perceptual loudness contours showing the listener response to varying pure tones and playback levels with reference to a 1kHz reference. The experiment is often referred to as loudness magnitude estimation, and in this case the ‘loudness level’ of a sound is defined as ‘the sound pressure level of a 1 kHz tone in a plane wave and frontal incident that is as loud as the sound; its unit is “phon”.’ (Zwicker & Fastl 1999)

A re-determination of these curves was carried out by Robinson & Dadson (1956) which became the basis for the widely accepted ISO 226 (1987) standard. The curves were later updated to form ISO 226 (2003) standard, which is referred to generically as the ‘equal-loudness contours’. These contours were revised again in 2007 to form the ANSI S3.4 (2007) standard. This incorporates a significantly lower sensitivity for low-frequency signals than the 1987 version. Salomons & Janssen (2011) provide comparative plots of these contours.

There are two approaches that can be adopted with respect to objective loudness measurement; these can be described as single-band and multi-band.

Single-band models of loudness can be constructed based on the contours outlined above. For example, if envelope detection is applied to the audio under test and frequency weighting is applied to the signal based on the equal loudness contour, a summation of overall perceived loudness can be achieved. As the loudness contours vary depending on playback level, a weighting curve IEC 60651(1979) is chosen e.g. A or B.

The  $L_{eq}$  measure specified in ANSI (1994) is the equivalent continuous sound level, or time-average sound level. The  $L_{eq}$  corresponds to an (energy domain) average over a time interval  $T$  during which the sound level is measured, in dBs. When used together with weighting functions IEC 60651(1979), the  $L_{eq}$  measure is often considered as a measure of loudness. In essence, when used in this way, the measure is a single-band model. For example,  $L_{eq}(A)$  would signify that an A-weighting function had been applied during the measurement. For long-term loudness measurement (perhaps of the entire length of the stimuli),  $T$  would be the same length as the stimuli under test. For short-term loudness measures, it would be made a fraction of the entire length of the stimuli.

The  $L_{eq}$  measure is defined as

$$L_{eq}(W) = 10 \log_{10} \left( \frac{1}{T} \int_0^T \frac{x_w(t)^2}{x_{ref}(t)^2} dt \right) \quad (8)$$

Where  $x_w$  is a frequency-weighted ( $w$ ) sound level at time  $t$  and  $x_{ref}$  is the reference signal.

Whilst the ISO 226 contours are widely accepted as being relevant to ‘pure tone’ measurement, extra consideration must be taken into account when measuring more complex audio sources, such as music. One major consideration is the frequency resolution of the ear, known as the ‘critical bandwidth’.

For example, for two tones less than one critical bandwidth apart, partial masking will occur, thus a more intricate algorithm, which takes this into account, is required. This makes the loudness summation a more complex process.

To counter these effects, some loudness models implement a multi-band approach, the two key methods are detailed in ISO 532 (1975). This initial standard stems from two proposals and is made up of method A (Stevens, 1957) and method B (Zwicker, 1960). Both approaches attempt to approximate the response of the human ear by way of auditory filter banks in octave and critical bands respectively. The bands are often weighted and the overall loudness is calculated as an average over a period of time, either based on short or long term windows.

Compared to the method by Stevens, the Zwicker method B, includes a spreading function closely related to the effect of simultaneous masking. Furthermore, frequency weightings for both diffuse and free sound fields are included. As far as implementation and uptake, this model has proved more popular (Scheuren, 2014) than Stevens. The Stevens method has been proposed to be replaced by the newer (Moore et al, 1997) method.

The principal difference between this and the Zwicker method is the use of equivalent rectangular bands (ERB). Like critical bands, these widen on a logarithmic frequency scale towards high frequencies. In the lower frequencies, the bands are narrower and more numerous than critical bands. The Moore et al. method also utilises equal-loudness contours that are closer to the newer 2003 ISO 226 contours than those used in the Zwicker model which are based on the 1987 approximations.

It is likely that two new standards will emerge as detailed by the International Standards Organisation (2016):

ISO 532-1 “Methods for calculating loudness – Part 1: Zwicker method” and  
ISO 532-2 “Methods for calculating loudness – Part 2: Moore/Glasberg method”

These, more complex loudness summation approaches, although modelled on the human auditory response, don’t always perform better than single-band methods (Skovenborg & Neilson, 2004; Ferguson et al., 2004).

Unfortunately, across all broadcast platforms, due to the long term lack of standardisation of both loudness levels and loudness monitoring, material proliferated with widely varying loudness thus causing listeners to experience undesirable changes in loudness between different sources, TV or radio stations and also between different programme segments within the same channel of TV or radio. This problem was exacerbated by what has now become known as the ‘loudness wars’ (Loudness Wars, n.d.) and led to movements such as TurnMeUp! (2009) to promote the opposite.

Audio material that is deemed fit for release has usually gone through a mastering process. This process is two-fold, firstly to ensure that the material fulfils certain aesthetic requirements and secondly to meet the technical requirements of the broadcast or delivery format e.g. bit rate. Loudness, is an aesthetic choice that since the mid-1980’s appears to have dominated in music production. Due in part to the record labels need to be the loudest on radio but also driven by the demand of the artist for their material to match that of their peers. In addition, the paradigm of loudness maximisation could also be attributed to the fact if two identical tracks are played back at differing loudness levels, the louder of the two will be perceived as better. (Vickers, 2011).

During loudness maximisation, material is compressed, resulting in a reduced peak to RMS level ratio and thus an overall reduction in dynamic range. This peak-level based processing makes material perceptually louder. This loudness maximizing is often achieved by aggressive application of dynamics compression, which may lead to undesirable artefacts, as well as technical problems (Katz, 2002; Neilsen & Lund, 2000). Whilst this continued decrease in dynamic range occurs, it is accepted amongst audio professionals that this is potentially detrimental to the overall audio quality of the music. By predicting perceived loudness of audio material, it is possible to measure and control the loudness of the delivered or broadcast product thus ensuring some uniformity. In addition, effective loudness measurement can help in music categorization and retrieval.

Many proposals and studies relating to loudness are documented and some have been outlined earlier. An additional aspect of loudness measurement that must also be taken into account, particularly within the broadcast and production environments is the simplicity of

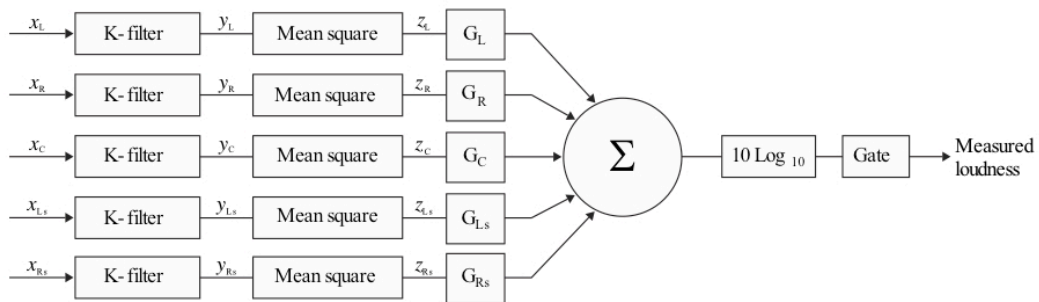
implementation which usually affords real-time operation. Single-band methods are implicitly simpler than their multi-band counterparts and provided that the resulting measure is greater or equal to 95% consistent in predicting subjective loudness levels it could be deemed viable for use over a more accurate multi-band model.

One such standard of model of loudness measurement is detailed in recommendation ITU-R BS.1770-4 (2015). This standard is now widely accepted as the loudness measure to be utilised for normalisation of broadcast audio.

A key benefit of this proposed loudness model is its simplicity. The algorithm is made up entirely of basic signal processing blocks, these relate to simple weighting filters and energy summation. Due to this simplicity, the loudness model can easily be implemented in the time-domain and run in real-time.

### 2.7.1 ITU-Recommendation BS.1770-4 (2015)

The ITU-R BS.1770-4 (2015) recommendation has been widely accepted as the loudness measure to be utilised for normalisation of broadcast audio. EBU-R 128 (2014) stipulates its implementation including specification of maximum target levels, momentary and short-term loudness measurement. In addition, further extensions to allow the effective measurement of consistency in loudness ranges are given. These extensions were originally proposed by Skovenborg & Lund (2008).



**Figure 2 - BS.1770 Loudness Model Outline**

The algorithm is relatively simple in nature, based around the summing of energy on each audio channel, this can be represented as shown in Figure 2. Made up entirely of very basic signal processing blocks, the algorithm is based on a simple  $L_{eq}$  measurement system which sees the stimuli pre-processed with a revised low-frequency-B weighting filter (RLB) (Soulodre, 2004) and a high shelving filter which aims to mimic the effects of the listeners head. Together these two filters are referred to collectively as a K-weighting filter. The low frequency enhancement (LFE) channel is neglected in the algorithm.

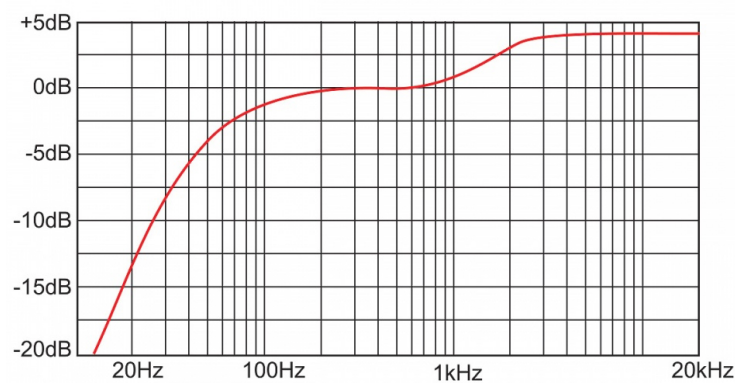
This loudness model can easily be implemented in the time-domain. In addition, since the contributions of the individual channels are summed as loudness values, rather than at the signal level, the algorithm is more generic and robust due to independence of inter-channel phase or correlation.

Another key benefit of the algorithm is its scalability. Since the processing applied to each channel is identical, it is very straightforward to implement a meter that can accommodate any number of channels. The LFE channel is an exception to this given the lack of accuracy in contouring at the frequencies involved. The output of the loudness model is given in Loudness Units (LU), whereby one LU is equal to one dB. Therefore, in simple terms the model output is a weighted dB measure.

### 2.7.1.1 *K-Weighting and filtering.*

As previously mentioned the K-Filter is made up of a revised low-frequency-B weighting filter (RLB) (Soulodre, 2004) and a high shelving filter, which is incorporated to mimic the effects of the listeners' head. The hi-shelving filter has a corner frequency of 1kHz and a boost of +4dB.

Overall the pre-filtering stage has the effect of attenuating the low frequencies (<100Hz), preserving the frequencies between 100Hz and 1kHz and boosting those above 1kHz by 4dB. The frequency response of the filter is shown in Figure 3.



**Figure 3 – Loudness K-weighted Filter**

The algorithm is designed to output a number of metrics outlined as follows:

**Program Loudness** – This is how loud a program is on average. The values of loudness are expressed in either LKFS (Loudness K-weighted Full Scale) or LUFS (Loudness Unit Full Scale). These are simply alternative names for the unit as the K-weighting filter is always applied to the audio stimuli. Both measures are given relative to digital full scale (FS) and one unit of LUFS is equal to one dB. The measurement and results are the same.

Recommended levels for broadcast are -23 LUFS however, AES Tech Doc (2015) recommends levels of no more than -16LUFS for streamed and networked file playback. This is somewhat of a relaxation to allow musical material to be played at a higher level than the broadcast standard. It is understood that the ‘mastered for iTunes’ level is also targeted at -16 LUFS rather than the lower -23LUFS (Lund, 2011).

**Loudness Range (LRA)** – detailed in EBU Tech Doc (2010) is a measure first proposed by Skovenborg (2009). It is calculated by measuring the ITU-R (2015) BS.1770-4 loudness within a 3 second time window and building a histogram of these values. The LRA is defined as the difference between the 10th percentile and the 95<sup>th</sup> percentile on the loudness histogram. The measure indicates loudness variation within a program and can indicate when compression is required, for example, when high LRA values are measured. Recommended LRA is -20LU (Loudness Units) for comfortable listening conditions.

**Maximum True Peak Level**, often abbreviated to just True Peak, indicates the maximum value of the audio signal waveform. The true peak should not exceed the recommended level of -1dBTP, where dBTP indicates reference to digital full scale i.e. 0dBFS. The ITU-R BS.1770-4 (2015) recommendation stipulates an over sampling rate of 4 times be adopted in order to allow for true peak measurement to take place.

### **2.7.1.2      *Gating Mechanism***

A gating mechanism is utilised in the loudness model in order to ignore audio that falls below a given threshold. The addition of a gating mechanism helps to maintain ‘overall’ program loudness between stimuli, particularly if they contain fade ins/outs. A two-stage process is adopted to perform the gating mechanism. Firstly, progressive loudness frames are compared to an absolute threshold of -70LKFS and frames below this level are ignored. From these frames an intermediate loudness is calculated. This intermediate loudness forms the relative threshold, which is stipulated at -10LU below the intermediate value. Finally the program loudness is calculated by the sum of the loudness levels above the relative threshold level.

There are 3 different loudness outputs given by the model, these are:

Momentary (M) – This corresponds to each 400ms window loudness with no gating.

Short-term (S) – This is equivalent to the above but uses a window size of 3 seconds.

Integrated (I) – This is also known as program loudness and is the final result of the summing calculation after gating.

The model has been implemented in broadcast loudness meters, (AC-R128, 2012; Vis-LM, 2012), but another desirable use is in sound quality research. For example, where a set of stimuli might need to be "loudness equalised" so that other (subjective) factors can be investigated independently of loudness (Aarts, R.M, 1992).

## 2.8 Audio dynamics

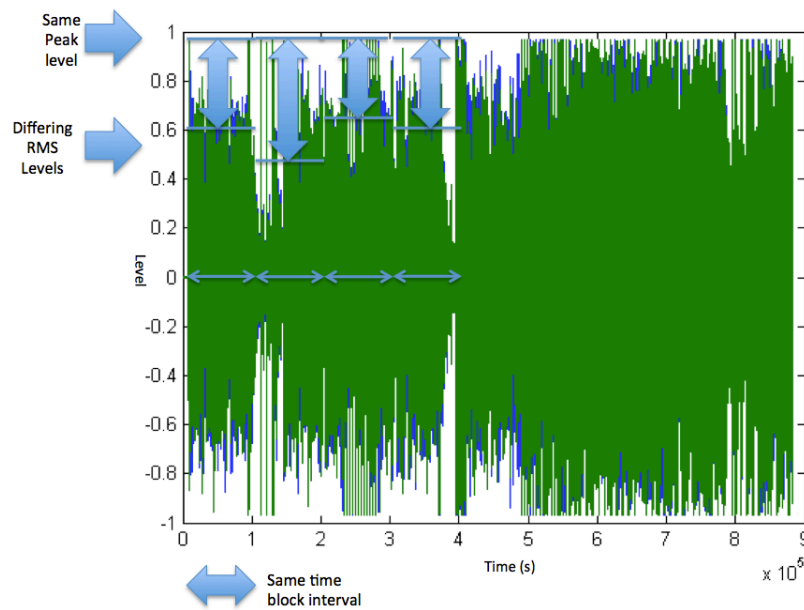
The term dynamic range is often quoted in dB when describing the performance of an audio system. The context of measurement is an important factor to consider when the interpretation of the dB value is evaluated. The context can either be categorised as that of a system or signal.

In the context of a system the measurement is used to describe the maximum range that is permissible, before distortion takes place (clipping), measured from the noise floor to the peak level. The Audio Engineering Society specification document, AES 6id (2006), specifies this measurement as "20 times the logarithm of the ratio of the full-scale signal to the root-mean-square (RMS) of the noise floor in the presence of signal, expressed in dB FS". This value gives an indication of the true headroom of a system and shouldn't be confused with SNR (Signal to Noise Ratio), which is often measured without the presence of a signal and can therefore give an inaccurate system measurement due to muting circuits.

When we describe the signal itself rather than the system under test, the dynamic range can be given as the ratio of the full-scale level of the signal to its lowest level. Given that audio signals under test are generally varying in level, particularly during fade ins, fade outs, interludes etc., an average level formed by the RMS is generally taken of a section of audio under test as being representative of the 'active' passage of music. This average level is then used to compute the dynamic range in conjunction with the peak level measured during the same passage.

This measure is referred as *Crest Factor* (CF), and is defined by equation 7 in Section 2.6

Audio dynamics, within the context of this work, therefore relate to the magnitude of change observed within the audio stimuli defined within a specific time frame. For example, considering Figure 4 – , shown below, the dynamic range measured at each time interval will vary due to the differing RMS levels measured at each interval despite the peak level being the same within each.



**Figure 4 – Audio Dynamics**

It is clear to see from this that the interval within which the dynamic range measurement is being taken has a direct impact on the measurement obtained. Therefore, careful selection of the interval dependent upon the source material is required in order to obtain meaningful results. A further point to highlight is that when taking this type of measurement, the signal RMS is being taken using the full audio spectrum of the signal. This may result in a ‘dynamic range’ measurement that may not correspond with listener perception as a result of particular frequency bands being more perceptually relevant than others. Consider a piece of audio whereby heavy compression of its high band components has taken place. Whilst this could have an impact on the perceived tonality of the audio, the overall quality perception of the audio by the listener may be less affected due to the dominance in low-

mid frequency energy being present. The hypothesis that dynamic range within frequency bands plays a role in listener perception is explored in the early experiments that are carried out within this thesis and forms evidence to support the author's definition of punch outlined earlier in Section 2.5.

A perceptual metric related to Crest Factor is the peak to loudness ratio or loudness crest factor defined by the peak amplitude of the signal divided by the loudness measured by ITU-R BS.1770-4 (2015). This measure can be used to quantify the level of audio dynamics present in the signal being measured. It has also been shown to correlate to integrative loudness models (Equation 8) due to decreases in peak to RMS levels that usually take place during audio compression processing.

A number of other proposals for the measurement of signal dynamics have been published (Skovenborg, 2014) and their link to perceptual attributes have been explored. (De Man et al., 2014; Wilson & Fazenda, 2013)

## **2.9 Compression, dynamics and audio quality**

The aim of this section is to describe the inter relationship between signal dynamics, compression and quality perception. Hierarchically, overall quality perception could be deemed higher in level than the underlying low-level parameters that are actually being judged by the listeners, such as punch. It is therefore good practise to begin to study the effect of compression on quality in the first instance. Section 2.8 introduced the reader to the context of audio dynamics and highlighted one form of objective measurement that can be employed to measure it called *Crest Factor*.

Further methodologies adopted in the measurement of underlying audio dynamics are reviewed and presented in Section 2.10 and their methods are contrasted. Changes of listener perception of audio quality relating to the level of audio dynamics contained within a piece of music are also explored. This section, in addition to the previous literature review, formed the basis for the experimental work that follows in the later chapters.

Within the context of this work, compression refers to the reduction in dynamic range of a signal rather than the reduction in bit rate, commonly found in the encoding of low bit rate audio. Dynamic Range Compression (DRC) is achieved either through the manual or automated ‘riding’ of faders or automatically through the use of compressors or limiters.

DRC is employed in music production for both practical and aesthetic reasons (Katz, 2002). The effects of DRC are to reduce the peak to RMS ratios of the audio signal and this could be for artistic effect such as to automatically control a vocal level during a performance or to de-emphasise attack transients of an instrument.

Limiting is often used in the music mastering process. This specific form of DRC produces a relatively consistent output level by utilising high compression ratios for sounds above the compression threshold. For an entire audio file, an overall increase in RMS level, as result of the limiting process, can result in a perceptual loudness increase in that file. That said, an interesting result was observed in a study by Moore et al. (2003) which demonstrated that even when the RMS level is kept constant, a change in loudness perception can be achieved depending on the degree of dynamics compression applied. This could be explained by other salient features of the audio being modified as a result of the compression being applied, for example, signal brightness increase may occur as a byproduct of loudness normalisation.

“Louder is better” (Vickers, 2011), and other works (Maempel & Gawlick, 2009) suggest that listeners prefer increased loudness. However, underlying perceptual effects enhanced or created by the processing involved, may be influential upon listener preference. In the latter study, high bass amplitude was cited as a cause. Hjortkjaer & Walther-Hansen (2014) found no evidence for preference of less-compressed music and with regards to ‘perceptual’ attributes, they failed to find differences in perceived ‘depth’ between original and more compressed audio.

Wilson & Fazenda (2013) found that listeners could identify reduced dynamic range as a determinant of reduced quality. In their study, CF was utilised as a measure of dynamic

range. They also identify that distortions, tempo, spectral features and emotional predictions show correlation with perceived audio quality. Some of these underlying features are clearly modified through the application of DRC.

Croghan et al (2012) using the genres of ‘rock’ and ‘classical’ music found that listeners preferred stimuli where DRC had been applied moderately. In contrast to the ‘louder is better’ paradigm, they state in their study that “louder is better... to a point”

It’s important when reviewing all of these studies to note that whilst the notion of *dynamic range* is often cited and used in audio experiments, its measurement is still somewhat poorly defined due to the measure itself having little standardisation. In the case of *Crest Factor*, values resulting from this calculation are largely dependent on the timeframe in which they are being measured as highlighted in Section 2.8. This timeframe, sometimes referred to as either micro or macro, also has no formal standardisation.

## **2.10 Audio dynamics measurement methods**

There have been a number of other proposals that attempt to measure the level of audio dynamics within a signal and associated perceptual artefacts. The following presents an overview of the approaches adopted.

Loudness Range, EBU Tech Doc (2010), outlined in Subsection 2.7.1.1 can be considered a measurement of ‘loudness dynamics’ albeit in an integrative sense, for example if any short term ‘spikes’ of dynamic activity occur in the signal, these are ignored by the algorithm.

The Pleasurize Music Foundation (Pleasurize, n.d) released the TT-Dynamics meter that calculates a different form of dynamic range. This algorithm calculates the ratio of the peak to the RMS, but limits the RMS to those values that occur in the top 20% of the histogram. The authors argue that only using the top 20% allows them to compare a variety of program types (genres of music, speech, etc.).

Ljudtekniska (2013), proposed an all-pass filtered crest factor method in his ‘MasVis’ analysis tool to measure or visualise audio dynamics. The all-passed crest factor measure is based on applying a set of all-pass filters in parallel, and then measuring the CF of each filtered version of the signal. Details relating to this algorithm are not available, however, from studying the manual it appears the author is ‘estimating’ the crest factor of the audio based on peaks that are measured across the all-pass filter outputs. For example, if there is a filter output that is particularly larger than the rest, it is assumed that this may be an indicator of the true dynamic before compression on the audio took place. The work suggests that for natural, unprocessed audio, crest factor may increase a little as loudness increases (assuming the audio is richer in spectrum and the number of sound sources increases). Conversely, if the audio is subjected to heavy limiting the crest factor will decrease as loudness is increased.

As loudness is increased towards the limiting threshold, the crest factors across the frequency bands should also normalise and begin to correlate towards each other.

Vickers (2001) proposed a measurement he called *dynamic spread*, which is a measure of the spread of time varying loudness levels. It doesn’t rely on a peak measurement but rather the p-norm of the signal. Vickers recommends  $p=1$ , such that the dynamic spread is just the mean absolute deviation of the signal. If  $p=2$ , a measure corresponding to standard deviation of the loudness levels is achieved.

Wilson & Fazenda (2013) utilised a modified probability mass function (PMF) with zero mean based on the assumption that as audio is compressed, more extreme amplitude levels assume higher probability. This measurement was utilised to provide a feature associated with audible distortion. Hard-limiting and DRC have been studied and linked to listener preference (Croghan et al, 2012) and since these are encompassed by PMF the measure is relevant in this work. They refer to this feature as ‘Gauss’ and whilst not being a direct measure of signal dynamics they utilise it to describe loudness, dynamic range and the related audible distortions in their study.

Deruty & Tardieu (2014) introduced a global descriptor called *High Level Sample Density* (HLSD). The measure indicates the proportion of samples that are present in the entire audio file that are above a -1dB Full Scale (dBFS) after normalisation. They show that HLSD is a good measure of the amount of brickwall limiting that has been applied to the audio content. Their study doesn't attempt to correlate this measure with listener perception of dynamic content or quality. Instead it uses the measure to timeline and predict what type of processing was applied via a studio practice timeline which groups technical innovation and studio practices between the 1960s-1997.

Tollerton (2008) proposed an algorithm he calls pfpf. This algorithm calculates ITU-R BS.1770-4 (2015) loudness on 3 different time scales (10ms, 200ms, 3sec) and defines short, medium, and long-term dynamic range respectively as the range from the 50th percentile to the 97.7<sup>th</sup> percentile.

An experimental descriptor called 'Density' was proposed by Skovenborg & Lund (2008). It measured the variation of loudness on a microscopic timescale. Microscopic means window sizes that are smaller than those typically utilised for momentary loudness measurement in ITU-R BS.1770-4 (2015). The use of a smaller window size results in the output of the model tracking the more local variations of the signal known as microdynamics (Katz, 2002).

Work on the 'Density' descriptor led to results being published in 'Measures Of Microdynamics' by Skovenborg (2014) and the measurement was termed Loudness Dynamic Range (LDR). In this study, LDR, based on the maximum difference between a "fast" and a "slow" loudness level, had the strongest correlation with the listener perceived dynamics of stimuli consisting of music and speech when compared to other types of measure such Peak-to-loudness ratio (PLR) and CF.

PLR is defined by the ratio of the true-peak of the signal and its loudness. This is a very similar measure to CF however the latter utilises the RMS of the signal in its denominator. Skovenborg (2014) indicated that whilst the PLR measure could be used to indicate

microdynamics within a signal, its robustness wasn't reliable as a result of the 'peak' of the signal not being typical amongst certain types of genre.

The issue with estimation of dynamic range of musical signals is that the results are largely dependent upon the window sizes utilised during measurement. This was highlighted by both Skovenborg (2014) and by Boley et al (2010) in their evaluation of algorithms for comparing and contrasting dynamic range.

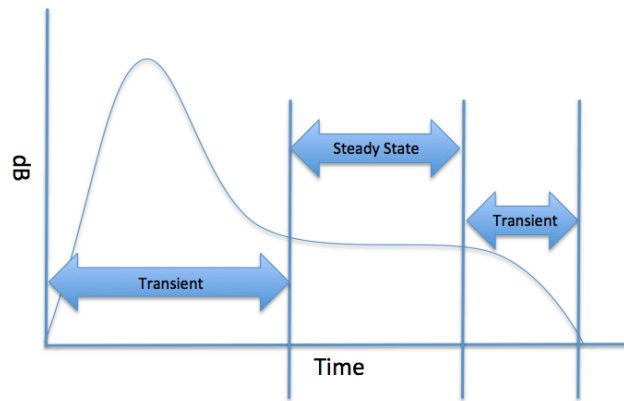
In addition, whilst it may be immediately intuitive to retain as much dynamic range in music as possible, the listening environment must also be considered to ensure that the listener experience isn't compromised. For example, airplane cabin noise may require the dynamic range of material to be reduced.

## **2.11 Signal transients**

Music can be considered to be a collection of complex tones where a complex tone consisting of a number of different harmonic components with varying magnitudes and phases. Each tone component consists of both steady state and transient parts. Previous work has identified that the transient portion of a complex tone contains a great deal of information with respect to perceptual attributes of the source (Grey, 1977; Rasch & Rasch, 1981).

Bello et al (2005) describe transients as short-time intervals within which the signal evolves quickly and unpredictably. These short intervals are often associated with the percussive elements of a musical piece and/or other aspects of the music that possess similar characteristics. They tend to show up on a spectrogram as vertically aligned points of energy distributed along the frequency spectrum. Percussive sounds tend to have a concentration of this energy during the attack phase of the wave. Their excellent tutorial describes different techniques for onset detection in signals characterised by either changes in amplitude or phase of frequency components or a combination of both.

The transient part of the signal can be loosely defined as the time interval in which the signal is evolving into its steady state, transients can be present in both attack or release phases of a signal as shown in Figure 5. Detection of transients can be useful in such applications as note detection, signal enhancement, dynamic range control and musical transcription (Collins, 2005; Avendano & Goodwin, 2004; Walsh et al., 2011; Wang & Tan, 2008).



**Figure 5 - Audio Transient**

The onset of a sound is the time that marks the beginning of the attack and is the earliest time at which the transient can be detected. Vos & Rasch (1981) define the perceptual onset time as the time a stimulus is first perceived and it is generally different to the physical onset. Gordon (1987) made this observation albeit with slightly differing results.

Almost all genres of music have significant transient content throughout as a result of differing tonal and noise onsets. Various methods of transient detection can be employed with varying degrees of success depending on genre and application. Some are psychoacoustic in nature (Collins, 2005; Klapuri, 1999) and incorporate sub-band approaches to mimic human hearing and detection mechanisms (Zaunschirm et al., 2012).

Modification of the transient portion of a sound source has been shown to modify the perception of the source by the listener (Collins, 2005; Goodwin & Avendano, 2004). Zaunschirm et al (2012) concluded that whilst the modification of transient had an effect on the perceptual attributes under test, the method of transient detection didn't have a significant effect.

Whilst the majority of literature tends to focus on the transient nature of an onset, it's important to highlight that the offset of a signal is also transient in nature and therefore will possess similar characteristics to the onsets. The usefulness of offset detection, whilst often ignored, is that of note length annotation and detection. This would prove useful to determine the integrated loudness of burst of audio if signal duration were short.

## **2.12 Summary**

In summary of this chapter, increases in loudness seen over the last decade has inevitably caused a reduction in dynamic range of produced music. Has our perception of both the subjective quality of the audio become somewhat distorted with regards to an acceptance of a louder product vs. a reduced dynamic range? There are numerous factors (Ronan et al., 2014) which suggest that the 'louder is better' paradigm may be more to do with other perceptually salient features being modified by the DRC rather than the headroom loss itself.

With this in mind and the recent move towards the use of loudness normalisation, it is important to continue to measure the other perceptually salient features both pre and post the loudness normalisation process in order to maintain both consistency and perceived quality.

As discussed, development of the measurement of audio relating to loudness has been undertaken however, work on measures based on dynamic range and other underlying low-level features and their correlation with the perception of listener quality needs to be extended. Audio dynamics have no clear perceptual counterpart and no standardised way

of measurement. This lack of definition leads to partially wrong or even contradictory conclusions in several publications.

It is the author's belief that whilst loudness and signal dynamics are inextricably linked, the latter should be considered as a separate measure when attempting to quantify both music quality and as a lower level feature contributing to punch.

In Section 1.2, a hypothesis that punch can be described as a short period of significant change in power in a piece of music or performance was proposed. The magnitude of change could be associated with and proportional to signal dynamics that are present and thus, productions that do not possess any transient or dynamic attribute cannot possess punch. Thus, punch is both related to transient change and the energy density at a moment in time and duration. These changes can be as a result of any sound source present in the signal, not just the drums and bass.

Further to the above, dynamic change in particular frequency bands may contribute to the perception of punch perceived by the listener and thus, an effect will be apparent upon both the overall sound quality perception and level of punch. The literature review found no measures that incorporated such multi-band measurements except for the all-passed crest factor (Ljudtekniska , 2013).

The current ITU-R BS.1770-4 (2015) loudness algorithm is based on an integrative approach thus can't really be utilised to quantify attributes such as punch, as the key components relating to this perceptual measure exist in the 'microdynamic' scale of the signal.

Fine time-scale approaches have been developed to measure microdynamics (Skovenborg, 2014; Fenton et al., 2011) however these approaches still consider the signal as whole when calculating peak and average levels loudness levels at different resolutions. By whole, the complete complex mix is being considered during measurement.

This work is motivated by a need to separate the signal under test into what is considered to be steady state, transient and residual components, allowing individual analysis of each. By utilising signal separation the true dynamics of the signal can be analysed whilst considering the relationship between each on signal perception. What is being considered here is the ‘transientness’ of the signal where the peaks in the signal are solely related to the transient component and nothing else. This has advantages over an integrated approach whereby ‘microdynamics’ within a signal can be considered independently of overall loudness or summed peak level.

Automatic loudness normalisation by broadcasters may hopefully have an impact on lowering the proliferation of low dynamic range material being offered to the consumer however, there still appears to be a reluctance to embrace this in music production; the trend being that loudness level meters are simply being used to match loudness to ‘current’ released audio rather than to the proposed broadcast levels. This trend contradicts the artistic desire of releasing music that possesses both dynamic range and spaciousness, all of which can be somewhat destroyed through ‘target’ driven mastering and to some degree mixing.

With this in mind, a metering tool that would aid the mixing and mastering engineer to gauge this perceptual parameter would help them to meet artist preference rather than rely on loudness alone. Indeed, further metering tools that are tuned to specific parameters within the complex musical signal may be of benefit to engineers and consumers alike.

A number of experiments were conducted as part of this work and these are detailed in the chapters that follow. The experiments are presented in chronological order and include testing of key metrics previously outlined that show promise in mapping to the punch parameter.

## **Chapter 3 Objective measurement of music quality using multiband dynamic range analysis**

The aim of this chapter is to investigate the perceptual effects of wide band dynamic range compression with respect to listener perceived quality. The purpose of the experiment was to investigate and contrast the results with a multi-band based measurement.

A listening test (Fenton et al, 2009) was designed to measure the subjective preference of listeners to changes in dynamic range caused by the maximisation of an audio signal. The objective was to extract the degree of perceptual degradation that a signal maximisation process could cause and investigate correlation with the standard measurement of CF. In addition, a new measure called Inter-Band Ratio (IBR) was introduced and investigated which appears to show a closer correlation to listener quality perception than the wide band CF measure.

The experiment involved playing a selection of audio stimuli to the subjects and allowing them to compare them against an uncompressed reference stimulus. Each subject was asked to compare each stimulus to the reference and grade its quality on a seven-point sliding scale.

The reference stimulus was unprocessed whilst the audio stimuli consisted of 5 versions of each, with progressively reduced dynamic range (-6dB steps). The dynamic range reduction was achieved through the use of an L2-Maximisation plugin (Waves, 2008) with ProTools. The only control utilised in the maximization process was the threshold control, which was reduced in 6dB steps. The attack and release controls are controlled automatically within the plugin using a combination of look ahead compression and automatic release control depending on the source material.

One effect that occurs when the dynamic range of a musical piece is reduced is that its overall perceptual loudness is increased. This is due to the RMS level of the audio becoming normalised towards the overall peak level of the signal and as either the make-up gain or playback level is increased, a perceived loudness increase is heard. If the correlation

between the peak and RMS levels were calculated, this correlation would approach a value of 1 if the peak and RMS levels evolved together. (Deruty & Tardieu, 2014).

In order to avoid biasing effects caused by differences in loudness level each excerpt had its loudness normalised to that of the reference sample. Measurements were taken using a BS.1170 loudness meter and the overall gain of each stimulus was reduced until it equaled that of the reference signal. This process enabled the subjects to give scores based on the perception of quality associated with the reduction of dynamic range alone and not the loudness increase. Arguably, this supports the notion that the increase in quality afforded by a loudness increase can be obtained simply by turning the volume control up and hence the need for DRC is reduced.

The subjects were given a training phase prior to the experiments taking place, this was to allow subjects to familiarise themselves with the test setup and the audio excerpts they were expected to listen to. This training process helps to reduce the contraction biasing that may occur during the testing process (Zielinski & Rumsey, 2008).

The subject scores obtained from the tests were combined resulting in a Mean Subject Score (MSS) for each excerpt. The tests were performed using MATLAB and based upon an existing script developed for performing the ITU-R (2003) BS.1534-1 MUSHRA based tests (Vincent, 2005).

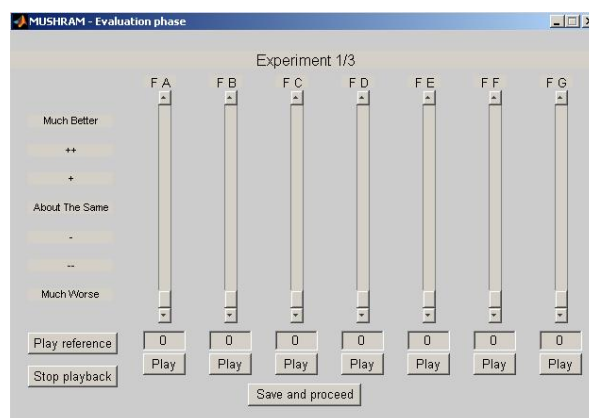
### **3.1 Test methodology**

Whilst the experiment was based upon the MUSHRA recommendation it was necessary to modify the test to facilitate the nature of the test being performed. The scales normally adopted in MUSHRA tests are specified as the five interval Continuous Quality Scale (CQS). This scale has intervals described from top to bottom as Excellent, Good, Fair, Poor and Bad. The sliders used by the user on these scales have an internal numerical representation in the range 0-100, where 0 corresponds with the bottom of the scale (Bad) and 100 with the top of the scale (Excellent).

The MUSHRA specification, ITU-R (2003) BS.1534-1, states that at least one of the excerpts under test should be a hidden reference, therefore its score should correspond to 100 when under test. This method, in conjunction with other hidden anchors is an attempt to gain consistent grading between subjects.

Whilst this scaling and numerical representation allows for the audio excerpts under test to be compared to the reference, it does not allow the subject to give a subjective quality score greater than that of the reference.

To accommodate this, the MUSHRA test was modified to incorporate a seven-interval scale to allow subjective scores to exceed that of the unprocessed reference. The interface is shown in Figure 6. In addition, the internal numerical representation range was increased to accommodate the larger seven-point scale, 0-140. ITU-R BS.1284-1 (2003) and ITU-T. P800 (1990) specifies a seven point scale called the Comparison Category Rating (CCR) ITU-T and this was used here. It has the advantage of allowing processing to be rated that either degrades or improves the quality. A score of 0 given by the subject would correspond to the bottom of the scale (Much Worse) and a score of 140 to the top of the scale (Much Better). A score of 70 would indicate that the listener thought the sample was the same as the reference.



**Figure 6 - Modified MUSHRA interface**

The length of the test was given consideration. The test methodology chosen enabled a large number (up to 15) test sounds to be evaluated alongside a single reference signal, thus keeping the test length to a minimum and ensuring fatigue of the listeners was not a biasing factor. Further to this, listeners utilise short-term memory whilst assessing music in qualitative tests (Koelsch & Siebel, 2005), therefore stimuli were limited to 7 seconds in length.

The audio excerpts were played back in random order during each experiment, thus every experiment can be classed as double blind with multiple stimulus, hidden reference and anchor.

### **3.1.1 *Biasing***

During any listening experiment, the effects of biasing must be taken into account in order to minimise their effects (Zielinski & Rumsey, 2008). The test interface was modified so that it did not contain any horizontal bars to prevent any interface bias effects. However, the interval scale remained to help the listener understand the grading process. As mentioned previously, the training process helps to reduce the contraction biasing that may occur during the testing process. In addition, the loudness of each excerpt was normalized to play back at a measured level of 72dB(A) so as to prevent this from being a factor contributing to the scores given by each subject.

### **3.1.2 *Stimuli***

After much consideration, three different audio stimuli were chosen, these were:

Excerpt 1 – “Acoustic Guitar” by Angus Barclay

Excerpt 2 – “Pop Music” by Eddie Rabbitt.

Excerpt 3 – “Dreadlock Holiday” by 10cc.

The stimuli were 16bit, 44.1kHz, stereo WAV format.

The reason for this choice of stimuli was to allow for a varied test set, thus testing the perception of the dynamic range across a number of different types of music, including transient and harmonically rich material.

The acoustic guitar was recorded by the author using an Audio Technica AT4033 large diaphragm condenser microphone and a Rode NT2 (Mk1) large diaphragm condenser microphone. No mastering (final bus compression) of the recordings took place. Pre-amps utilised for the recordings were Calrec (M-Series) PQ1789s.

The Eddie Rabbitt stimulus was obtained from the EBU SQAM test CD (SQAM, 2005). As such it can be considered a standard stimulus for subjective testing. In the context of this study it is well suited as it contains a main vocal line, is well balanced and has not been subjected to any excessive DRC.

Dreadlock Holiday by 10cc was chosen as it represents a produced piece of music that hasn't been subjected to what could be called over compression. The song, released in July 1978, could be considered an album that avoided the forthcoming 'loudness wars' that commenced around the mid-late 1980's and is perhaps one that would be familiar to most experienced listeners.

The tests took place in a critical listening room in the University of Huddersfield utilising a PC with a Realtek HD sound card. All the excerpts were auditioned on Sennheiser HD650 headphones and therefore biasing effects caused by both room acoustics and background noise were eliminated.

### **3.1.3    *Test subjects***

A total of 10 test subjects participated in the experiment. All were experienced listeners. These were selected from University staff members, engineers and music producers, and doctoral and final year students.

The listeners were pre-screened to ensure that they were suitable to take part in such a test. The pre-screening involved the subjects taking part in both a hearing test and listening experiment to determine that they were a) sound of hearing and b) could detect impairments in audio excerpts that had been subjected to processing.

Each subject was given a training phase in which they could play all the samples selected, which included the reference and anchors. This enabled the listeners to familiarise themselves with the audio being presented.

Each subject, following the training phase, was given an explanation of the experiment and was told to listen to the excerpts and grade each with respect to the reference in terms of ‘overall quality’. A hand out was also given to each subject also detailing the test and guidelines.

### 3.2 Discussion

In total 21 audio excerpts were listened to and graded by each subject. Scores for each experiment were collected and collated by order of maximisation level and stimuli type.

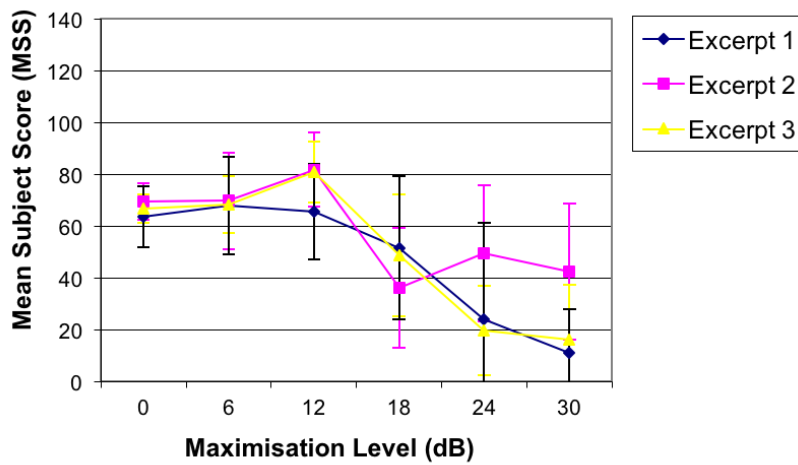
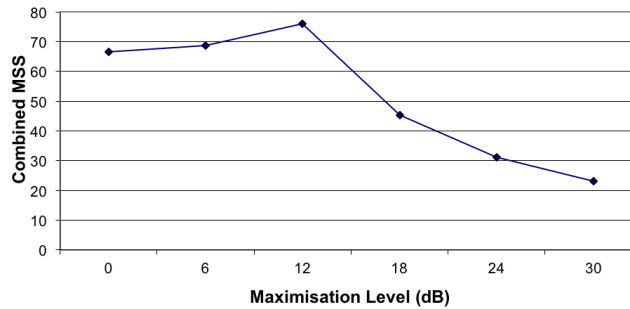


Figure 7 - Mean Subject Scores vs. Maximisation Level

Figure 7 shows the MSS vs. Maximisation level. Maximisation level 1 corresponds to the reference (i.e. no compression applied) and maximisation is applied in steps of 6dB. The results suggest that quality degrades as increasing levels of maximisation are applied. In all cases, the 3.5kHz low pass filtered anchor was rated as ‘worst’ quality by the panel and isn’t included in these plots. The whiskers on the plot show 95% confidence in intervals in the means scores achieved.

Figure 8 shows the combined MSS for each maximisation level when the differences between audio stimuli have been disregarded, i.e. the scores are combined.



**Figure 8 - Combined MSS vs. Maximisation Level**

A MSS of 70 represented a rating whereby the subject rated the stimuli as being ‘About the same’ quality to the reference.

One can observe a slight increase in the MSS as the DRC level is increased upto the 12dB point. This appears to contradict the notion that listeners might prefer a wider dynamic range in music production. Indeed, it seems that the listeners preferred a level of DRC ‘up to a point’. Movements such as the ‘Pleasurize Music Foundation’ (Pleasurize, n.d) advocate the maximum use of dynamics within music production. The maximum MSS value of 75.29 equates to a mean 7.56% increase in perceived audio quality from the reference, based on the listeners’ subjective perception of quality. As one can see, the general trend is an almost linear reduction on MSS as the maximisation level is increased beyond 12dB. Figure 7 shows that excerpt 2 is the exception to this, showing a further increase in quality at maximisation level of 24dB and 30dB

The reference, in all cases, does not appear to be associated with maximum quality according to our test panel.

A 2-way analysis of variance test (ANOVA) was performed on the data in order to determine the significance of each test factor – i.e. excerpt and dynamic range reduction. The ANOVA results from the study (Table 3) indicated that the effect of the reduction in dynamic range (labelled “Rows”) is highly significant ( $p=0$ ). This is a strong indication that the subjects consistently perceive a change in quality as the dynamic range of the samples is varied. In addition, the effect of the audio excerpt (labelled “Columns”) could be considered as being significant ( $p<0.05$ ), suggesting that the particular excerpts used have some influence on how subjects rated the quality of perceived audio across the different maximisation levels. However, this marginal result, with such a low F-ratio from the ANOVA and combined with a significant level of interaction between excerpt and dynamic range ( $p=0.0026$ ) make a generalisation of results somewhat difficult.

Source	SS	df	MS	F	Prob>F
Columns	2756.9	2	1378.4	3.79	0.0243
Rows	123968.1	6	20661.4	56.8	0
Interaction	11598.3	12	966.5	2.66	0.0026
Error	68753.5	189	363.8		
Total	207076.8	209			

**Table 3 - 2 Way ANOVA Test**

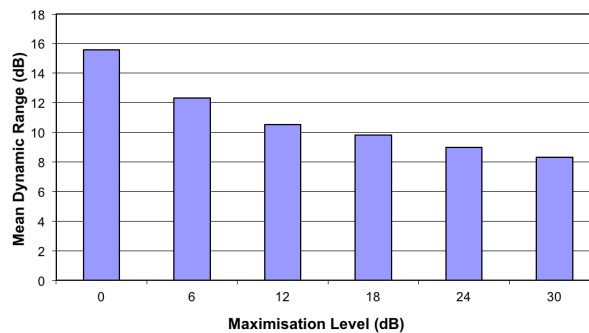
### 3.3 Dynamic Range Analysis

One could argue that the peak levels of each of the excerpts, prior to any maximisation process being applied, would dictate the overall reduction of dynamic range achieved during maximisation, and indeed they do. All three excerpts used contained differing peak signal levels, however, given the results shown in Figure 8 and the results of the ANOVA test, one can observe that there is some correlation between the maximisation level MSS, irrespective of excerpt and therefore peak level in each case.

### 3.3.1 Wideband dynamic range measurement

Wideband dynamic range (WDR) is the dynamic range measured when considering the entire audio file and its full bandwidth. Dynamic range in this study was calculated using CF and the value converted into decibels.

In the same way that the MSS was combined for each maximisation level, the WDR of each stimulus was combined to indicate a ‘mean’ dynamic range reduction taking place (Figure 9).



**Figure 9 - Mean Dynamic Range vs. Maximisation Level**

Using Figure 8 and Figure 9 an optimal mean dynamic range was derived corresponding to a WDR of 10.51dB. 30dB maximisation corresponded to a mean WDR of 8.32dB. With reference to Figure 9, showing the MSS at each maximisation level, it appeared that maximisation of 12dB (WDR of 10.51dB) was preferred; suggesting that compressing the WDR by more than this value was undesirable. Interestingly, this maximisation level is also shown to be preferred over 0dB and 6dB, having mean WDR values of 15.59dB and 12.33dB respectively.

### 3.3.2 *Multiband dynamic range*

From the results obtained, it was apparent that listeners preferred some level of dynamic range compression rather than none at all. However, there is a preference towards uncompressed when more extreme levels of maximisation are applied. This concurs with a later study made by Croghan et al (2012). In order to examine these preferences further the dynamic range across critical bands was investigated and the interaction of each band against the combined MSS was observed.

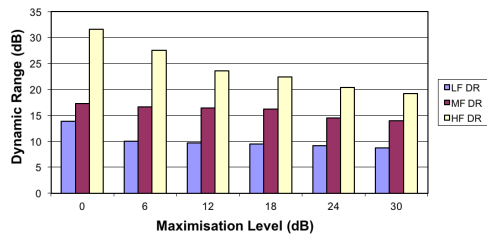
In the study each excerpt was filtered using a 3 band linear phase FIR filter. Their respective cut-off frequencies are shown in Table 4. All filters were 24dB/Octave.

Filter Type	Lower Fc (Hz)	Upper Fc (Hz)
Band Pass LF	20	947
Band Pass MF	947	3186
Band Pass HF	3186	15447

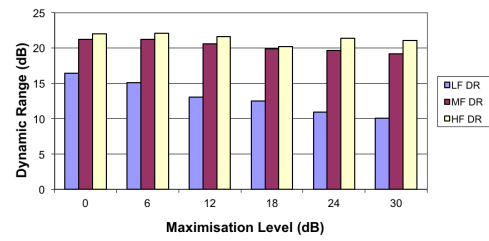
**Table 4 - Three Band Filter Corner Frequencies**

These frequencies were chosen as they approximate to the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> set of 8 Bark scale critical bands in the auditory system.

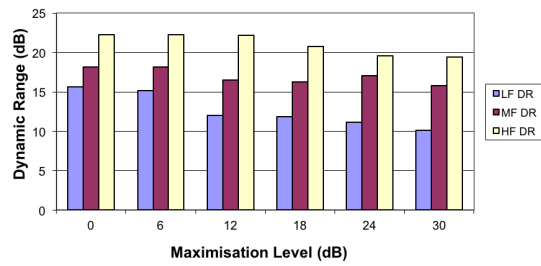
Following this filtering process, dynamic range was measured for each band. The three measurements obtained across the three bands form what is referred to as Multiband Dynamic Range (MDR). Figure 10, Figure 11 and Figure 12 show the dynamic ranges within each frequency band, LF, MF and HF respectively.



**Figure 10 - Ex1 MDR vs. Maximisation Level**



**Figure 11 - Ex2 MDR vs. Maximisation Level**



**Figure 12 - Ex3 MDR vs. Maximisation Level**

The study considers that the general trend of frequency balance within produced music follows that of the response of the ear i.e. the mid to high frequencies will be balanced at a lower level than that of the low frequencies. As a result, it is assumed that a loss of low frequency headroom would be experienced prior to other bands as DRC was applied. In the previous three figures this can be observed by the larger downward trend present in the low-frequency band as the maximisation level is increased when compared to the mid and high frequency bands.

Low frequency content of produced pieces of music contribute greatly to the spectral energy of the piece, therefore a loss in this energy could result in a perceptual loss of audio quality by the subject. Indeed, overall warmth perception is likely to be affected which may also contribute to a loss of perceptual quality.

### 3.3.3 *Inter-Band Ratio (IBR)*

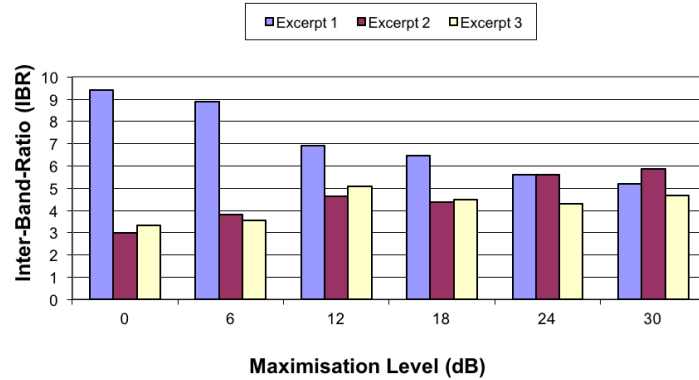
In order to investigate and quantify the relationship between the dynamic ranges in each frequency band, a simple standard deviation equation was adopted to form an Inter-Band Ratio measure (IBR).

$$IBR = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Dr_i - \overline{Dr})^2} \quad (9)$$

Where  $Dr_i$  represents the dynamic range calculated for band  $i$ .  $\overline{Dr}$  represents the mean dynamic range.

The IBR measure (standard deviation of dynamic range between each band) is suggested as an alternative to WDR based on the measurements made in the study. By plotting the IBR measure (Figure 13), the study indicated that there was a trend of deviation increase up until the point of 12dB maximisation, this was in contrast to the ever declining WDR figure. In the case of excerpt 2 & 3, a Pearson correlation test confirmed this with correlation coefficients of 0.89 and 0.99 respectively when considering up to the 12dB maximisation level.

Whilst the WDR measure indicates an overall reduction in dynamics is occurring during the maximisation process, it's concluded that it isn't a measure that can be utilised to indicate listener preference.



**Figure 13 – Inter-Band-Ratio vs. Maximisation Level**

### 3.4 Conclusions

The experiment represented a pilot study into the effects of dynamic range reduction on the perception and measurement of audio quality. The results suggest that a multiband approach to dynamic range measurement might be more effective than a wide band approach as an indicator of overall audio quality. Ljudtekniska (2013) later incorporated a multi-band approach that also focused on the CF measurement across multiple bands, albeit using all-pass filters to separate frequency bands but inter-band correlation wasn't considered in his measure.

Low frequency content of produced pieces of music contribute greatly to the spectral energy of the piece, therefore a loss in this energy could result in a perceptual loss of audio quality being perceived by the listener. Maempel & Gawlick (2009) suggest that listeners prefer increased loudness however underlying perceptual effects, enhanced or created by the processing involved, may be influential in listener preference. In their study, high bass amplitude was cited as a cause.

As observed in this study, two of the three excerpts exhibited a reduction if LF dynamic range to a greater degree than other bands, as the maximisation process took place.

Due to the wide variation in spectral content between pieces of produced music, in addition to fade-outs and fade-ins a single WDR figure is not accurate enough to quantify overall music quality. It may however, be utilised as a general 'figure of merit' score.

## **Chapter 4 Inter-Band Ratio and music quality perception.**

The experiment detailed in Chapter 3 identified a potential correlation between the Inter-Band Ratio (IBR) measure and the subjective quality of produced music. The testing was performed with 3 audio excerpts and a small number of listeners. In order to test the IBR measure more rigorously, and to establish its relationship to listener perception of music, a more comprehensive experiment was undertaken. This involved testing the IBR with real-world music excerpts and a greater number of listening subjects.

It is widely accepted that the response of the human ear and therefore listener perception differs across the ear's frequency range. It is therefore argued that a single wideband dynamic range figure would be inaccurate in reflecting the perceived audio quality of a signal, although it could be used to represent an overall mean 'figure of merit' score. This was shown to be the case in the previous chapter. Additionally, Croghan et al (2012), using genres of 'rock' and 'classical', found that listeners preferred stimuli where DRC had been applied moderately.

### **4.1 Test Methodology**

Analysis was made of the dynamic range in three critical bands and the interaction of each against the overall MSS was examined. Each excerpt was filtered using a 3 band linear phase FIR filter. Three filters were used and their respective cut-off frequencies and Q settings were the same as utilised in Chapter 3 (see Table 4 - Three Band Filter Corner Frequencies).

Following this filtering process, dynamic range analysis was performed. Calculations of the dynamic range were derived from the samples in each band as follows:

$$Srms = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad (10)$$

Where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (11)$$

and  $x_i$  represents the value of sample  $i$  between +1 and -1.

$$Spk = \max (x_{1..n}) \quad (12)$$

$$Dr = 20 * \log \left( \frac{Spk}{Srms} \right) \quad (13)$$

The Inter-Band Ratio (IBR) is derived in the same way as outlined in Chapter 3 using Equation 9.

Effectively, the IBR represents the standard deviation measured between the dynamic ranges existing across the three bands. A low value would represent a smaller variation in dynamic range measured across the bands, whereas higher values would represent higher degrees of variation.

Factors that may contribute to a low IBR score are the application of wide band hard-limiting. This could cause the dynamic ranges across bands to normalise and cause a perceptually less dynamic production. Particularly ‘dense’ sections of an arrangement in terms of them having similar broadband frequency content could also result in low IBR measures, for example, broadband noise-like signals that result in similar peak and RMS values.

High IBR scores can be measured when there is a de-correlation of dynamic ranges between bands. This may be as a result of strong transient content being present in say the mid-range, whilst the dynamic of the low band may be small in comparison.

#### **4.1.1    *Subjective listening test***

This study involved 57 experienced listeners listening to and grading 5 excerpts from commercially available produced music. A listening test was designed to measure the subjective preference of listeners grading produced music of varying genre. No knowledge of the engineering and production techniques involved with any excerpts was made available to the listeners.

The listeners were asked to listen to each excerpt with respect to analysing their relative punch, clarity, overall tone and balance. Each listener was then asked to grade each of the excerpts out of 10 with respect to their overall production quality (1 being low quality, 10 being the highest quality).

The listening conditions and equipment varied between subjects however, this was deemed satisfactory for such a wider study. This would in fact, be how the music would be received and auditioned by the general population of listeners. Whilst the listening level between subjects was not maintained, the relative listening levels between excerpts could be assumed to be constant, thus the quality scores given by each of the subjects would be relative between excerpts.

A hand out was given to each subject detailing the test and guidelines.

#### **4.1.2    *Stimuli***

The five excerpts were from the following songs, each of which can be classed as contemporary productions.

Excerpt 1: “The End Of The Line” by Metallica

Excerpt 2: “Mr Brightside” by The Killers

Excerpt 3: “Freak Like Me” by Sugababes

Excerpt 4: “Animals” by Nickelback

Excerpt 5: “Seldom Seen Kid” by Elbow

The excerpts were 16bit, 44.1kHz, stereo WAV format and 20 Seconds in length.

The reason for the choice of excerpts was to allow for a varied test set, thus testing the perception of the dynamic range across a number of different styles of music, including transient and harmonically rich material. Given that all the excerpts were released on major labels, it can be assumed that all had been professionally mixed and mastered.

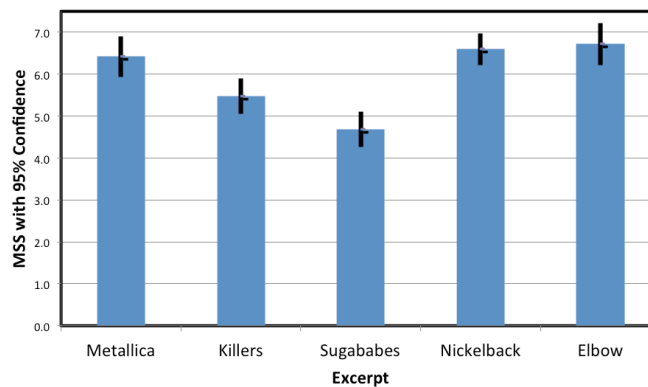
Elbow were advocates of the ‘Turn It Up’ movement, therefore one can expect a greater level of dynamics to be present in the “Seldom Seen Kid” excerpt.

## 4.2 Results

The subjective scores obtained during the test were averaged resulting in a MSS for each excerpt. From this, an order of preference (ranked best quality to worst) of the excerpts was extracted.

### 4.2.1 Listening test results

The results of the test are shown in Figure 14 along with 95% confidence intervals.



**Figure 14 - Mean Subject Scores**

Excerpt	Rank	Title
5	1st	"Seldom Seen Kid" by Elbow
4	2nd	"Animals" by Nickelback
1	3rd	"The End Of The Line" by Metallica
2	4th	"Mr Brightside" by The Killers
3	5th	"Freak Like Me" by Sugababes

**Table 5 – Rank Score Order of Excerpts**

It should be noted that whilst an order of preference was extracted (Table 5), the scores given for the top three placed excerpts were indeed very closely grouped. Looking at the 95% confidence intervals shown in Figure 14, there is a high degree of overlap between the scores obtained by the top three placed excerpts.

Reasons for this could be due to biasing factors in genre preference in addition to all excerpts being professionally produced and mastered, thus making their relative ranking scores group together.

At this stage we could consider relaxing the ranking and suggest that this test has extracted the top three, 2<sup>nd</sup> place and last place excerpts in terms of audio quality. Indeed, if we consider the 95% confidence intervals shown in Figure 14, there is a clear differentiation between the tiers identified.

Considering that the excerpts are all 'release' quality one would expect that all the excerpts would achieve a relatively high MSS. However, there is clearly some differentiation between their perceived audio qualities reflected by the different MSS received. In order to establish whether the variation of preference scores achieved were due to the excerpts under test or simply down to chance an analysis of variance was undertaken with the following null hypothesis:

$H_0$  = There is no difference in audio quality between excerpts.

$H_1$  = There is a difference in audio quality between the excerpts.

The results of an ANOVA test for this (see Table 6) show that the  $F$  value is highly significant and the resultant  $p$ -value obtained is  $\ll 0.01$ , suggesting that the differences found in the MSS across the samples is more than would be expected by chance alone.

The  $F$  value obtained is significantly larger than the  $F_{crit}$  value therefore we can reject the null hypothesis with a large degree of confidence.

ANOVA						
Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	174.225	4	43.556	15.1	3.43E-11	2.404
Within Groups	807.649	280	2.8846			
Total	981.874	284				

**Table 6 - ANOVA table**

#### **4.2.2 Objective results**

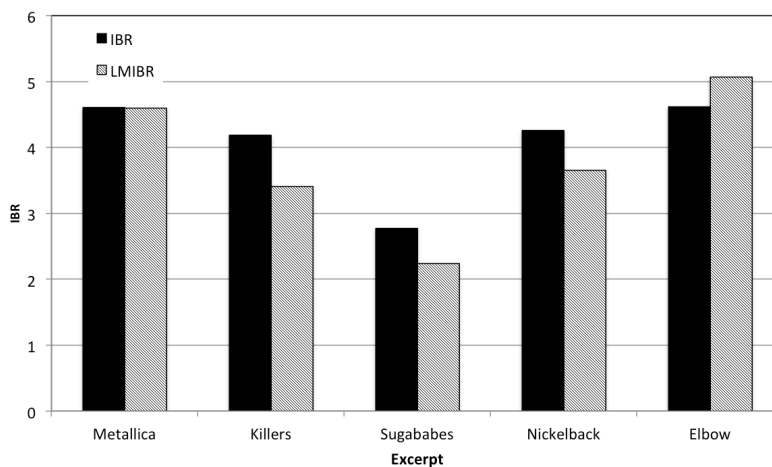
The excerpts were each analysed with respect to the IBR as detailed in Section 4.1. The results of the analysis are shown in Figure 15. The calculation of the IBR measure was based on the entire sample length.

The order of the excerpts, based on the largest to smallest IBR measurement, was extracted. The order is shown in Table 7.

Excerpt	Rank	Title
5	1st	Seldom Seen Kid” by Elbow
1	2nd	“The End Of The Line” by Metallica
4	3rd	“Animals” by Nickelback
2	4th	“Mr Brightside” by The Killers
3	5th	“Freak Like Me” by Sugababes

**Table 7 – Rank Score Order based on IBR**

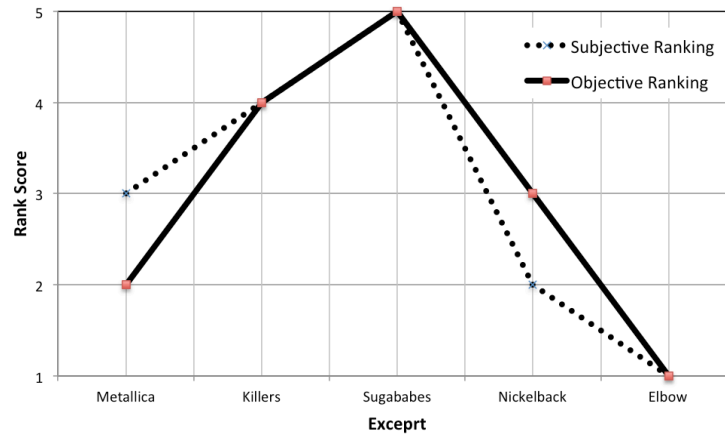
A second measurement, based on calculating the IBR based only on the Low-Mid bands was extracted. This is shown in Figure 15 as LMIBR.



**Figure 15 - Inter-Band Ratio Measurements**

### 4.3 Discussion of results

If the extracted subjective and objective rank order scores are compared (Figure 16) a large degree of correlation between the two can be observed.



**Figure 16 - Subjective & Objective Rank Order Comparison**

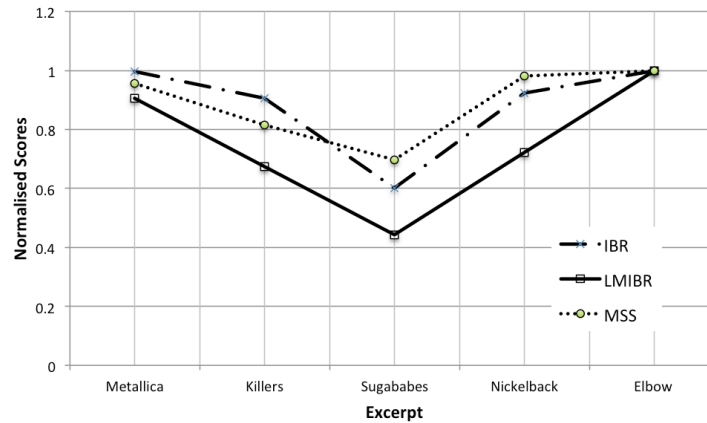
Running a Spearman correlation test between the subjective and objectively extracted rank scores a correlation coefficient of 0.9 is obtained along with a significance value of 0.037, this is significant at the 0.05 level (2- tailed)

Killers and Sugababes were all ranked identically with placements 4<sup>th</sup> and 5<sup>th</sup> in order of quality. Therefore, the IBR measurement successfully identified the 2 excerpts that were graded as having the lowest quality subjectively by the listeners.

The top three excerpt rankings based on the IBR score and the MSS differed with Metallica and Nickelback being reversed in order. This was the only difference in ranking score.

Differences in the placement of Nickelback and Metallica shown in Figure 16 could be attributed to biasing due to personal preference in genre. Also, subjectively, the top three excerpts scored very similarly with a degree of overlap evident in the confidence intervals. The excerpts themselves contained variation in transient content and tempo. These factors could be affecting the overall rankings given by the subjects. The order of playback of excerpts was left up to the subjects therefore allowing continuous and multiple comparisons to be made between each excerpt.

If one normalises both the mean subjective scores and the IBR score for each of the excerpts tested, the correlation between the two can be observed, see Figure 17.



**Figure 17 - Normalised Subjective & Objective Scores vs. Excerpt**

If we consider the deviation between low and mid-range only, shown on the plots as LMIBR, the objective quality ranking becomes more pronounced.

The low and mid-range frequency bands could therefore be attributable to having a greater effect on the perception of quality in music production. Low and mid-range frequencies are certainly attributable to energy and warmth being present in a mix. Interestingly, by examining the dynamic range measurements made on each of the excerpts across the three bands (see Table 8), it can be seen that whilst the lowest placed excerpt has a greater degree of dynamic range in its low frequency band, its IBR score is the lowest due to its mid and high band correlation.

Excerpt	LF Dr	MF Dr	HF Dr	IBR	LMIBR
Metallica	8.1577	14.6563	17.0584	4.6048	4.5952
Killers	10.3824	15.2067	18.7134	4.1828	3.4113
Sugababes	11.7525	14.9279	17.2871	2.7773	2.2453
Nickelback	9.8965	15.0646	18.3508	4.2619	3.6544
Elbow	8.7291	15.8984	17.3616	4.6199	5.0695

**Table 8 - Dynamic Range and IBR Measures**

With reference to Table 8, the low frequency band dynamic range of the two lowest ranking excerpts is greater than that of the top three ranked excerpts. Considering the top three ranked excerpts, the differential between their respective low and mid-high frequency band ranges is

greater, resulting in a greater IBR score. Further to that, a reduced dynamic range in the low frequency band suggests a higher level of compression may have been applied to these productions.

This could suggest the importance of controlling the dynamics of the low frequency bands, often referred to as ‘tightening’ resulting in a subjectively powerful / punchy mix. This technique, common to the Rock/Metal genre, would probably have been applied to Metallica & Nickelback. Interestingly, the Elbow sample, despite not being in the same genre, was a production that attempted to adhere to the ‘Turn It Up’ movement by keeping as much of the dynamic range intact during mastering.

Low frequency content of produced pieces of music contribute greatly to the spectral energy of the piece, therefore a loss in this energy could result in a perceptual loss of audio quality by the subject.

As the IBR measure is based upon frequency band correlation, it can be assumed that during fade ins and outs the relative measure between peak and RMS levels across all frequency bands used in the calculation would remain constant. Therefore, the resulting IBR score would be unaffected. In addition, since the IBR score is derived by measuring the correlation between relative dynamic ranges across frequency bands, it is unaffected by the overall playback level selected by the listener. Thus allowing for a qualitative measure to be made prior to amplification taking place.

## 4.4 Conclusions

This experiment presented a wider study into the IBR model output variable. The results of the test indicate that the IBR measure could be effective in the assessment of audio quality.

A reduction in dynamic range in a single frequency band does not necessarily result in a perception of low quality by the listener; rather, the relationship between the dynamic ranges in bands has been shown to correlate to this score. Reduced dynamic range in the lower band may have afforded an increase in perceptual warmth, punch or tightness being perceived by the listeners.

For each excerpt a single IBR measure was extracted, this utilized a fixed window size based on the entire sample length. Previous work by Skovenborg & Lund (2008) proposed a measure of ‘consistency’ which measures *the variation of loudness on a macroscopic timescale*. They describe the application of a loudness-correction processor increasing the consistency of the musical material. Empirical study suggests that this is the typical approach applied to music during the mastering stage and often results in loss of dynamic range across the frequency range.

Their measure of consistency is based upon statistical distribution of measured loudness utilising the ITU-R BS.1770-4 (2015) loudness algorithm as a starting point. It uses a statistical distribution to prevent the measure from being skewed by short but loud musical sequences and/or fadeouts. In order to achieve this, a smaller measurement window size is employed along with an overlap. The basic IBR method does not employ any method of statistical distribution or weighting, therefore shortcomings would be that ‘micro dynamic’ changes would be obscured by the averaging process.

In the following chapter, an experiment was employed to investigate the use of a windowed IBR methodology and to investigate the possible uses of a statistical based output.

## **Chapter 5 Profiling of punch and clarity using Inter-Band Ratio**

The results outlined in Chapter 4 indicated that there was a potential correlation between the IBR measure and the listener perception of overall quality of produced music excerpts. This chapter details a test (Fenton & Wakefield, 2012) which extends and evaluates the IBR measure temporally. Clarity is considered relevant in this study as lack of clarity in music could relate to the inability of a listener to clearly define transient components within it. Without transients, and in line with the hypothesis of this thesis, punch would not exist.

### **5.1 Clarity and punch in music production**

In acoustic space it is possible to objectively determine the clarity and intelligibility achievable as an alternative to the traditional RT60 measurement (Ballou, 2005). Measures such as early to late arriving sound ratio (C50) and Early Decay Time (EDT) can all be combined for this purpose however, if one considers a completed music production, the task becomes very difficult for two key reasons.

Firstly, clarity in a musical extract becomes somewhat subjective in context of the production. For example, a classical 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 solely on the clarity of transients perceived by the listener as a result of the drum/percussive elements. Spaciousness may not be as important in the context of the rock recording whilst this attribute could be perceived by some as an important aspect to obtaining clarity in the classical example.

Secondly, due to the combination of spectral components from a number of sources, it's very difficult to utilise traditional acoustic measures to grade the extract. Blind source separation (Barry et al, 2005; Every, 2008) could be useful in determining overall clarity of individual instrumentation contained within a production however, this remains a complex process and often the sources extracted contain residual artefacts. Their use in qualitative measures is therefore limited.

In order for listeners to detect individual notes, instrument timbre and rhythm, it's important that enough elements of the mix conveying this information are clearly audible. Masking is a phenomenon that often occurs during mix down when harmonic components of one source mask that of another source. Masking can occur in the temporal domain in addition to the frequency domain.

Considering the spectral nature of the individual sound sources during mix down, it is possible to determine the level of masking taking place and modify the relative balance between sources to minimise this (Gonzalez & Reiss, 2009). In a musical context, this anti-masking process will allow the listener to clearly hear the individual sound sources and thus, the overall production could be deemed to have higher clarity. Due to the varying nature of audio, and moreover the harmonic content within each sound source, this process is not without its difficulties and the process of spectral balance is left to the skill of the engineer.

A reduction in dynamic range, as a result of applying maximisation/compression techniques, has the effect of raising the noise floor whilst at the same time increasing the level of spectral components contained within a source that were previously balanced in relation to their counterparts. Therefore, the process can cause additional masking to occur. This can occur whether the compression is being applied to individual tracks within a mix or the entire mix itself. This is perhaps one of the reasons lower subjective scores were given to audio samples that had been subjected to high compression levels in the previous studies conducted in the earlier experiments.

In contrast, reverb tails and other sources that may be relevant to depth cues such as those mixed lower in the mix, could become more perceptible to the listener, therefore resulting in a subjectively more pleasing mix. Thus, moderate levels of compression may in fact be beneficial, in line with previous findings. On the other hand, Hjortkjaer & Walther-Hansen (2014) stated in their study that there was a weak link between perceptual depth cues and the influence of compression. Unfortunately, their study didn't identify or state what depth was other than the subjective ability to discriminate the musical sounds.

Whilst anti-masking plays an important role in determining the ability of 'tonality of sources' to be clearly defined in a music production, onset or transient detection is also key (Every, 2008).

Within a musical context, temporal changes in frequency component amplitudes within the piece allow us to detect instrumentation (Lagrange, 2009). A produced piece of music must contain various elements of information, which include instrumentation and transient content in order to convey such things as emotion, and energy in the piece.

Dynamic range takes a leading role in allowing these elements to play their role in this process. If dynamic range is reduced, perhaps through excessive use of compression, important information in the piece is detrimentally affected, in particular transient information.

The ability of the listener to detect transients in a piece of music is fundamental to the determination of instrument type, note detection and rhythm. There are a number of automatic methods that have been proposed and evaluated (Bello et al, 2005; Hainsworth & Macleod, 2003) that attempt to detect onsets (and subsequent transients). These can include computation in the time, frequency and phase domains. The use of these techniques is often employed in beat extraction to determine rhythm, instrument identification and genre classification.

With respect to polyphonic music productions, where there could be a number of competing audio sources in the overall spectrum and thus overlapping attributes, some onsets events could be promoted as being more important than others (Collins, 2005) and/or wrongly identified. Thus, in order for highly accurate onset detection to take place, a complex algorithm is often required that utilises, for example, particular frequency bands for analysis of different onset types. Even in these cases, current onset detection algorithm performance varies when presented with near simultaneous events and in-distinct spectral signatures.

The previous experiments conducted have shown that a reduction in dynamic range on a piece of music does have an impact on the overall subjective qualitative score given by listeners. Given that transients within a production can be affected by this dynamic range reduction, it holds true that their associated measurement and detection, both by the listener and by objective measurement may be affected.

By monitoring the temporal changes in dynamic range across three key frequency bands, representing bass, mid and treble from a production viewpoint, it is proposed that the method may yield results that relate to the perception of clarity within a completed musical production.

Further to that, in line with the hypothesis stated in Section 1.2, dynamic change in particular frequency bands may contribute to the perception of punch indicated by the listener. By measuring the magnitude of change in dynamic range across bands and temporally, these changes may correlate with the punch attribute.

## **5.2 Temporal Inter-Band Ratio**

By plotting the IBR measure against time, relative dynamic content *across* frequency bands can be observed and allow engineers to identify sections of audio that possess differing dynamic attributes. These attributes may correlate with both clarity and punch.

The IBR is a standard deviation score, such that strong correlation in dynamic range across the bands yields a low IBR score and vice versa. When calculating the IBR score it's also important to consider the dynamic range measurement itself within the time frame in which the IBR is measured. For example, where high dynamic range is measured across all bands this results in a strong correlation and therefore a low IBR score would result. This may indicate a fast transient within the windowed time frame and/or a significant change in loudness level across all frequency bands within that time frame. On the other hand, a very low dynamic range measurement across all the bands will also result in a low IBR score. Low scores indicate correlation, therefore in the latter example this would perhaps indicate very loud or steady-state passages of audio. Therefore, in addition to plotting the IBR temporally, it would be useful to collate all the IBR frame scores and look at them statistically thus giving an indication of the overall underlying nature of the dynamics present in the audio. As a single point in time can't be considered as having dynamic range, the windowed time frame will also have an impact on the measurements obtained. This is the case for both the current ITU-R BS.1770-4 (2015) loudness model and other integrative models such as the LDR proposed by Skovenborg (2014).

Previous experiments outlined in Chapter 3 and Chapter 4 identified a correlation between the IBR and the subjective scores given by subjects with respect to overall quality. In those studies, excerpts tested were 7 seconds in length and the IBR score for each excerpt was calculated using a 7 second window size.

If the variable nature of musical content over time is considered, it is highly likely that the IBR measurements would vary as different window sizes are utilised and at different points in the

audio under test. As such, this work investigates this by profiling two songs using varying window sizes to calculate the IBR. These profiles are then compared to subjective listening test results where the listeners were asked to identify and map the audio with respect to their possible points of perceived punch and clarity. The IBR plots are also compared to ITU-R BS.1770-4 (2015) loudness measurements obtained using a NuGen Vis-LM loudness meter (VisLM, 2012).

The aim of this work was to identify trends and relationships between the dynamics contained within a musical signal, its temporal measure and how this relates to the perception of clarity and punch throughout the piece. In addition, the IBR frames are collated, which allows statistical analysis of the stimuli to take place.

### **5.3 Method of testing**

#### **5.3.1 *Subjective testing***

A listening test was conducted comprising 8 expert listeners. Each listener was asked to listen to 6 stimuli and score them along a time axis with a 400ms resolution. The listeners were given a timeline plot of each audio sample segmented into 400ms blocks and asked to score each block. The stimuli was played back using a DAW which allowed the listeners to re-audition sections as required. This block size used relates to the short-term window integration time defined in the ITU-R BS.1770-4 (2015) standard. Resolutions less than 400ms were deemed impractical with respect to how the listeners would enter their grades.

This test allowed the users to score the audio according to their own perception. However, in order to ensure some consistency in scoring and allow the data to be collated and averaged, the listeners were given a training sheet detailing the attributes that were to be assessed within the audio. The points detailed on the training sheet were as follows:

If the audio is clear and punchy - Give a score of 1. This can be defined if 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.

If the audio is unfocussed, lacks punch and clarity - Give a score of 0. This may consist of a large collection of harmonics or unrelated frequency components, noise, there is evidence of masking, no single element is clear, no dynamics present, distinct lack of transient content.

If you feel the audio at any point is neither of the above, give a score of 0.5.

All scores given by the subjects were then averaged to produce a Punch/Clarity score for each excerpt, varying between 0 and 1. This formed the subjective data to compare to the objective IBR measures.

The listening test took place in a professional control room environment, commonly found in music studios and the excerpts were auditioned on Genelec 8040 speakers at an average listening level of 74dB(A). The results of the listening tests were collated as an average punch/clarity score (P/C Average) and a profile plot was created which represented the perceived punch and clarity of each excerpt.

### **5.3.2 Objective testing**

MATLAB was used to calculate the IBR using window sizes of 400ms, and 3s respectively. A relative IBR ‘threshold’ was chosen based on initial studies (Chapter 3 and Chapter 4) and IBR scores attained in the ‘best’ scoring excerpts, in this case 4 or more. This value was based on the average IBR score attained for the two highest scoring excerpts in previous tests. The threshold determines the point at which the IBR score is deemed large enough to relate to a significant de-correlation in dynamics across the frequency bands tested. Choosing a lower threshold would mean the resultant IBR score would increase if smaller deviations were apparent across the bands, a larger threshold would only increase the IBR score if larger deviations were evident.

Where there was a significantly low score for the IBR throughout the excerpt, this threshold was adjusted to accommodate the reduced dynamic correlation.

The IBR values for each window size were measured against time. In addition, measurements were taken which detailed the short and long term loudness and loudness variation against time.

For all IBR measures, a 75% window overlap was adopted unless otherwise specified.

An overall IBR score was calculated by taking the average of the two window size measurements such that:

- *Neither IBR above threshold then the IBR = 0*
- *One of the windowed IBR scores exceeds the threshold then the IBR = 0.5*
- *Both Windowed IBR scores exceed threshold then the IBR = 1*

From this calculation, an objective profile was produced which represented a moving window average based on the two IBR averages. The objective IBR scores were then compared to their subjective counterparts.

In ‘Measures Of Microdynamics’ by Skovenborg (2014) a measurement of Loudness Dynamic Range (LDR) is proposed. That study utilised the maximum difference between a “fast” and a “slow” loudness level. This temporal IBR approach is similar with respect to the utilisation of two window sizes however, the key differences are summarised as:

- *LDR uses a temporally integrated loudness measure based on the entire frequency spectrum.*
- *IBR uses a temporally integrated measure based upon dynamic range correlation across frequency bands.*

### **5.3.3 Stimuli**

2 different audio stimuli were chosen

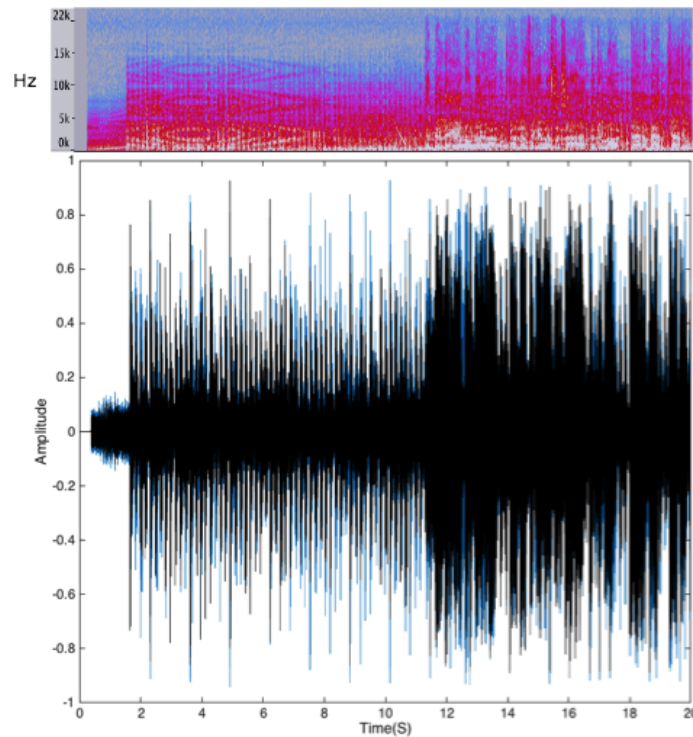
- *Excerpt 1 – “Freak Like Me” by Sugababes*
- *Excerpt 2 – “Animals” by Nickelback*

Both were in 16bit, 44.1kHz, stereo WAV format.

The reason for this choice was to allow for a varied test set based upon the best and worst performing productions in the previous experiment. The Sugababes excerpt was considered to be the worst overall and the Nickelback excerpt as one of the best both subjectively and objectively based on having the highest average IBR measure.

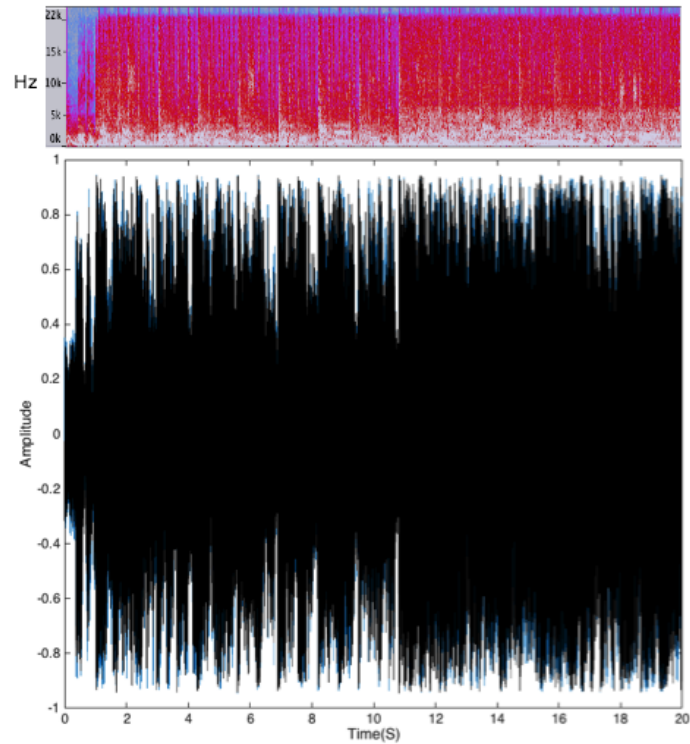
The songs were broken down into three 20-second excerpts that represent key sections in the production: introduction, verse with drums and breakdown. In order to familiarise the reader with the productions the excerpts are now described.

#### 5.3.3.1 *Sugababes excerpts*



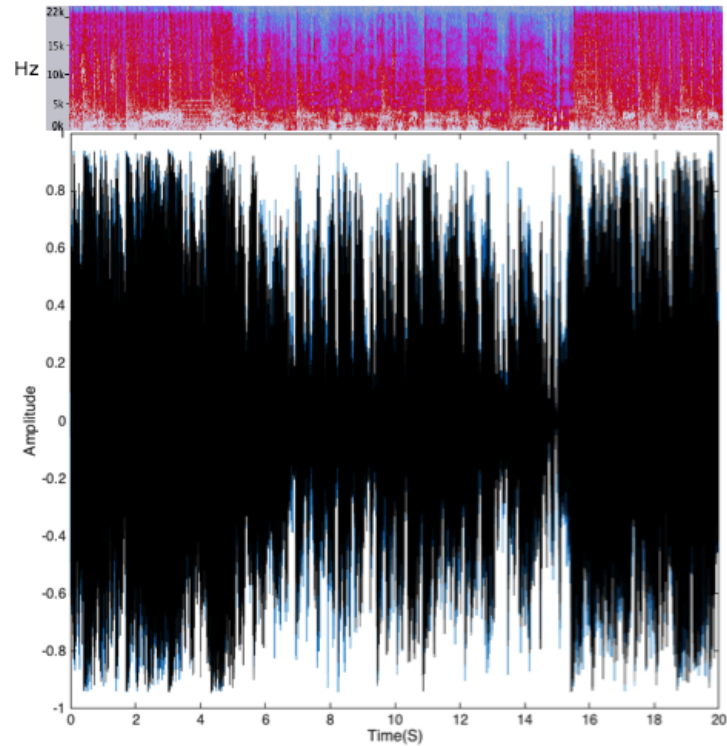
**Figure 18 - Sugababes Introduction Time Domain**

The Sugababes introduction (see Figure 18) commences with a sound effect from the video game 'Frogger'. At approximately 1.6 seconds the main synth hook from Tubeway Army's "Are 'Friends' Electric?" is introduced along with a low pass filtered drum track. This lasts for approximately 11 seconds and during this time the section is heavily processed with a form of sample rate reduction effect (Lo-Fi) and flanging. In addition, the main synth hook is panned between left and right channels. At approximately 11 seconds the main vocal introduces the first verse with no change to the music backing. Upon examination with a spectrogram, it was observed that up to the 1.6-second point there is reduced energy above 8kHz.



**Figure 19 - Sugababes Verse with Drums Time Domain**

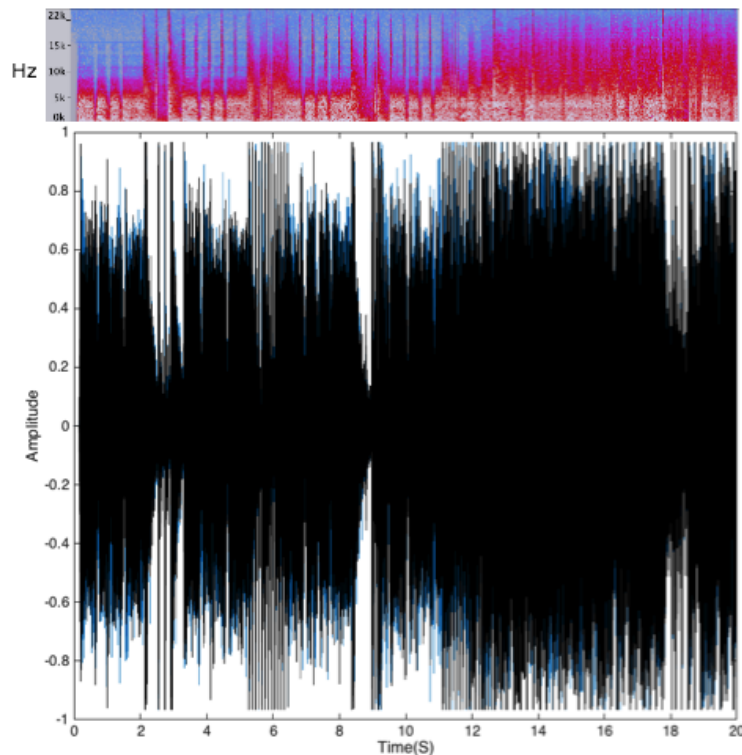
The Verse with Drums section of the Sugababes song (see Figure 19) begins with a small siren effect without drums and backing. The main synth, drums and bass then begin along with the main vocal verse. The verse continues until the 12-second point, at which time a heavy guitar riff is introduced, along with an additional lead synth and the chorus begins.



**Figure 20 - Sugababes Breakdown Time Domain**

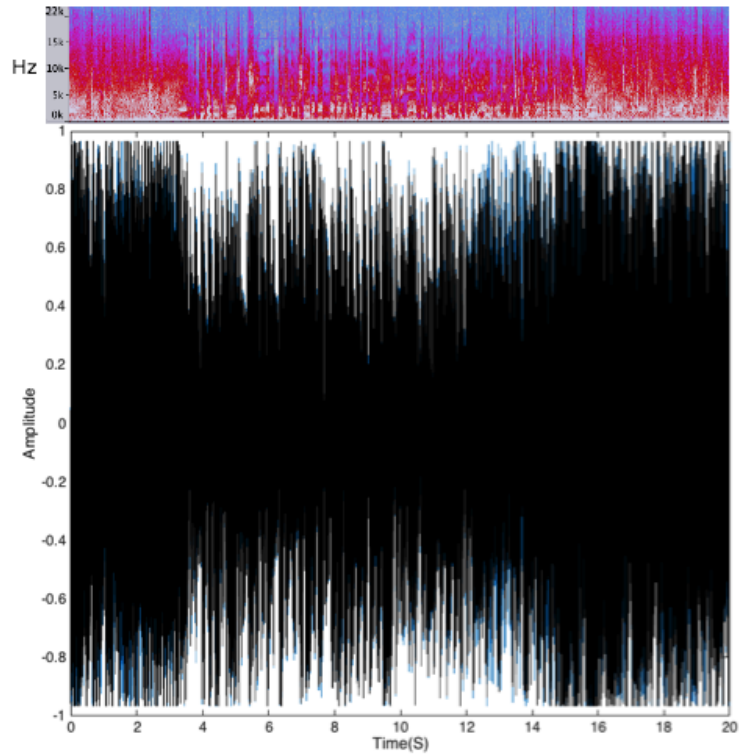
The Sugababes breakdown section (see Figure 20) begins with the end of the chorus before and at the 6 second point drops to the basic vocal and effected backing present in the intro section. This breakdown section continues before the song comes back in with a section identical in arrangement to that of the verse with drums first 12 second section. This occurs at the 16.5-second point, where there is a significant audio cut out.

### 5.3.3.2 *Nickelback excerpts*



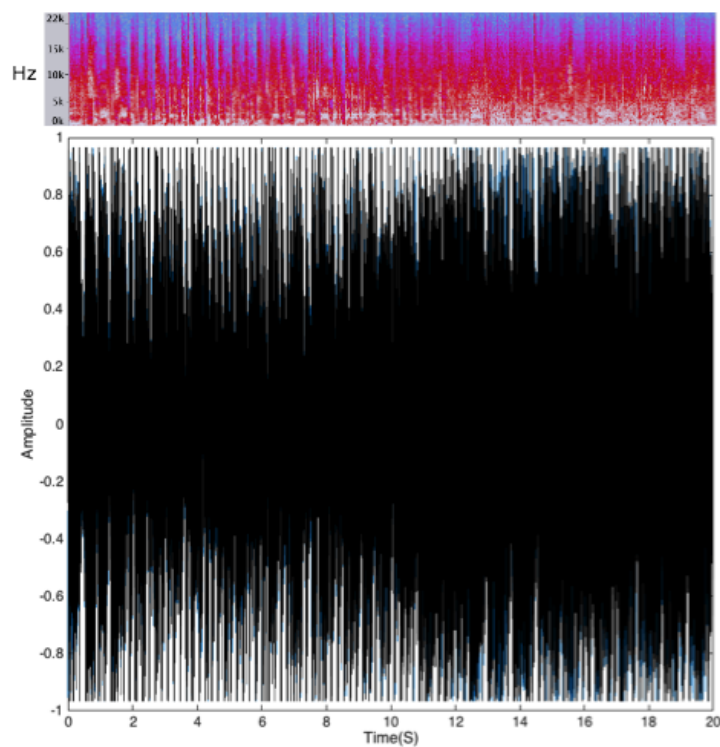
**Figure 21 - Nickelback Intro Time Domain**

The Nickelback Intro section (see Figure 21) opens with a heavy rhythm guitar four chord sequence, with strong low-mid frequency components. At the 2 second point a short drum fill occurs lasting until 3.2 seconds. The guitar riff continues including underlying hi-hat quarter note hits. At 8.8 seconds a significant tom fill occurs which includes a brief audio dropout. At 12 seconds a major drum fill occurs and the guitars, bass and drums play the main hook from the 14 second point. At 17.7 seconds a short drum fill / guitar drop out occurs before the main hook continues.



**Figure 22 - Nickelback Breakdown Time Domain**

The Nickelback Breakdown section (see Figure 22) opens with a heavy rhythm guitar four chord sequence, with strong low-mid frequency components. At the 3.5 second point the production drops the guitars out of the mix and features the vocal and hi-hats as a breakdown. A drum fill occurs at 15 seconds, followed by the full drums, bass, vocal and guitar mix at the 17 second point.



**Figure 23 - Nickelback Verse with Drums Time Domain**

Figure 23 shows the Nickelback Verse With Drums section of audio. This section of audio represents the section of the track that contains bass, drums and vocal up to the 16-second point, at which the guitars are re-introduced.

## 5.4 Results

The following charts compare the punch/clarity average scores with the 75% overlap IBR scores for each excerpt.

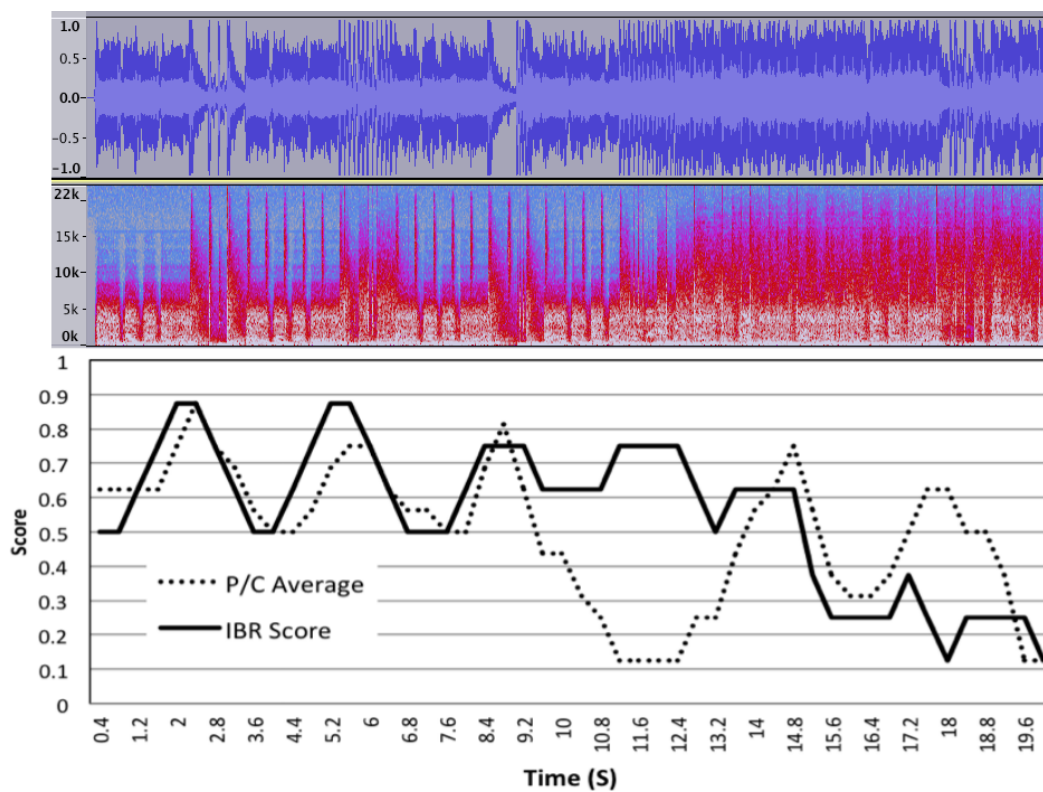


Figure 24 - Nickelback Intro - IBR vs. Subjective

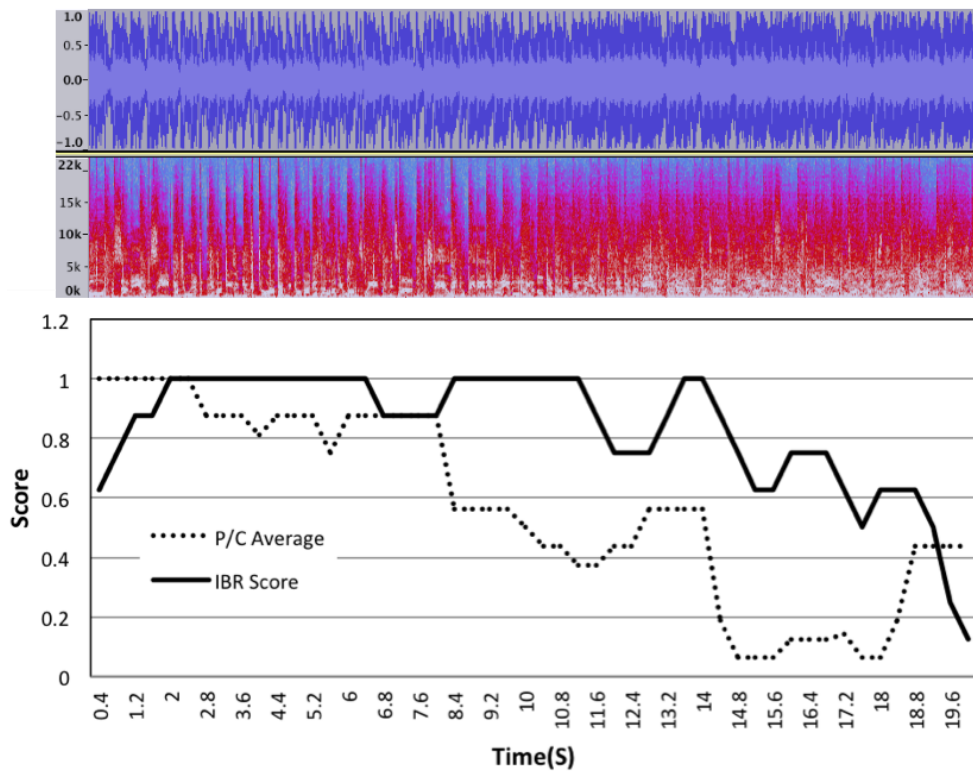


Figure 25 - Nickelback Verse with Drums - IBR vs. Subjective

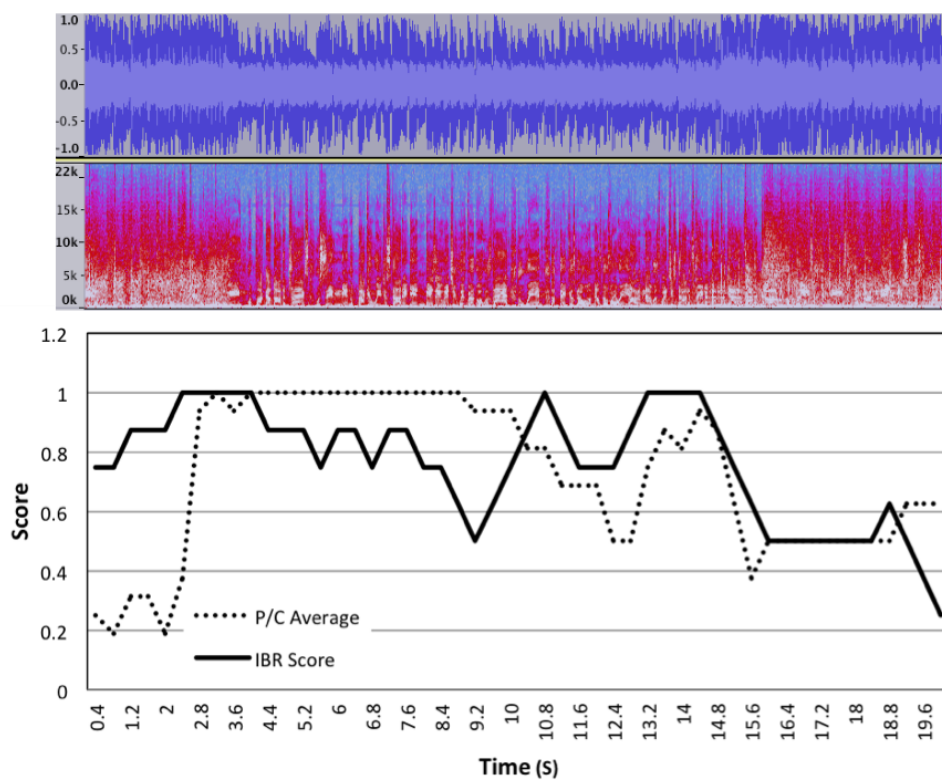


Figure 26 - Nickelback Breakdown - IBR vs. Subjective

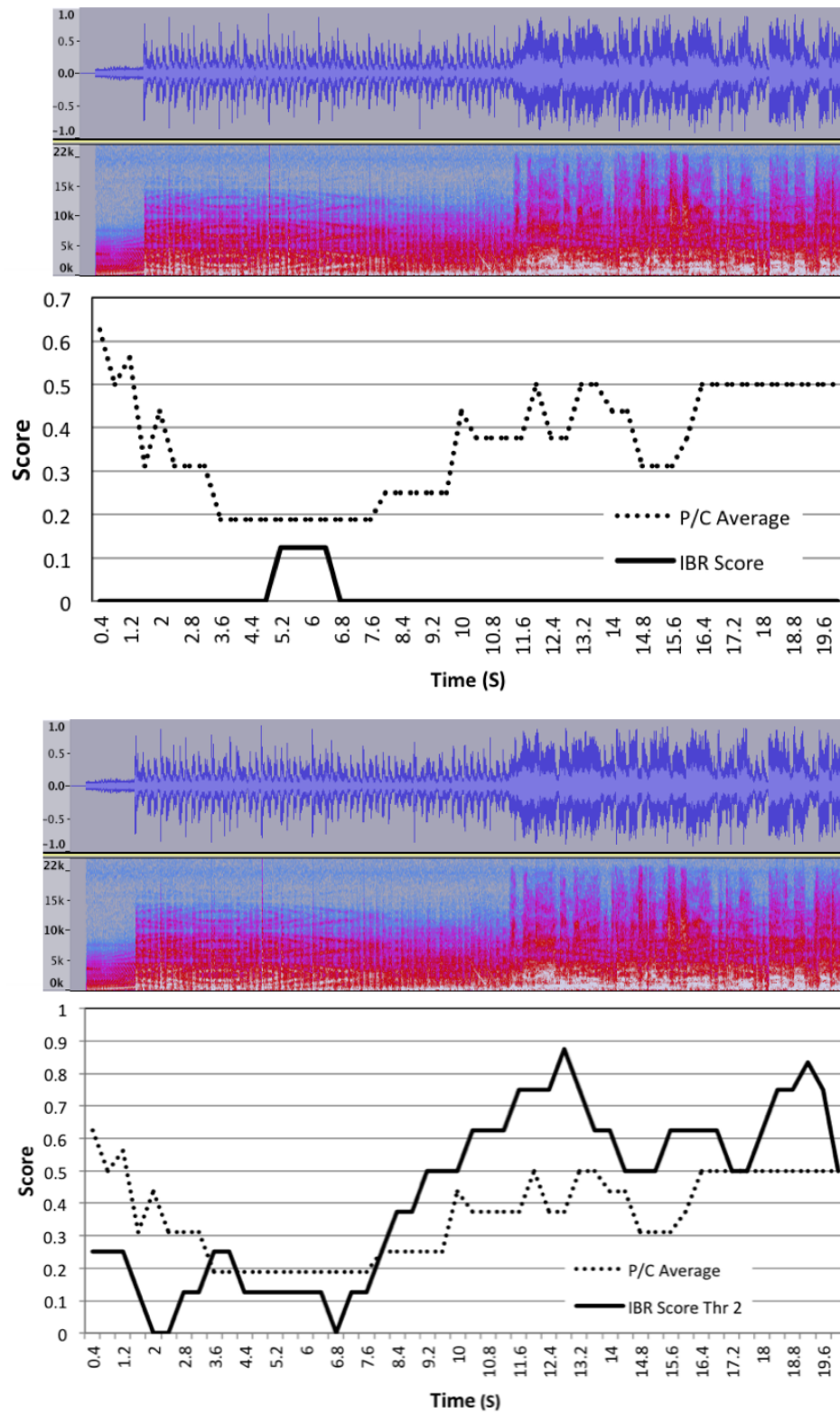


Figure 27 (a)/27(b) Sugababes Intro - IBR vs. Subjective

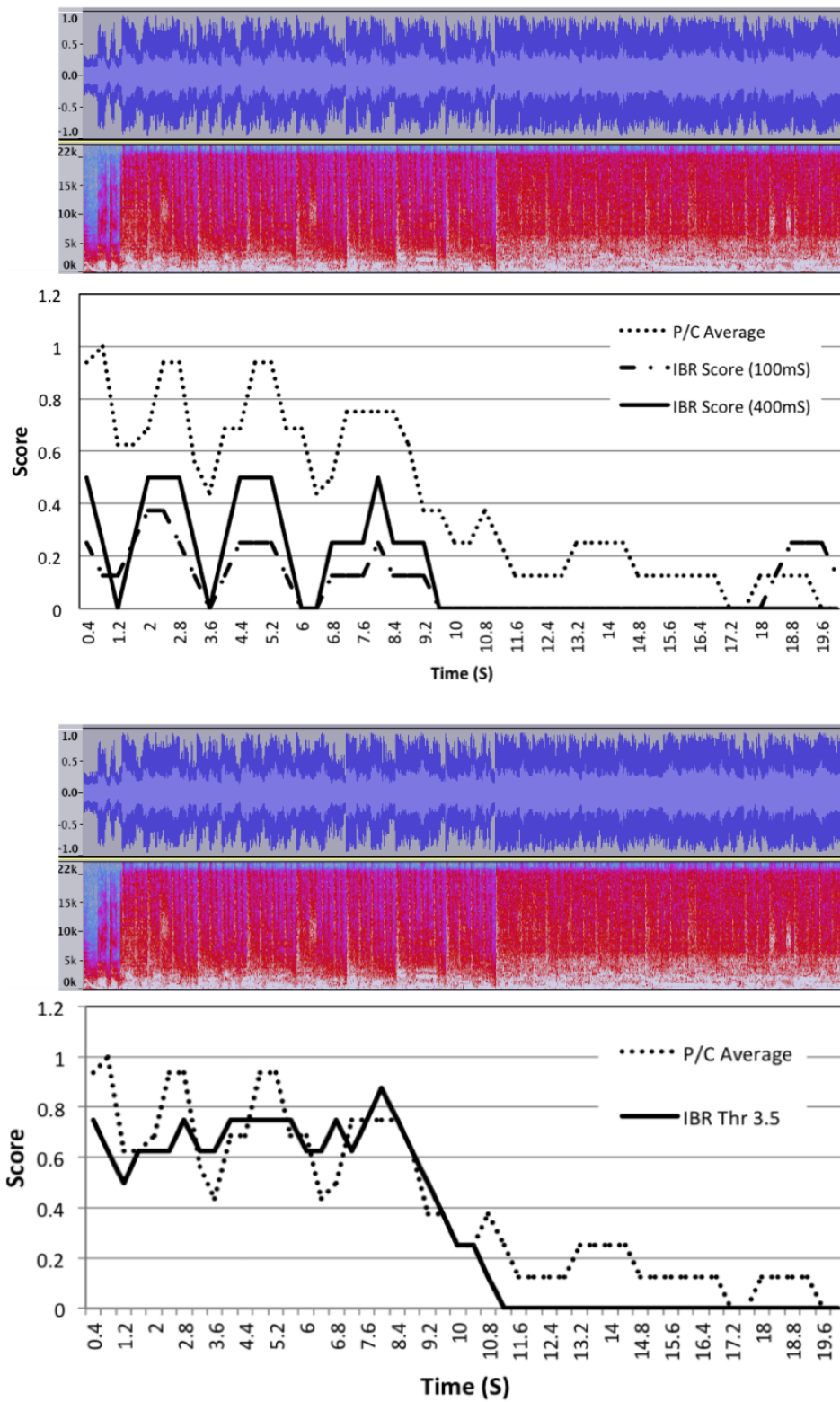


Figure 28(a)/28(b) Sugababes Verse with Drums - IBR vs. Subjective

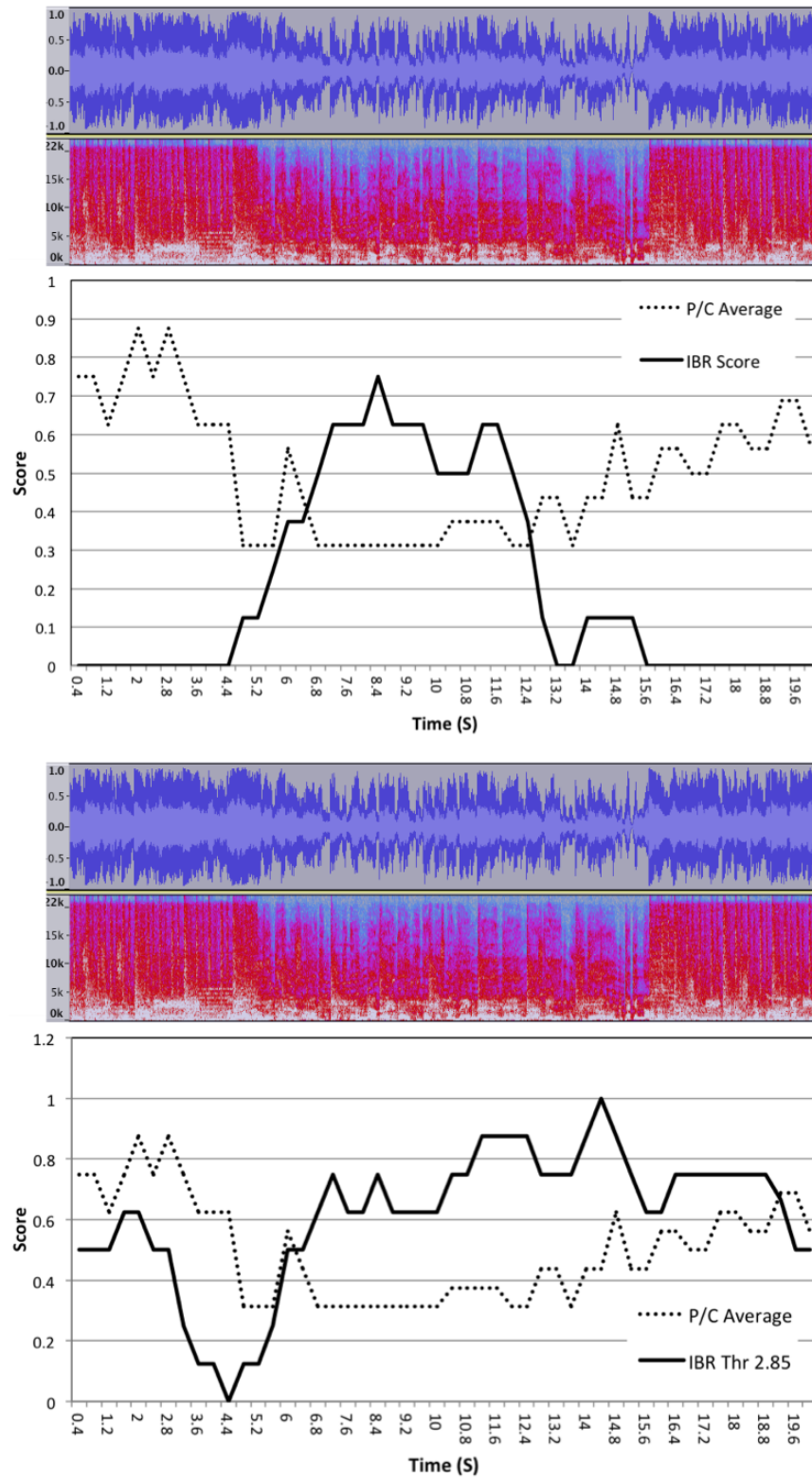


Figure 29(a)/29(b) - Sugababes Breakdown – IBR vs. Subjective(with and without threshold)

The following table presents the Pearson coefficient calculations for each of the excerpts.

Excerpt	Pearson Coefficient
Nickelback – Intro	0.330
Nickelback – Verse With Drums	0.485
Nickelback – Breakdown	0.334
Sugababes – Intro	Figure 28a 0.118 Figure 28b 0.354
Sugababes – Verse With Drums	Figure 29a 0.605 (400ms) Figure 29b 0.917
Sugababes – Breakdown	Figure 30a -0.684 Figure 30b -0.240

**Table 9 - Pearson Correlation per Excerpt**

## 5.5 Discussion of results

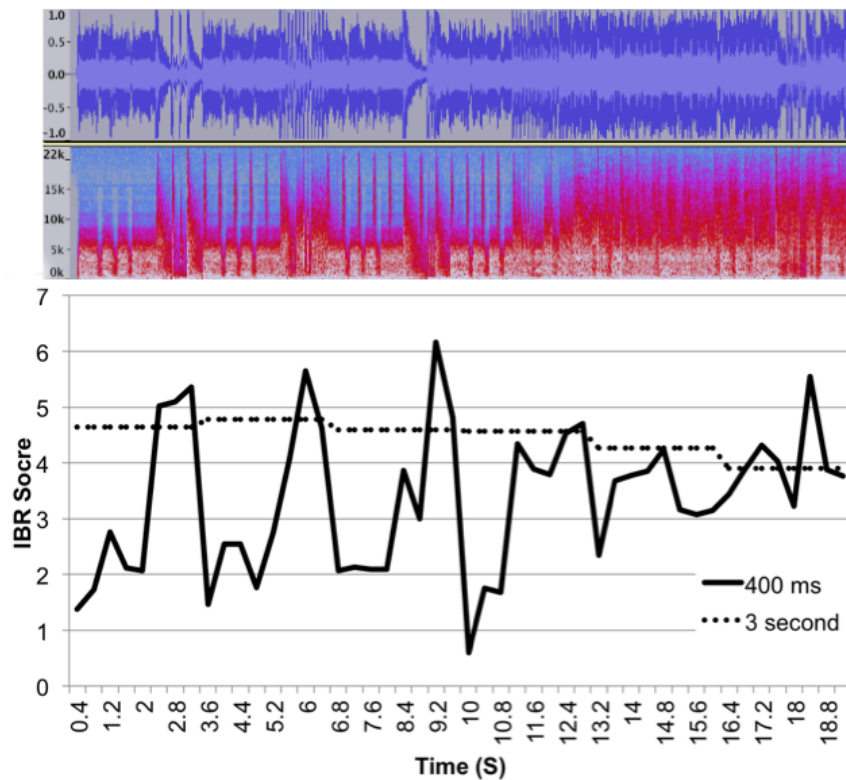
As can be seen from the Pearson tests in Table 9, there is some degree of correlation evident between the subjective and objective scores, these range from  $r=0.917$  to  $r=0.118$ . However, the correlation varies across the different sections of the audio under test., in some cases the correlation also depends on choice of IBR threshold chosen. What follows is a discussion and analysis based on the two excerpts which have very different musical structure and perceptual content, i.e. that the Nickelback Intro and Sugababes Breakdown excerpts.

Visual inspection of Figure 24 shows a high correlation between the IBR and subjective scores up to approximately 8.8 seconds at which point the two trends deviate. This error margin begins to decrease around the 13 second point of the audio.

During the initial 8.8 second period of this excerpt the elements that are prominent are those of the drums and guitars at different times in a call and response pattern. Major drum fills occur centered around the 2, 5.2 and 8.4 second mark. These fills relate to the points at which the listeners have graded the audio as punchy and clear. These points also correlate well with the objective IBR score. Further points where the error is minimised between the two measures are at the 14 second point, again a point at which a major drum fill occurs

Significant errors begin to occur at 8.8 seconds where there is a major tom fill and an audio drop out. This could explain the de-correlation in subjective and objective scores within this period and the overall weak Pearson coefficient of 0.330. The tom fill and audio drop out is seen by the algorithm as a highly transient event and therefore a high IBR score is calculated whilst the loss of audio could be considered by the listeners as lacking both punch and clarity. Measurement of the excerpt loudness at this point indicated a loudness range increase of approximately 14dB, this lasts for around 3 seconds which relates to the point at which the error margin begins to decrease. This period of de-correlation is caused by a period of low-level audio and this issue may be overcome by making use of a gate in the same way as used in ITU-R BS.1770-4 (2015) loudness model.

Whilst the IBR score indicates transient behaviour is occurring during the 8.8-13.2 second period, this is graded with a low score by the listeners. The IBR 400ms plot for this excerpt (see Figure 30) shows the 400ms IBR score falling below the threshold value of 4 during this time period. The calculated IBR compared to the subjective measures is based on a 75% overlap window and also incorporates the 3-second IBR score. This suggests that in order to increase the accuracy of the IBR score, one might consider a calculation based solely on the 400ms windowed IBR, a smaller window size or a weighting factor being applied to the smaller window size measure. This confirms that the window size used in the measure does impact on the resulting IBR score. A smaller window size being able to track finer time scale changes as one might expect.



**Figure 30 - Nickelback Intro IBR (400ms and 3s Window Sizes)**

With reference to Figure 28 (a), this technique was applied to the Sugababes Verse With Drums excerpt as can be seen and the trends begin to map more closely.

Overall, when comparing all six of the Nickelback and Sugababes excerpts, it was noted that despite the Nickelback excerpts having the lowest overall loudness ranges they possessed on average, higher instances of dynamic content, indicated by higher IBR scores than that of the Sugababes excerpt.

The Nickelback excerpts regularly exceed the threshold whilst the Sugababes excerpts don't. The associated loudness plot of the Nickelback Intro excerpt indicates that the audio measured is well in excess of the proposed -23 LUFS loudness level. Of interest, is that despite the Nickleback samples indicating a smaller average loudness variation, they still possess enough dynamic fluctuation to assert a high IBR score, which suggests frequency band de-correlation is present.

The Sugababes excerpts, whilst not having the reduced loudness range of the Nickelback tracks do not exhibit the same frequency band de-correlation, hence the lower average IBR scores obtained.

The nature of the subjective test was non-comparative i.e. the listeners were asked to grade each excerpt separately. Therefore, it's likely that the listeners were grading 'punch and clarity' by comparing points in time of the audio they were listening to for that particular grading phase. The original threshold grading of 4 was perhaps inappropriately chosen as it represents a comparative value that originally distinguished overall between good and bad quality excerpts in previous experiments.

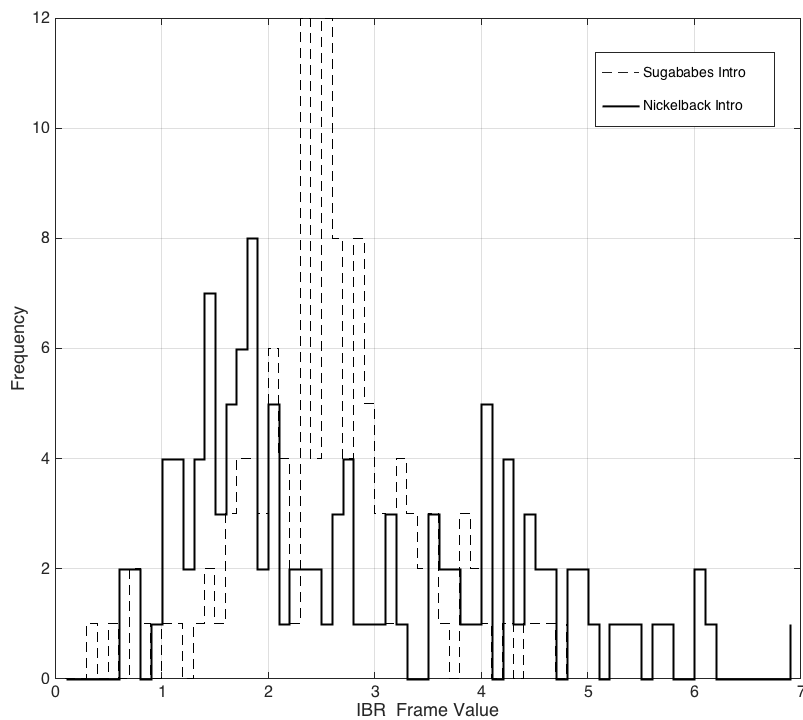
Table 9 shows that with a Pearson coefficient of -0.684, the Sugababes Breakdown excerpt was the excerpt with a strong negative correlation. With reference to Figure 29 (a), the IBR measure rises from zero and becomes closer to the subjective scores. This corresponds to the point at which the vocal line is prominent in the mix. This time period has a high IBR score both in the 3s and 400ms window time frame due to varying dynamics in the mix i.e. de-correlation in the low, mid and high frequency bands.

The listeners appear to grade the excerpt differently, giving higher scores to the sections of audio outside this region. This could be due to them gauging drums, bass and synth as punchy and a single vocal as neither punchy or clear. This could be the cause of the negative correlation of -0.684. A point to raise here is that whilst there is clearly dynamic content within the piece signified by the IBR scores, there is a drop in loudness level by approximately 13.5dB, which could also correspond to the listener perception of punch and clarity. This might suggest that a combination of both a loudness and dynamics measure could improve the accuracy of the punch and clarity score.

By varying the threshold used in the IBR calculation, it is possible to profile the excerpts with a greater *dynamic sensitivity*. Figure 27 (b), Figure 28 (b) & Figure 29 (b) show an IBR profile that has been calculated using threshold values of 2, 3.5 and 2.85 respectively. Looking at Table 9, the Pearson coefficients indicate that in most cases, a change in IBR threshold improves the trend correlation. The Sugababes - Verse With Drums excerpt shows an improvement from 0.605 to 0.917 which could be considered highly correlated. Having said that, as can be seen in Figure 28, as the IBR threshold is lowered resulting in an overall higher correlation score, the resulting plot can track, by visual inspection, the general trend of the P/C scores. In addition, although some correlation improvement is seen between Figure 29 (a) and Figure 29 (b) the correlation is still negative, albeit visually, the objective and subjective plots look more alike.

## 5.6 IBR statistical output

The IBR measure, along with suitable metering ballistics, may prove more useful in offering a real time indicator of the dynamic nature of the material, for example, to indicate elements of the material that is changing. A more useful application of the IBR measure could be to evaluate the measure statistically with respect to time, for example, in a histogram. The histogram can be analysed, giving statistical data relating to the stimuli under test allowing overall trends to be observed and compared. Trends such as the overall ‘correlation’, or mean IBR between bands may offer an insight into how the measure may correlate with attributes of the stimuli perceived by the listener.

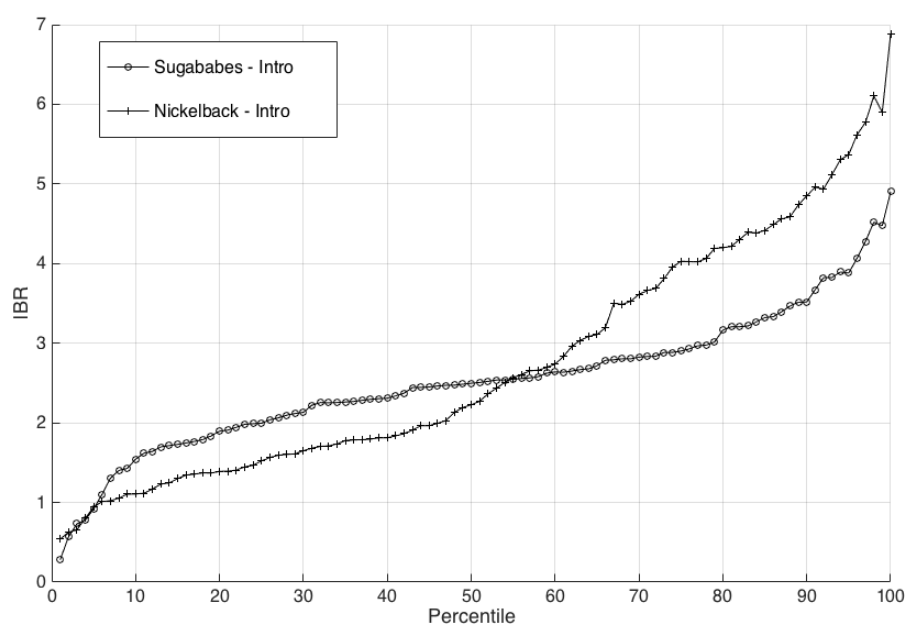


**Figure 31 - IBR Histogram - Nickelback Intro vs. Sugababes Intro**

The histogram shown in Figure 31 represents the data extracted from the Nickelback Intro and Sugababes Intro excerpts. It shows the magnitudes of a particular IBR frame, quantised into 0.1 intervals, and its frequency within the entire music sample. By examining the data in this way, maximum, minimum, median, standard and deviation measures can be extracted.

It can be seen in Figure 31 that the Sugababes excerpt (shown as the dotted line) has a high number of IBR frames around its IBR mean of 2.5233. In fact, the distribution of IBR frames for the Sugababes excerpt is also normally distributed about this mean (which was confirmed by a

Kolmogorov-Smirnov test). The Nickelback intro excerpt on the other hand, is somewhat bimodal in its frame distribution. Whilst it has an overall IBR mean of 2.7351, there are a larger number of frames below the IBR value of 2 if compared to the Sugababes excerpt. This suggests that overall the Nickelback Intro excerpt has a higher number of frames that exhibit a higher correlation of dynamic range between the frequency bands employed in the IBR measure. In addition to this, there is a greater spread of higher value IBR frames within it. The reasons for this could be the transient nature of the Nickelback Intro excerpt with its staccato guitar, drums and dropouts. These parts all contain strong transients therefore it's likely that there is strong correlation between bands if measured at their peak, especially when combined with hard limiting as you might find in this style of production. The spread of frame values could be as a result of the differing dynamics evident in the sample. Contrast that with the somewhat uncorrelated and distorted Sugababes Intro excerpt that doesn't have particularly strong transients within it.

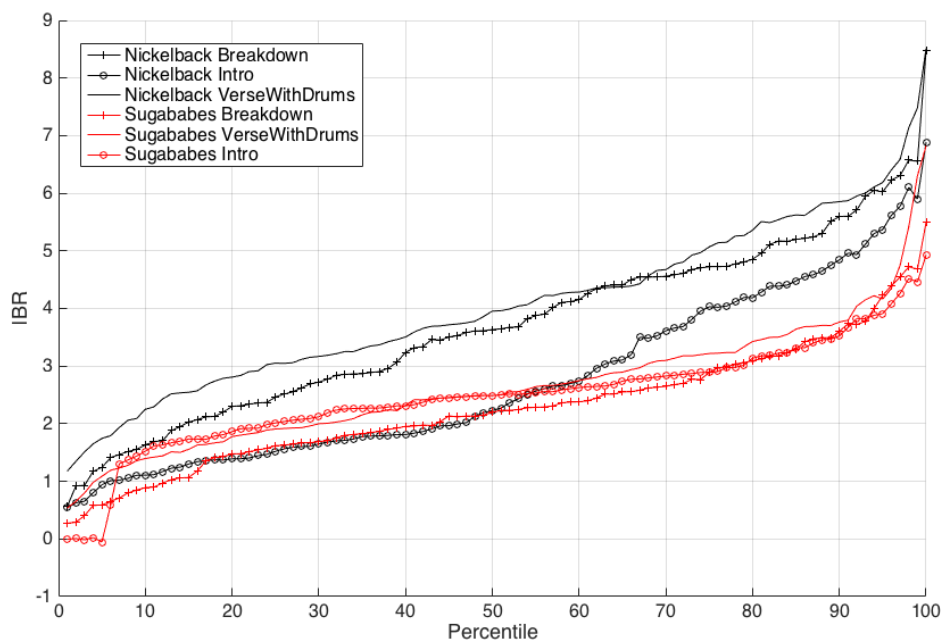


**Figure 32 - IBR Percentile Nickelback Intro vs. Sugababes Intro**

This data can also be shown in the form of a percentile plot (see Figure 32) which gives an effective insight into the underlying trend of the music with respect to the IBR metric. In Figure 32 clear differences in IBR distributions are evident. The Nickelback Intro, above the 55 percentile exhibits a higher number of larger IBR frame measures. These larger IBR frame measures could be as a result of the somewhat sparse arrangement employed in the sample, for example guitar intro followed by drum fill, followed by guitars and drums and as mentioned

previously, the perceptual dynamic this affords. The measures obtained below this point show a general trend of being less than that of the Sugababes Intro sample but only marginally. These lower IBR frames could be a result of the correlation between frequency bands due to the hard limiting or compression that could have taken place in the mastering process. In a similar way to the ‘Loudness Range’ measure (Skovenborg, 2012), upper and lower percentiles could be ignored therefore resulting in an IBR\_diff measure being extracted based on this data.

If for example the difference between the 1<sup>st</sup> and 95<sup>th</sup> percentile is calculated, Nickelback Intro, in this case, would exhibit the largest IBR\_diff with a figure of 4.8305 whilst the Sugababes Intro has a figure of 3.6027.



**Figure 33 - IBR Percentile Nickelback vs. Sugababes**

Figure 33 shown above details all 6 excerpts measured and their relative percentile distributions of IBR measures. Overall, it can be seen that all the Nickelback excerpts have a higher IBR\_diff than their Sugababe counterparts, measured between the 1st and 95th percentile. For Nickelback Intro, Breakdown and Verse With Drums the values are 4.8305, 5.4614 and 5.0134 respectively. This shows clear differentiation between the dynamic content (or contour) measured at the three different points in the song with the introduction being the most contrasting. The Sugababes excerpts all show a very similar trend with respect to dynamic contour with IBR\_diff values of

3.6027, 3.9546 and 3.6311 for the Intro, Breakdown and Verse With Drums respectively. This corresponds with the somewhat less dynamic differences evident throughout the Sugababes song if compared with the Nickelback song.

## **5.7 Conclusions**

This experiment presented the temporal IBR as a measure to quantify audio quality with respect to punch and clarity. The results indicate that despite a musical piece having a smaller loudness range, it is the transient content and dynamic range de-correlation between frequency bands that could relate to higher subjective scores being given by the listeners.

A degree of correlation was observed between subjective test scores and the objective IBR descriptor suggesting it could be used as an additional measure to describe punch and clarity with a piece of music. Limitations of the measure were identified which highlight that further consideration is required with regards to the choice of threshold adopted based on the range of dynamics detected within the musical extract and the possible inclusion of a gate as utilised in some loudness algorithms.

The IBR statistical output, both in terms of percentile and histogram representation, is an improvement on the integrative-based method proposed in previous experiments. With reference to Experiment 2 in Chapter 4, the statistical IBR output affords more insight into the underlying dynamic contour of the sample under test than an overall dynamic range based measure. It also shows a clearer differentiation between the best and worst subjectively scoring excerpts used in this experiment.

## **Chapter 6 Elicitation and grading of punch in music**

The experiment conducted in Chapter 5 investigated the relevance of dynamic range and associated correlation of this measure between frequency bands, with the listeners being asked to grade excerpts in terms of perceived punch and clarity. It was concluded that the transient content and dynamic range de-correlation between frequency bands could relate to higher subjective scores by listeners when judging punch and clarity.

Whilst it's apparent from this study that listener perception of punch and clarity is related to the presence of dynamics, it's important to establish which components of the audio signal contribute greatest to this and establish if other low-level parameters play a role. Therefore, a reverse elicitation method was utilised.

The experiment is outlined in detail in the following sections.

### **6.1 Method of testing**

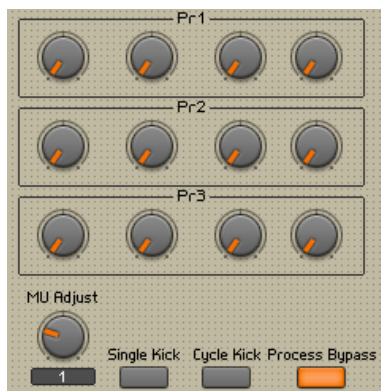
As discussed in Section 2.5 the term punch is a subjective term, which is often used to characterise music or sound sources that exhibit a sense of dynamic power or weight to the listener. A reverse elicitation experiment was conducted in order to establish low-level characteristics of a signal deemed as 'punchy' by expert listeners. In this experiment, expert listeners were asked to create audio samples that they perceived as having punch using a multi-band wave shaping process. They then graded the generated punchy audio samples in a controlled listening test.

The samples created were then analysed with respect to various low-level features such as Spectral Centroid, Log Attack Time, Signal Intensity, Intensity Ratio, Rhythm Strength, Spectral Flux and Spectral Spread.

### 6.1.1 Elicitation exercise

Twelve expert listeners took part in the initial elicitation exercise where they were asked to create audio samples by modifying a sound source using a multi-band wave shaping interface.

A synthesised kick drum was chosen as the sound source for two key reasons. Firstly, the spectral and temporal components of the synthesised sound source could be carefully controlled, also the resulting kick sample did not suffer from room coloration. Secondly, due to the transient nature of a kick drum and its frequency range, it is often an instrument that's used to add 'weight' or 'punchiness' to music production and it contains a strong transient component that can be measured.

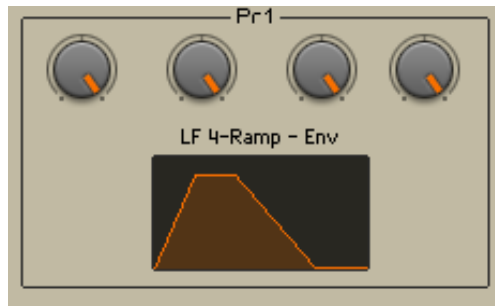


**Figure 34 - Test Interface Wave Shaper**

The kick drum source was synthesised using a T Bridge oscillator type model found in the TR-909 drum machine. It was then fed through a 3-band linear phase filter, with respective cut-off frequencies and Q settings chosen to be the same as in the experiment described in Chapter 3 (see Table 4). Each sub-band was then fed into a temporal shaper, the interface of which is shown in Figure 34 with the Pr1, Pr2 and Pr3 sections having 4 control knobs for each shaper.

The test interface, shown in Figure 34 was intentionally left unlabelled and was merely a collection of control knobs in a random arrangement. Despite the knobs being in groups of 4, their functionality was also assigned randomly, therefore any pre-conceptions of typical audio wave shaping controls or production preference biasing effects were avoided. The listeners were asked to modify the sound source until they felt the audio exhibited an increased sensation of punch. They could continue modifying controls as long as they wanted until they thought they had achieved maximum punch. The exercise took place in a soundproofed control room using

headphones to eliminate room colouration and speaker influences. Playback levels were set to 76dB(A). This playback level was aligned utilising the measurement of the RMS signal level and the headphone sensitivity of 112dB/1V<sub>RMS</sub>.



**Figure 35 - Example Waveshape Setting**

Figure 35 shows the relationship of the knobs to the waveshaping settings in an un-randomised order. From left to right, the knobs controlled attack time, peak level, peak-hold time and finally release time respectively. This detail was not shown to the listeners.

During the wave shaping process, the experts were asked to maintain the loudness levels between the processed and unprocessed signals at all times. This was achieved by the use of Make-Up gain control marked MU adjust on the interface. The level was monitored using a NuGen Audio Loudness Meter (VisLM, 2012) and was set to a level of -32 LUFS.

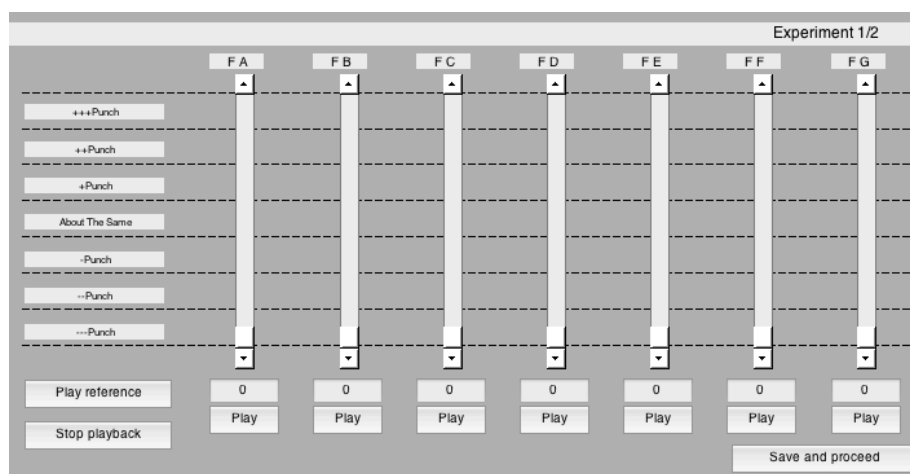
In order to achieve a compression like response, the reciprocal of the wave shaper output was used to shape the respective sub-band. The rationale behind wave shaping by envelope rather than modelling of a specific audio compressor was to reduce the number of experimental variables and prevent ‘equipment signatures’ being considered during the process by the listeners.

Each listener was asked to process two separate instances of the sound source. The difference between the two was the inclusion of an instantaneous attack in the first source. The reason for this was to investigate the effect of batter head or beater change. In total 24 samples were created, 1 by every listener for each of the two sound sources. Each listener confirmed that when they referenced the processed sample with the original sound source, additional punch was perceived. All samples, including the sources were 44.1kHz, 16 bit, mono WAV format.

### 6.1.2 Subjective testing

Eleven expert listeners took part in a controlled subjective listening test. They were asked to grade the ‘punchiness’ of the audio samples created during the stimuli elicitation exercise. The listening test took place in a sound proof control room using headphones to eliminate room coloration and the playback level was fixed at 76dB(A). This playback level was aligned utilising the measurement of the RMS signal level and the headphone sensitivity of 112dB/1V<sub>RMS</sub>.

A modified MUSHRA test formed the basis of the listening test. The test being modified to allow the listeners to rate the samples as either less punchy, more or the same as the hidden reference, in this case the unprocessed sample. The scale was continuous and ranged from 0 to 140, with samples rated the same as the reference being scored as 70. A hidden anchor was utilised which was a 3.5kHz hi-pass filtered version of the reference. A section of the modified MUSHRA interface is shown in Figure 36, the full interface consisted of all 14 samples visible across the screen, consisting of the 12 listener created samples, hidden reference and anchor. The experiment was undertaken based on source 1 and then repeated with source 2.



**Figure 36 - Modified MUSHRA Interface**

Following the analysis detailed in ITU-R BS.1534-1 (2003), the individual scores were collated and a Mean Punch Score (MPS) profile for each sample was obtained along with 95% confidence intervals, for both experiments. In addition, the listeners were asked to describe what

they perceived the punch attribute to be, and these were collected as a set of verbal punch descriptors.

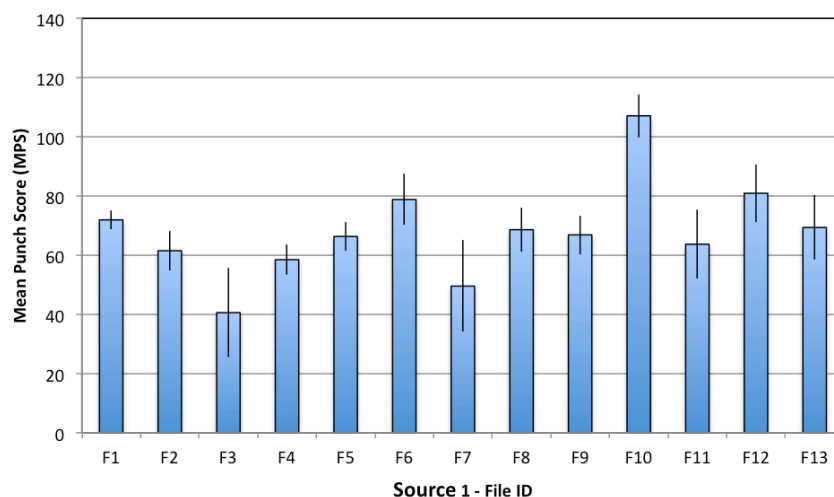
### 6.1.3 Objective measurement

A number of extracted parameters were analysed using the best and worst samples based on the normalised MPS achieved. The choice of parameter was guided by both the interpretation of the verbal descriptors given by the listeners and a choice of low-level audio descriptors described in the MPEG7 standard (MPEG 7). The signals were analysed using a combination of MATLAB scripts using an 1024-point STFT with a variable step size and Sonic Visualiser (Cannam et al., 2010).

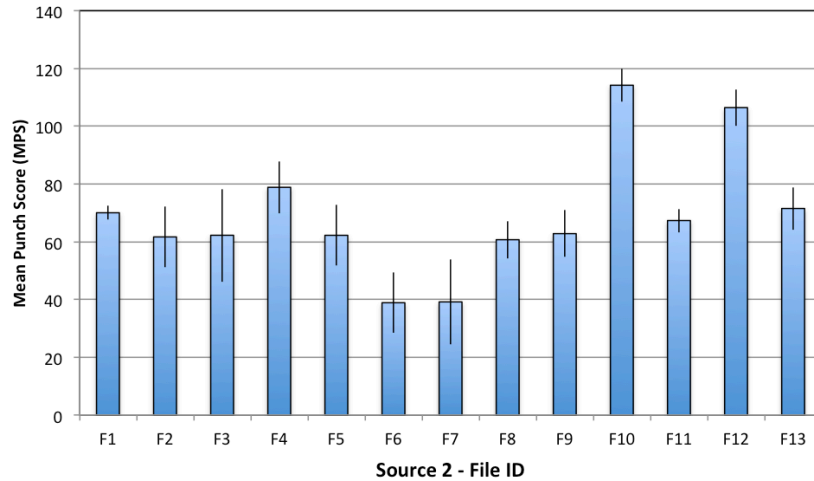
Parameters measured were Spectral Centroid, Log Attack Time, Signal Intensity, Intensity Ratio, Rhythm Strength, Spectral Flux and Spectral Spread. A description of these measures can be found in Subsection 2.6.1

## 6.2 Subjective listening test results

Figure 37 and Figure 38 show the normalised MPS along with 95% confidence intervals. The x-axis shows each wave-shaped file. File 1 (F1) is the unprocessed reference. Three listeners failed to identify the reference and therefore their results were not utilised. File 14 (F14) was the hidden reference and was identified by all listeners with a grading of 0, this file is omitted on the graphs.



**Figure 37 - Source 1 (Instantaneous Attack) - MPS vs. File**



**Figure 38 - Source 2 - MPS vs. File**

### **6.2.1 Verbal punch descriptors**

Each listener was asked to describe the sensation of punch and what they were making their choices based on. The following is a list of the descriptors collected.

“Thud, Weight, Fast Attack, Thump, Gated Feel, Energy Burst, Hard, Dense, Focussed, Tight, Narrow, Defined”

### **6.2.2 Statistical analysis**

A Repeated Measure ANOVA was performed on the subjective data set. The results showed that the samples had a significant effect on the results ( $p < 0.01$ ,  $F = 26.703$ ). The source itself was found to be insignificant ( $p = 0.676$ ,  $F = 0.190$ ).

Multiple linear regression analysis of the high-level control settings chosen by the expert listeners was carried out, choosing the Mean Punch Score as a dependent variable and the three band control settings as independent variables. The reason for this analysis was to establish the key temporal parameters that could have the majority of the effect on the perceived punch attribute. Initial analysis showed that the make-up gain parameter had a large and incremental effect on the punch score albeit with a large standard error based on its model coefficient.

The make-up gain was utilised to re-normalise loudness levels after the other parameters had been modified. It was therefore deemed unnecessary to include it in the regression analysis in order to obtain correlation coefficients for the parameters of interest.

Firstly, the Pearson correlation coefficient for each parameter was calculated using SPSS considering each parameter independently. The output of this analysis is shown in Table 10. In addition, the relative p-value is shown for each which gives an indication of statistical significance rather than correlation by chance alone.

	Source 1 – Instant Attack		Source 2	
Parameter	Pearson	Sig(1-tailed)	Pearson	Sig(1-tailed)
LF Attack	-.339	.045	-.439	.067
LF Release	.360	.035	.509	.038
LF Peak Level	.307	.063	.238	.217
LF Peak Hold	.071	.364	.140	.324
MF Attack	-.414	.018	-.401	.087
MF Release	.245	.113	.325	.139
MF Peak Level	-.048	.407	-.197	.259
MF Peak Hold	.032	.439	.333	.133
HF Attack	-.287	.077	-.448	.062
HF Release	.324	.053	.239	.216
HF Peak Level	-.559	.001	-.663	.010
HF Peak Hold	.027	.447	.160	.300

**Table 10 – Pearson Correlation of Wave-shaper Parameters to Punch Score**

Dark grey highlights p-value significance level of 0.05, whilst light grey indicates a relaxed p-value of 0.10. Considering the correlation levels specified by Evans (1996) along with 0.05 significance level the LF Release was shown to have a weak positive correlation  $r=.360$ ,  $p<.05$  (1-tailed) for source 1 and a moderate correlation  $r=.509$ ,  $p<.05$  (1-tailed) for source 2 respectively. If taken in isolation, the respective  $r^2$  parameter could therefore account for approximately 13% and 26% of the variation in punch score for source 1 and 2 respectively.

In the case of source 1, LF Attack was significant but again with only a weak negative correlation  $r=-.339$ ,  $P<0.05$ . Looking at the same parameter for source 2, there is a stronger correlation however at this sample size it is not deemed to be significant. MF Attack shows a moderate negative correlation with a high significance level  $r=-.414$ ,  $P<.018$  (1-tailed) for source 1.

The HF Peak Level parameter showed a moderate to strong correlation across samples 1 and 2 cases and both could be deemed significant. As the HF Peak Level is reduced an increase in punch is observed. In the wave shaper model, this is valid as the actual level reduction is the reciprocal of this value.

If the significance level is relaxed to 10%, the attack parameter across all three bands can be seen to have a correlation to the punch attribute; the significant parameters are shown in light grey. The negative relationship between the attack parameters and the punch attribute indicates that an increase in punch is observed when attack times are reduced. Conversely, a decrease in release times results in a decrease in perceived punch.

The Pearson coefficients begin to indicate possible relationships between the perceived punch attribute and the underlying audio with respect to the wave shaping that has been employed. However, the coefficients shown are being considered independently and it's clear that no single parameter has an overriding impact or indeed shows very strong correlation to the perceived punch attribute. Multiple linear regression was conducted and analysis results were dismissed due to a high degree of multi-collinearity (correlation between the parameters). This was indicated by the variance inflation factors of the parameters (VIF) having values  $> 1$ . A cause of this could be that two or more parameters have an equal effect on the variance of the punch score and/or some listeners may have been modifying a combination of parameters to achieve a level of punch equal to that if they had chosen a single parameter.

### 6.3 Objective measurement results

The subjective experimental results were examined and through post statistical analysis of the data best and worst samples were identified with respect to the punch scores obtained. The naming conventions used in the objective feature measurement results relate to the subjective test results files as follows.

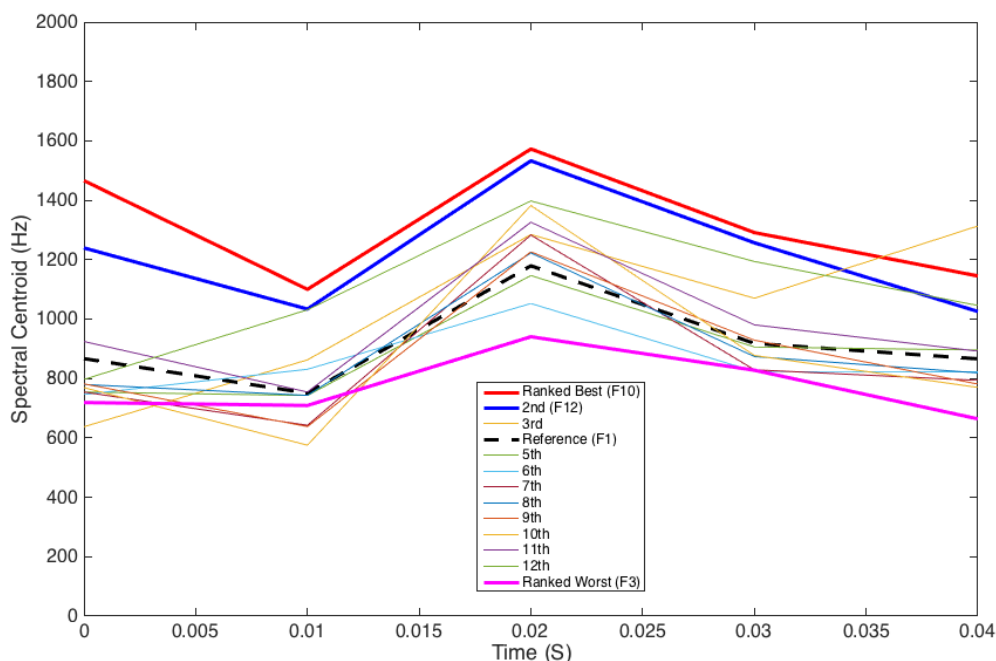
Referring to Figure 37:

- *F1: Reference 1*
- *F10: Source 1 – Best*
- *F3: Source 1 – Worst*

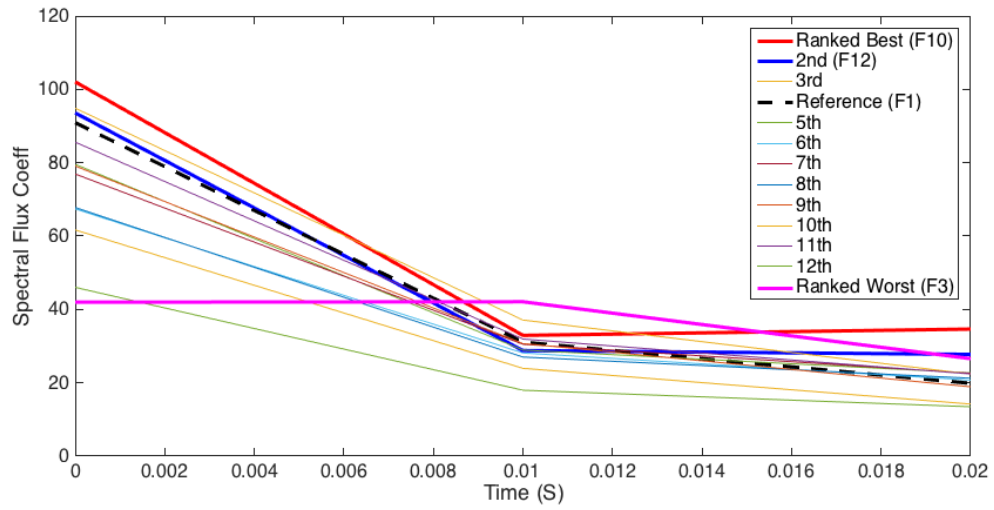
Referring to Figure 38:

- *F1: Reference 2*
- *F10: Source 2 – Best*
- *F6: Source 2 - Worst*

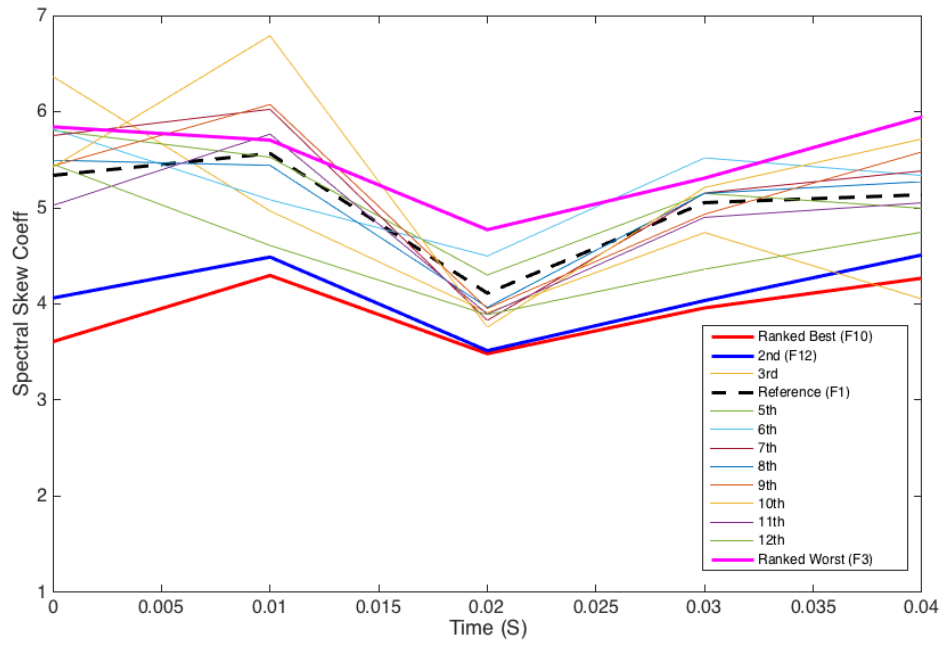
The following plots show the temporal values extracted for source 1. Values extracted for source 2 showed similar trends. To aid in the interpretation of these plots the best, 2<sup>nd</sup> place and worst are plotted as red, blue and magenta respectively and the reference is shown as the dotted black line. In addition, only the initial 40ms or 20ms of the analysis is shown. This is due to the fact that after this point in time the signal to noise ratio became too low to warrant any useful measure.



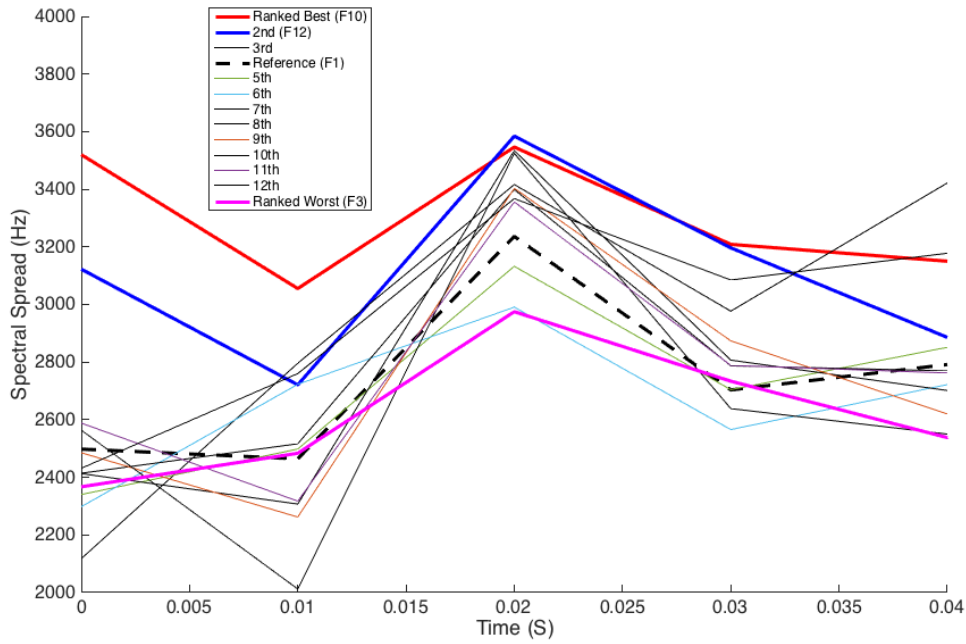
**Figure 39 –Spectral Centroid of Source 1 Kick Drums vs. Time**



**Figure 40 –Spectral Flux of Source 1 Kick Drums vs. Time**



**Figure 41 –Spectral Skewness of Source 1 Kick Drums vs. Time**



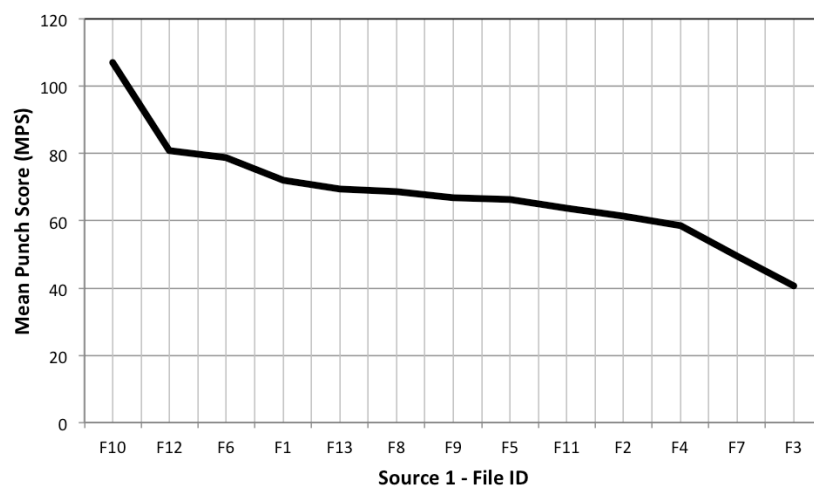
**Figure 42 –Spectral Spread of Source 1 Kick Drums vs. Time**

#### 6.4 Discussion of results

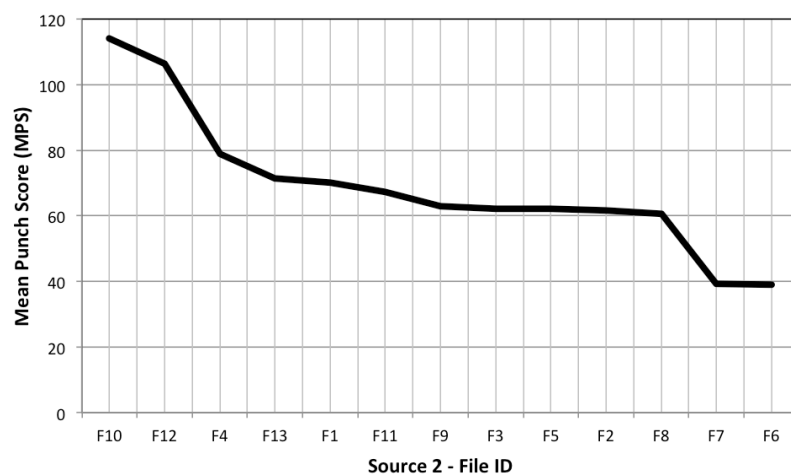
With reference to Subsection 6.2.2 detailing the statistical analysis made on the wave shaping parameters, it's evident that there are some parameters that show a degree of significance with respect to the subjective punch scores, namely the LF Attack and Release times, the MF and HF Attack times and the HF Peak Level parameter. However, as previously discussed, the majority of parameters exhibit correlation with one another with respect to modelling of the punch score. As such, if a measure or prediction of the punch score is to be made based on the effects of the parameters highlighted, it would be beneficial to elicit the perceptual change based on the parameter in isolation.

Figure 37 and Figure 38 show the mean punch scores for each sample for the two sources. With reference to these scores (for both sources) the highest score was obtained by F10 with an MPS of 107.07 and 114.2 respectively.

Due to the ranking nature of the MUSHRA test, it is possible to rearrange the data as shown in Figure 43 and Figure 44. The samples are shown in rank order from left to right with the highest first for each source tested, e.g. F10 has the highest rank score in each plot whilst F3 and F6 have the lowest in each respective plot.



**Figure 43 - Source 1 - Rank Scored**



**Figure 44 - Source 2 - Rank Scored**

Sample	Spectral Centroid (Hz)
Source 1 – Best	1263.11
Source 2 – Best	1242.91
Source 2 – Worst	1089.4
Reference 1	809.54
Reference 2	726.79
Source 1 – Worst	575.14

**Table 11 - Spectral Centroid (1024-point FFT)**

Table 11 shows the Spectral Centroid for the reference sources and best and worst samples for each experiment. One can observe that the highest MPS ranking sample achieved a spectral centroid value of 1263.11 Hz, contrasting with 575.14 Hz in the worst case sample. With reference to a typical percussive instrument timbre (MPEG 7), a spectral centroid of approximately 1217.34 Hz would be expected. When comparing MPS for each sample, the top two samples in each experiment have centroid measures around this figure.

The spectral centroid measures in Table 11 were obtained by analysing the full temporal response of the kick samples and averaging the data. An additional analysis was run to measure the spectral centroid at the point of maximum intensity within the onset period. A value of 41.9 Hz was obtained and was typical amongst all samples. This suggests that the majority of onset power is centered at this frequency, which is closely related to the 47Hz tuning of the kick sample. As the samples were produced using a simple T-Bridge arrangement, which is fundamentally a modulated sine wave without any complex modelling of the membrane, beater, drum shell or dampening, the spectral centroid measure outlined above would be expected. Further testing to establish variation in punch perception upon modification of the centroid at the point of maximum onset could be beneficial. If the spectrum of the onset was made more complex, one might expect the maximum onset centroid to change.

Figure 39 shows the spectral centroid measurements made on source 1 reference and samples with respect to time. The best scoring sample is shown to have a higher centroid value than all of the other samples throughout the sample timeframe except for falling below the centroid for the 3<sup>rd</sup> ranking sample, in the final 5ms. The lowest scoring sample shows a general trend of having a lower centroid value throughout its timeframe. Looking at Figure 43, which shows the extracted rank scores, there is clear differential between the best, worst and 2<sup>nd</sup> place samples.

This ranking is also evident in Figure 39 between these samples taking the centroid as a metric. Inspection of both Figure 37 and Figure 43 shows the remainder of the samples have a high degree of overlap in their respective MPS values. Likewise, the temporal spectral centroid measures obtained for these samples show a clustering of centroid value. This suggests, in these cases, that the centroid measure shows some correlation with the mean punch scores obtained.

Subband (Hz)	Ratio
1 (0-344)	0.643
2 (345-689)	0.067
3 (690 – 1378)	0.071
4 (1379-2756)	0.067
5 (2757-5512)	0.055
6 (5513 – 11025)	0.046
7 (11026 – 22050)	0.042

**Table 12 - Intensity Ratio of Source 1 reference**

Subband (Hz)	Ratio
1 (0-947)	0.736
2 (948-3186)	0.1206
3 (3187– 22050)	0.1434

**Table 13 - Intensity Ratio of Source 1 reference Using Wave-shaper Bands**

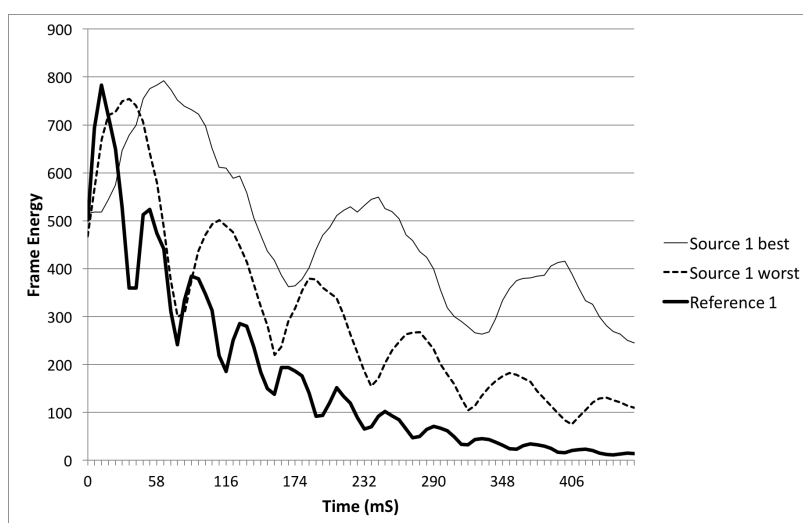
Intensity ratio is an indication of which frequency bands constitute the majority of a signal's power. From the Intensity Ratio measures taken, the first sub-band had the highest value for all of the samples. Typical values measured exhibited a pattern similar to the experiment 1 reference sample shown in Table 13. As the samples were synthesised with a fundamental tuning of 47Hz, this pattern might be expected. Given the majority of power exists within the low band of the samples under test, it could be possible that this band has a higher weighting contribution to the punch scores given by the listeners. When the signal energy ratios were examined using the bandwidths used in the wave shaper, typically there was equal mid and high frequency energy present in all of the samples tested. Table 13 shows how the energy is distributed for the reference in experiment 1 within the bands allocated for the wave shaper.

Sample	Rhythm Strength
Source 2 – Best	5200
Source 1 – Best	4970
Source 1 – Worst	4800
Reference 1	4790
Reference 2	4670
Source 2 – Worst	4230

**Table 14 - Rhythm Strength**

Table 14 shows the rhythm strength measured for the references and best and worst sample for each source. Both the best MPS samples measurements were greater than both the reference samples and worst MPS rated samples. Rhythm strength is effectively the sum of the magnitudes of the lowest sub-band power spectrum of the signal, a more detailed overview of this measure can be found in Subsection 2.6.1. This could be intuitively linked to the rating of the perceived punch of the sample however, the spectral centroid measures suggest that features other than those associated with the lowest sub-band dictate the level of punch perceived.

The log attack times measured for the source 1 best, worst and reference samples are -0.856, -1.092 and -1.468. This measure is based on the logarithm of the detected onset time in seconds. The intensity over time for these three samples are shown in Figure 45. The smaller the log attack time corresponded with the shorter onsets of intensity shown.



**Figure 45 -Signal Intensity vs. Time**

The best scoring sample has a longer onset but larger rhythm strength than the other two. Hence, it has more spectral energy in the onset portion of the sample. A shorter onset time is often associated with a punchy signal however the results suggest that the overall intensity contained within the onset shows a stronger relationship.

If one examines Figure 45 it can also be seen that the best scoring sample has a much larger intensity throughout the timeframe observed. This results in a reduced crest factor as the RMS level of the overall signal is larger. Crest factors for the best, worst and reference in this plot are 8.358, 10.559 and 12.428 respectively. With reference to the MPS obtained, the reference scored better than the worst case sample with source 1, thus suggesting that crest factor alone may not correlate well with punch perception.

Figure 40, showing the spectral flux measure over time for source 1, indicates that there is more flux in the best scoring sample than the worst during the initial 9ms. However, there isn't a significant correlation between the flux measure of the other samples and the MPS rankings obtained. This could be due to the MPS rankings having a high degree of confidence interval overlap within the best and worst sample boundaries. Inspection of the plot does show that the spectral flux is effective in signifying the periods during which the audio signal is evolving but in the worst sample, this is less evident.

Looking at the temporal measures of spectral skewness, Figure 41, there is a clear differential between the best and worst ranking punch samples. The best-ranked sample shows a lower skewness value than the worst ranked sample in addition to it possessing a more uniform horizontal trajectory with respect to time. This lower skewness values indicate biasing towards the low frequency ranges of the magnitude spectrum.

Figure 42 shows the spectral spread measure with respect to time. Again, there is a clear differential between the best and worst ranking samples. Spectral spread indicates how distributed the spectrum is about its centroid value. The best scoring sample has a highest spectral spread measure up to the 18ms point, this suggests that the onset in this case has a wider distribution of frequencies about the centroid. Typical of percussive sounds, the onset portion of the signal is noise-like, as such it has a wide distribution of signals within it and (taken in

isolation) and is less tonal. Comparing the other samples to the best scoring, their onsets could be considered as being less noise like, perceptually being voted as having less punchy.

The spectral centroid measures shown in Table 11 indicate significant variation between the samples tested. Given the initial sample creation exercise involved only temporal wave shaping and no direct modification of frequency spectrum (i.e. use of equalisation) took place, the resulting change in centroid (and other spectral attributes measured) was a direct result of the envelopes used. This is expected, as the temporal modification results in harmonic distortion and therefore additional frequency components appearing in the frequency domain.

## **6.5 Conclusions**

The experiment suggests that there is a possible correlation between the perceived punch attribute and the measure of rhythm strength, however, the measure is deemed not resolute enough with respect to the frequency bands it employs. The crest factor of the signal does not correlate well with the punch perceived. Additional experimentation is required to establish which sub-bands play a more dominant role with respect to onset strength and the perceptual link to punch. As it stands the rhythm strength measure is considered not resolute enough. Subsection 6.2.2 indicated statistically that the attack and release times used in the wave shaper did have a bearing on the punch perceived by the listeners but unfortunately linear regression was unable to establish effective coefficients due to multi-collinearity being present.

The results confirmed the author's belief that rather than any one particular control setting being responsible for punch modification, a number of low-level features must be attributable. The important factor lies in not the process involved in audio modification but the inter-relationship between the controls resulting in the final signal. As such, further elicitation based on low-level feature modification is required.

As was shown, a high centroid value and greater spectral spread indicated some correlation with punch perception. Moments within a musical piece that exhibit greater spectral spread would generally correspond with percussive type components; these in turn would correspond with moments of punch perceived within the signal. Higher centroid values could be an indication of the culmination of energy present at the onset itself. It is therefore assumed that onsets within sub-bands and cross-band summation of energy within these onsets could be a useful predictor in

punch perception. In the case of the samples used in this experiment (kick drum), the overall spectral centroid is important in establishing the timbre at least lies within the boundaries expected of a percussive instrument.

Other low-level features identified that showed a degree of correlation with the punch attribute were attack and release times, spectral skew and rhythm strength.

The attack time and release times were separated into frequency bands chosen in the wave shaper, it would be beneficial to analyse the perceptual changes at a higher resolution in respect of number of bands tested, and indeed test each independently. Further to that, subsequent objective measures were taken based on both the full bandwidth and temporal response of the signal. Due to the correlation shown in attack and release times with the punch attribute, there are advantages in instigating signal separation prior to performing these measures. Signal separation in this context is the separation of the attack and release portions of a piece of audio from the somewhat steady state or harmonic portion (sustain portion). The rationale behind this was first mentioned in Section 1.2 as follows:

“...The onset of the transient present across octave bands affects the listener perception of punch, with the lowest octave attributing the most punch as the onset is decreased and vice-versa. Punch is therefore related to transient change and the energy density (summation across frequency bands) occurring at a particular moment in time and duration.”

The following chapter presents a hybrid multi-resolution technique for the signal separation of a mixed musical signal (and subsequent measurement of attributes contained within it).

## **Chapter 7 Hybrid multiresolution analysis of punch in music**

Decomposing music into simpler percussive, harmonic and noise components could enable a more detailed and focused measurement of signal attributes. For example, extraction of the percussive elements within a complex mix would enable the independent analysis of objective measures with respect to perceptual attributes, such as punch.

A hybrid multi-resolution technique that initially decomposed the musical signal using a quadrature mirror filter bank (QMF) before applying a short time Fourier transform (STFT) to each band was explored. QMF filters were chosen as their symmetry allows for perfect reconstruction of the audio during playback. By adopting this technique, it is possible to segment the signal energy into discrete bands and tune the STFT window size based on the frequency range of interest. The adoption of a hybrid system offers advantages over a single transform method. One advantage is a high degree of resolution can be achieved in both the time and frequency domains, this is explained later in Section 7.4.

Following the initial transform process, transient, steady state and residual components (TSR) are extracted. The method of separation uses iterative median filtering to achieve a high degree of separation into the TSR components. Median filtering is a technique utilized in image processing for edge detection and has been shown to give good results with low computational overhead when used for TSR separation.

Each of the components is then analysed using well-established spectral and time based measurements, e.g. spectral centroid. In addition, new measurements are investigated which explore the relationship between each component.

## 7.1 Sines, transients and residuals

Music can be considered to be a collection of complex components each with differing harmonic and non- harmonic attributes. These components can be categorized as a steady state, transient and residual.

The transient portion of a complex tone contains a great deal of information with respect to perceptual attributes of the source (Rasch & Rasch, 1981; Collins, 2005). In addition, given the transient information is inherently related to defined moments of change in a piece of music, this information is paramount in determining a punch measure. The experiment described in Chapter 5 also reinforces these points.

The transient part of the signal can be loosely defined as the initial time interval in which the signal is evolving into its steady state. The transient definition within the context of this thesis is defined in Section 2.11

Detection of transients can be useful in such applications as note detection, signal enhancement, dynamic range control and musical transcription (Avendano & Goodwin, 2004; Walsh et al., 2011; Wang & Tan, 2008; Zaunschirm et al., 2012). Various methods of transient detection can be employed with varying degrees of success depending on genre and application (Avendano & Goodwin, 2004; Zaunschirm et al., 2012).

Almost all genres of music have significant transient content throughout as a result of differing tone onsets. Onsets can be considered to have differing onset rates, e.g. drums would result in fast onset times whilst a bowed instrument such as a violin may have slower onset times. Despite having a slow onset, it can still be considered as having a transient characteristic initially.

Generally, transient information can be considered as the non-stationary components of a signal. Non- stationary being defined as a component that has a degree of magnitude or phase change within a particular time frame. Once transients have been detected, they can be enhanced or removed from the signal. The latter would result in the steady state and residual part of the signal remaining. This separation process is discussed in Subsection 7.4.2

The steady state components of the signal are usually related to pitched instrumentation. It's shown in Section 7.6 that analysis of this information independently can reveal parameters such as note length, scale and magnitude.

Residual components can be classified as neither steady-state nor transient. Consider noise within a signal, having both a random distribution of magnitude and phase within a time frame. The residual components relate therefore to the noise floor of the signal under test. Much in the same way that images can be de-noised, it's possible to de-noise audio signals resulting in the potential for increased clarity and to improve perceived audio in audio compression algorithms.

## **7.2 Source separation**

To precisely discriminate between transient, steady-state and residual components is not an easy task. Much work has been performed in this area and as such, excellent reviews and tutorials on the subject are available (Bello et al, 2005; Daudet, 2005). Considered opinion is that for sharp onset transients, the results of extraction are largely independent of the method chosen. It therefore makes sense to utilise methods that have minimum processing and latency load when considering audio metering applications.

The focus of this thesis is the modeling of the punch attribute, with a view towards its use in audio metering therefore, the more complex soft onset detection was not considered.

## **7.3 Fast onset detection method**

Fitzgerald (2010) proposed an efficient method of transient and steady state separation that utilised median filtering. This approach, inspired by Ono et al. (2008) considers that transient components will be broadband in nature with highly concentrated energy in time, whereas steady-state sources are taken as discrete narrow-band components with smooth magnitude temporal behavior. These components can be seen in spectrogram as vertical and horizontal ridges, respectively.

Further investigation utilising this method was performed Iraragay et al. (2013). Their work incorporated the use of a Wiener filter stage and a Stochastic Spectrum Estimation (SSE) method

proposed by Laurenti et al. (2007) which replaces the median filtering stage of the above with an alternative non-linear filter.

Through evaluation of the differing approaches with respect to relative performance and keeping in mind the need for simplicity, Fitzgerald's approach was adopted to detect fast onsets. However, the separation algorithm was modified to reduce spill between components. This modification, proposed by Driedger et al. (2014) incorporated separation factors which allow for the tightening or reduction of steady- state or transient bleed.

#### 7.4 Implemented analysis model

The chosen analysis model was implemented using MATLAB and is shown in Figure 46. It incorporates a filter bank in its first stage, which decomposes the signal into sub-bands. The advantages of this approach are that the subsequent processing can be tuned to the bandwidth of each sub-band (i.e. allow variable time and frequency resolution as required) and the sub-bands can be psychoacoustically tuned to the auditory response.

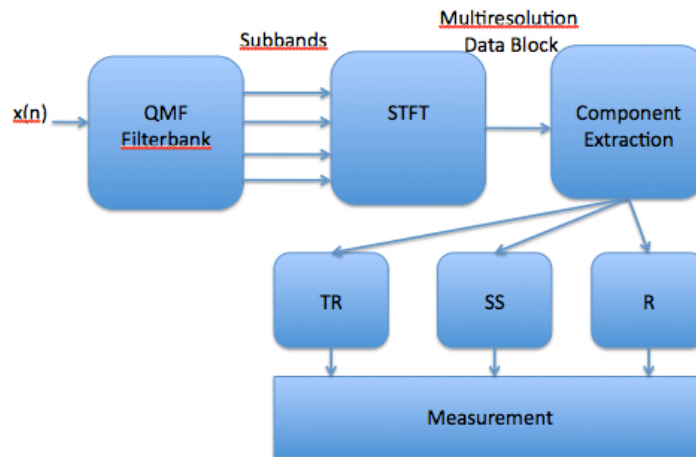


Figure 46 - Analysis Model

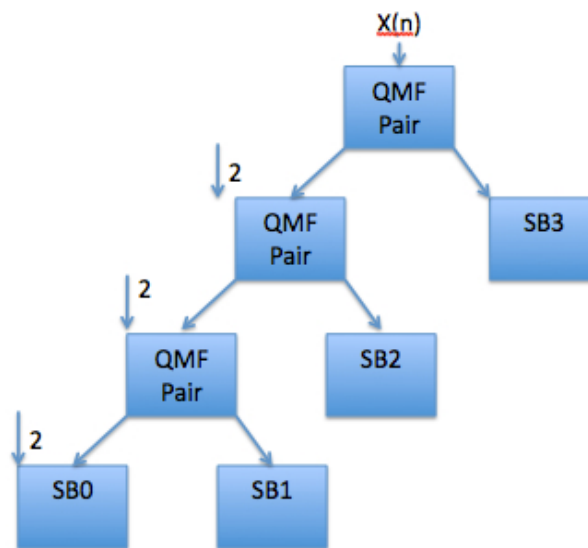
The choice of sub-band filtering was based on various factors, processing speed, possible reconstruction of the signal with minimal artefacts and also time alignment of resulting data.

Initially, stationary packet based wavelet decomposition was investigated (Learned & Willsky, 1995). The decomposition resulted in sub bands that were aligned in time and signal reconstruction was possible with no artefacts. However, this approach is highly redundant given

that each resultant packet contains all components between 0 and  $F_s$ , where  $F_s$  is the sample rate of the signal under test. If one considers a full packet wavelet tree at the lowest level of decomposition, each packet contains *equal* bandwidth components of  $F_s/L+1$ , where  $L$  is the level of decomposition. Given that the bandwidth of interest varies at each decomposition level, it makes sense to employ down sampling at each level thus reducing the data storage requirements whilst also increasing the frequency resolution at the lowest scale.

Utilising a full packet tree does have some advantages for signal classification (Learned & Willsky, 1995) for example an energy map of wavelet packets can be computed resulting in a feature set of a particular sound. This feature set can then be compared against a library of known sets resulting in identification or classification of the signal itself. This approach could be adopted, for example, in the case of a bass drum to detect not only whether a ‘hard beater’ or ‘soft beater’ had been used, but also the type and size of kick drum used during recordings.

For this work, a full packet tree decomposition was deemed unnecessary. As a model based on the auditory response requires lower resolution at higher frequencies, sub-bands could be chosen to reflect this. A critically sampled constant-Q filterbank of quadrature mirror filters (QMF) was employed to implement the filtering process. QMF filters are pairs of matched but reciprocal filters that are symmetrical about  $0.5\pi$ . By being matched they allow for perfect reconstruction should the original audio be required. Down-sampling by a factor of 2 is employed after each QMF filter stage, thus reducing data redundancy. To keep processing overhead to a minimum, 3 level decomposition into 4 bands took place as shown in Figure 47.



**Figure 47 - Filterbank of Cascaded QMF Filters**

This filtering results in four sub-bands, as shown in Table 15.

Sub band	Frequency Band (kHz)
SB3	11.025-22.05
SB2	5.5123- 11.025
SB1	2.756 – 5.5123
SB0	0 – 2.756

**Table 15 - Measurement Frequency Bands**

Each sub-band is then processed to give a time-frequency representation computed using the Short-Term Fourier Transform:

$$S(t, k) = \sum_{n=-\infty}^{\infty} w(n)s(n + tH)e^{-j\omega kn / N} \quad (14)$$

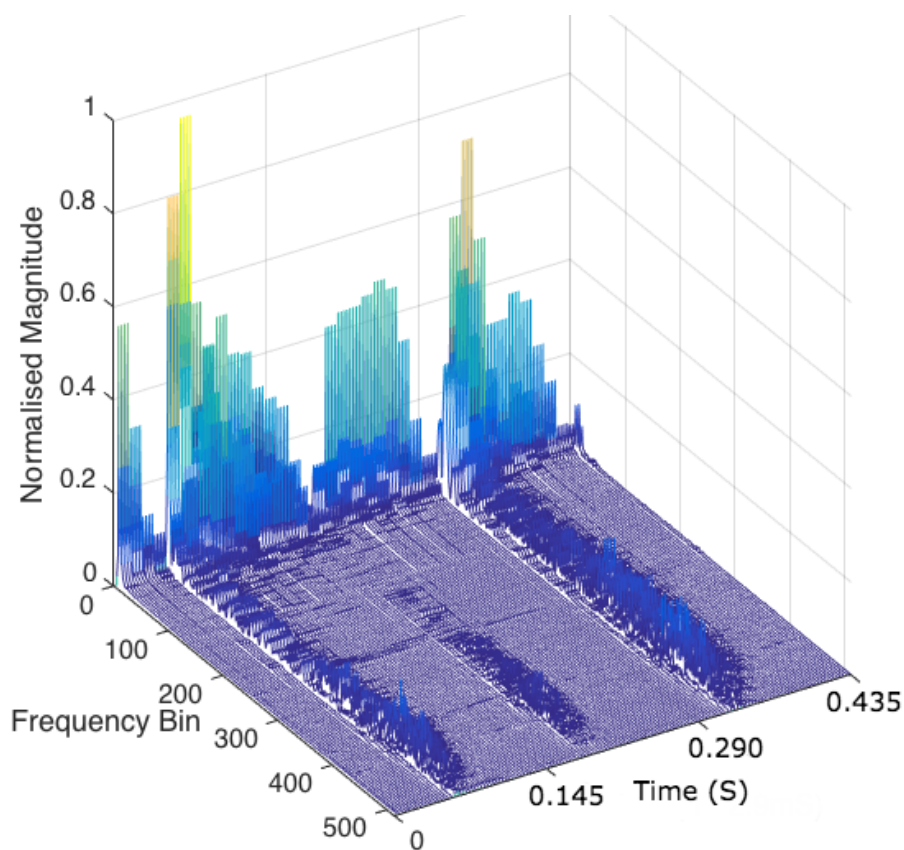
with  $t \in [0:T-1]$  and  $k \in [0:N]$ .  $k$  represents the number of bins  $N/2$ , where  $N$  is the DFT frame size.  $w(n)$  is a hann window and  $H$  is the hop size. The hop size was chosen to enable a 50% overlap.

#### 7.4.1 Multiresolution analysis

Due to the down-sampling nature of the QMF filterbank, the STFT window size is actually self-optimising with respect to the separation process. As explained in Section 7.1, strong percussive onsets tend to spread across the spectrum in a broadband nature, this spread tends to narrow in time in the upper frequency bands as the frequency components decay quicker. In order to capture this information in time, a shorter STFT analysis window is required. On the contrary, with respect to the low frequencies, these evolve much more slowly over time and require longer STFT analysis windows.

If one keeps the STFT frame size fixed, due to the signal down sampling, we are in fact able to analyse the signal on a multi-resolution basis in time. Thus, we achieve a system that has good time resolution in the upper sub-bands and good frequency resolution in the lower sub-bands, which is conducive to a psychoacoustic model.

The signal under test had sample rate of 44.1kHz. The chosen frame size was  $N=256$ . This resulted in fast computation and a hop size equating to 2.9ms. As outlined earlier, the same  $N$  frame size was adopted for each sub band, resulting in a hop sizes equating to 5.8ms and 11.6ms respectively. The lower 2 bands having the same hop size. The number of bins allocated for each frequency remains fixed relating to the frame size and is therefore 128. The resulting frequency resolutions for each of the four bands are 43Hz, 21Hz, 11Hz and 11Hz. The resulting STFT coefficients are then re-combined into an overall multi-resolution data block, a waterfall example of which is shown in Figure 48, before being passed through the median filters.



**Figure 48 - Multiresolution STFT of 'Animal' WAV.**

### 7.4.2 Separation of components

After the multi-resolution data block is produced it is fed into the component extraction block shown in Figure 46. Median filtering takes place within this block. The median filter operates by replacing a given sample in the block by the median of the sample values in a window around the sample. If the window size used is odd, the original sample is simply the middle value of the sorted window of samples. If the window size is even, the sample value is obtained by the mean of the two values in the middle of the window. Median filters used in this fashion suppress impulse noises whilst enhancing the steady-state and transient components. The window size used in this experiment was chosen as 13, which empirically proved to offer the best separation.

Median filtering performed across the time axis results in a steady-state enhanced data block, in addition transient outliers are suppressed. Likewise, filtering across the frequency axis tends towards suppressing the steady state components and enhances the transients.

This results in two median filtered STFT data blocks,  $S_t$  &  $S_{ss}$  representing the transient and steady-state components respectively. These data blocks are then used to produce two binary masks utilising steady state and transient thresholds defined as  $\beta_t$  and  $\beta_{ss}$ . Following the proposal outlined by Driedger et al. (2014), separation factors of 3 and 2.5 were chosen for  $\beta_t$  and  $\beta_{ss}$  respectively. These gave good separation when tested on a variety of sources.

The mask equations are defined as follows:

$$M_{ss}(t,k) = (S_{ss}(t,k) / S_t(t,k) + e) > \beta_{ss} \quad (15)$$

$$M_{tr}(t,k) = (S_t(t,k) / S_{ss}(t,k) + e) \geq \beta_t \quad (16)$$

Where  $t$  and  $k$  are the time and  $k^{\text{th}}$  frequency bin respectively. Separation is achieved by applying the masks to the overall multi-resolution data block which results in two separate transient and steady state data blocks:

$$TR(t,k) = S(t,k) * M_{tr}(t,k) \quad (17)$$

$$SS(t,k) = S(t,k) * M_{ss}(t,k) \quad (18)$$

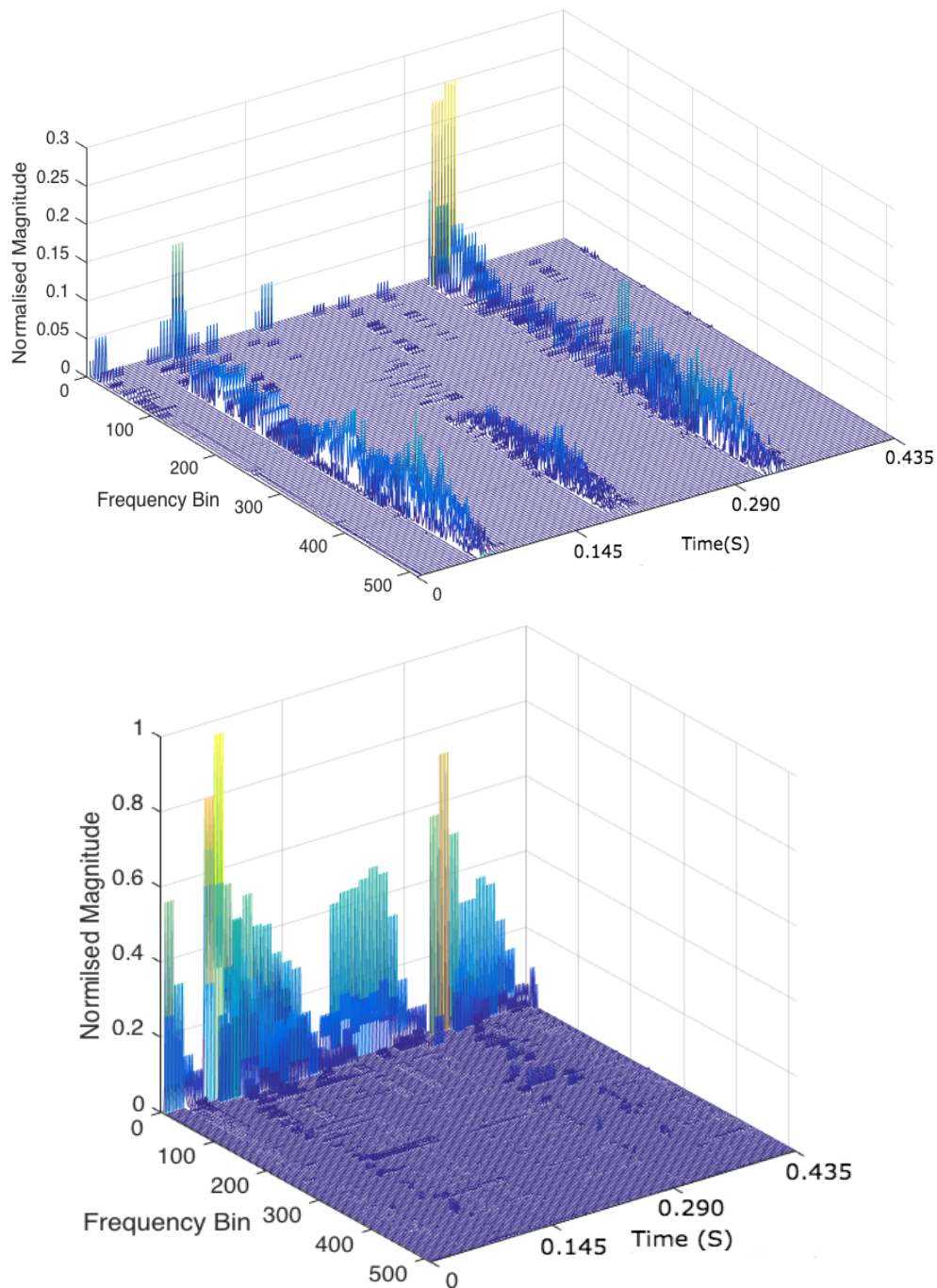
In addition, the method also enables the extraction of the residual components. The mask of which is defined as:

$$R_m(t,k) = 1 - [M_{tr}(t,k) \mid M_{ss}(t,k)] \quad (19)$$

The residual components are then extracted as:

$$R(t,k) = R_m * S(t,k) \quad (20)$$

An example of a separated file is shown in Figure 49 (a) and Figure 49 (b) which shows the transient and steady-state components respectively. The audio source formed the multi-resolution data block shown in Figure 48. The residual isn't shown as the resolution required in an image is insufficient.



**Figure 49(a) Transient (b) Steady State Median Filtering**

## 7.5 Analysis parameters

As outlined in Section 2.10 common metering tools used during the mixing and mastering consider the signal as a whole. Whilst being accurate for measures such as absolute peak level and overall RMS, a subsequent calculation of dynamic range (whatever the integration window size) is likely to be somewhat meaningless other than allowing a ‘loudness driven’ metric for target mixing or mastering. This is due to the RMS calculation used being the sum of all the signal components, those being the transient, steady state and residual.

The primary use of integration in dynamic range measures is to stabilise variations caused by the individual component parts of the signal. This is fine as a representation of ‘overall’ or ‘macro’ dynamics, but does nothing to represent the true nature of the audio with respect to microdynamic activity and listener perception. For example, during moments of true dynamic activity, one would expect a measure based only on the components that relate solely to this activity. such as the drums, or other onsets in the signal. Currently, dynamic range meters would utilise the steady state components of the signal in either the RMS or loudness calculations.

Through the use of component separation, it’s possible to measure elements within the complex musical signal either individually or grouped. The hypothesis being that this approach will give a more accurate objective representation of listener perception. In this analysis, the overall frame intensity of each component is calculated as a summation of each frequency bin for every STFT hop. Following this summation, each intensity block can then used as a separate or group measurement. The energy summation results in Transient (TR), Steady-state (SS) and Residual (R) intensity components.

A proposed measure of interest is the Transient to Steady State ratio (TSR). Considering the hypothesis outlined in Section 1.2 and Chapter 6 that punch perception is related to transient change at a particular moment in time in addition to the overall loudness at that time, this measure considers all three.

Should the steady state component intensity be significant at the timeframe of measurement, the transient components will inevitably be somewhat masked by the steady state components resulting in overall punch perception being affected. Conversely, should there be minimal steady

state component, the transient component has the potential to increase punch perception and itself, will not be masked.

In addition, it should be possible to measure the steady state signal without the detected transient components, thus determining the potential for masking.

The measure is given as:

$$TSR(t) = 10 * \log[TR(t) / SS(t)] \quad (21)$$

where  $TR$  &  $SS$  are the sum of the  $k$  magnitude bins at time  $t$  of the transient and steady state components respectively. The measure is expressed in dB. An additional parameter can also be measured which takes into account the residual component, as follows:

$$[TSR + R](t) = 10 * \log[TR(t)/[SS(t)|R(t)]] \quad (22)$$

This parameter can be likened to a dynamic range measurement in the presence of a signal i.e. with no noise-gating present. The level of noise or residual component is expected to affect the punch perception in addition to clarity within a complex mix.

Further to these parameters, spectral centroid measures were taken on a frame by frame basis of the transient and steady state components. Equation 23 shows the transient component centroid measure, where  $f(n)$  is the bin centre frequency. Steady state version is calculated in the same way:

$$SCtr(t,k) = \frac{\sum_{n=-\infty}^{\infty} f(n)TR(n,k)}{\sum_{n=-\infty}^{\infty} TR(n,k)} \quad (23)$$

Considering the spectral centroid of a complex mix of components, one would expect the measure to vary wildly and thus its use is somewhat limited for audio classification or mix/mastering purposes. It's expected that focusing the measure on isolated components may yield a more useful metric.

All the measures utilised are summarised in Table 16

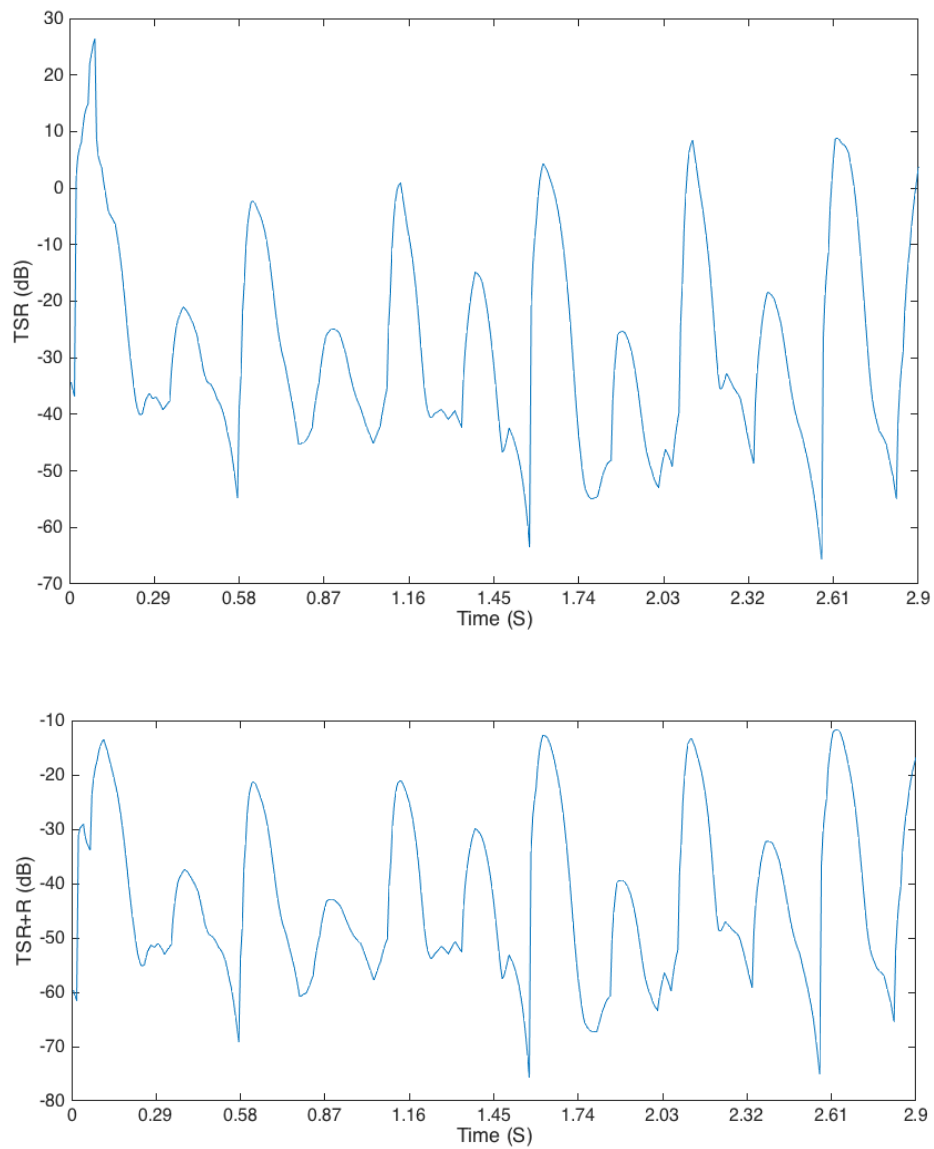
Parameter	Description
TR	Transient Component Intensity
SS	Steady-state Component Intensity
R	Residual Component Intensity
TSR	Transient to Steady State Ratio (dB)
TSR+R	Transient to Steady State Ratio plus Residual (dB)
SCtr	Spectral centroid of transient frame (Hz)
SCss	Spectral centroid of steady state frame (Hz)

**Table 16 - Measurements Performed**

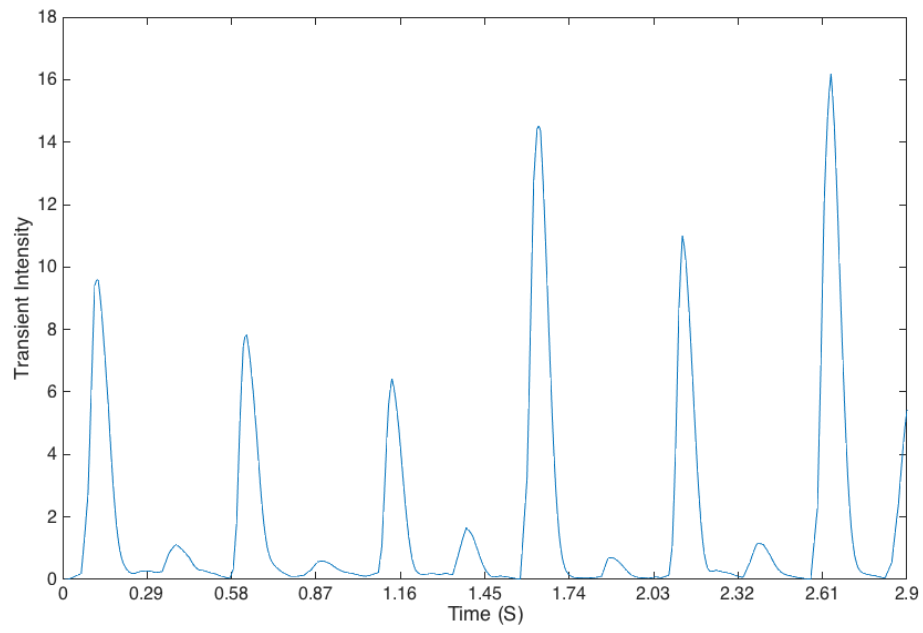
A raised-cosine (Half Hanning) filter can be applied to the resulting measures which further approximates to the integration present in the auditory response. A window size of approximately 100ms was chosen for this. Plots in Section 7.6 that have this filter applied are shown as ‘Smoothed’.

## 7.6 Results and discussion

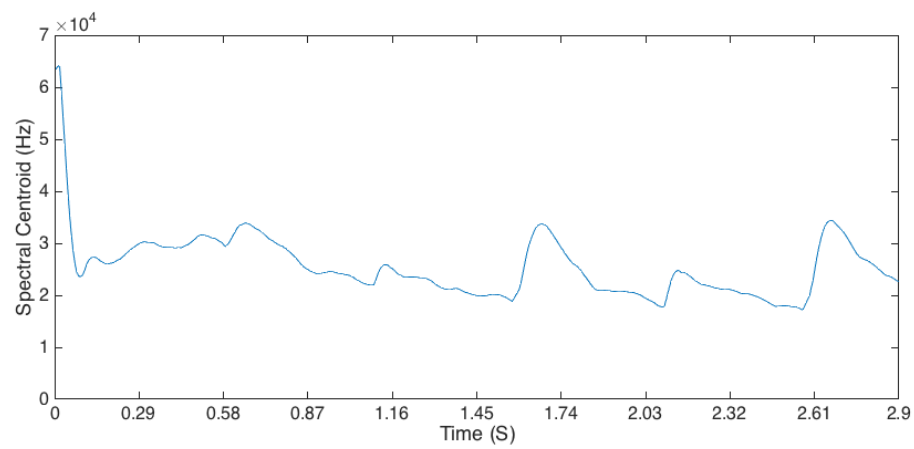
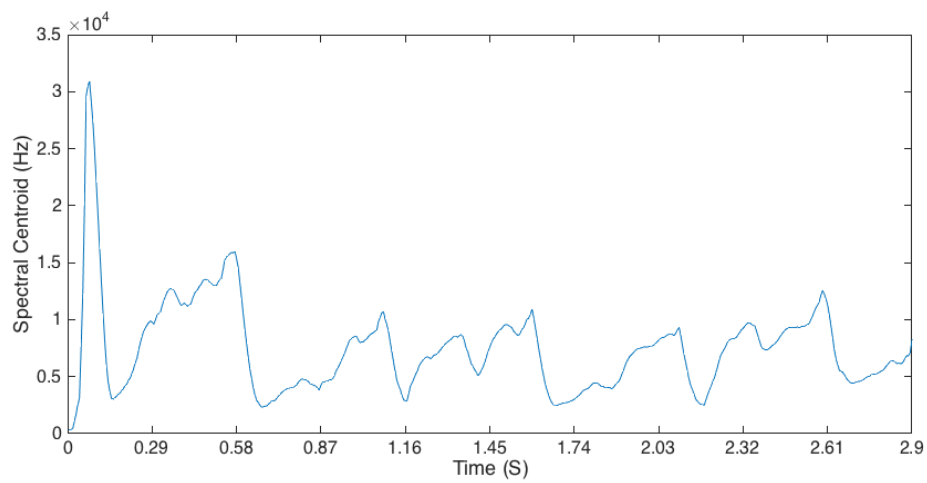
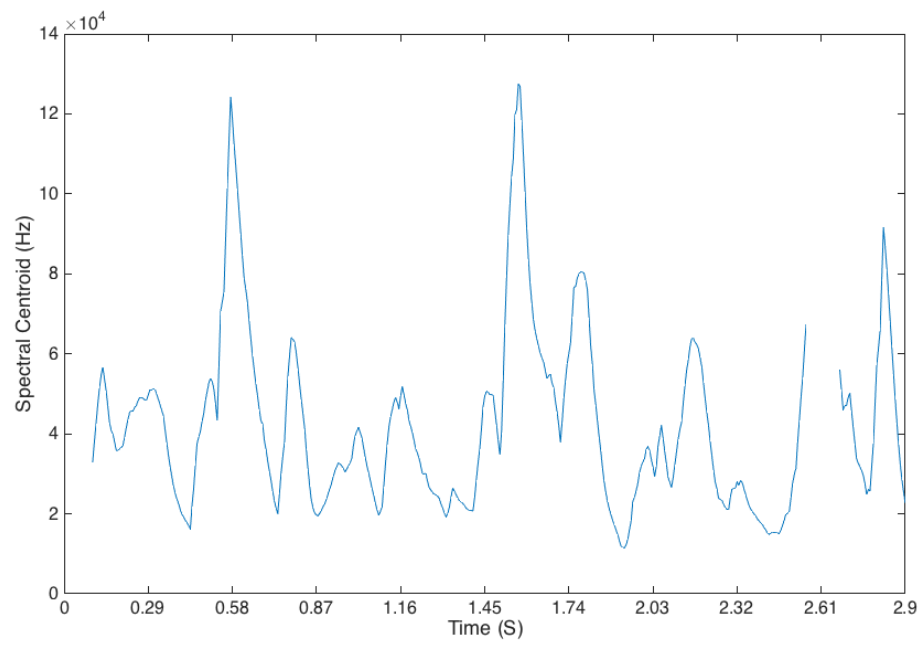
The sound sample under test was a 44.1kHz WAV file of the Def Leppard's song "Animal". The sample was converted to mono and normalised prior to measurement. The opening bars of the song were the point of measure.



**Figure 50(a) TSR and 50(b) TSR+R vs. Time**



**Figure 51 - Transient Intensity Summation vs. Time**



**Figure 52 - Spectral Centroid (a) Transient (b) Steady-state and (c) Overall**

With reference to Figure 50 showing the transient to steady state component ratio with and without the presence of the residual, one can see that the measure of dynamics is greatly increased. In the case of the non-residual calculation the maximum peaks average around -5 to -10 dB and in some case rise above the 0dB point whereas when the residual is considered the associated levels fall to between -13 to -22 dB. In addition to this, a small kink is evident at the start of the trace when the residual is included. This suggests that there is residual component energy fluctuation at the beginning of the sample.

The dynamics of the signal are clearly visible resulting from the presence of the intensity peaks within the extracted transient components. Of note is the addition of peaks present at 75 block intervals. These are as a result of the small intensity peaks in Figure 51 at the corresponding points in time. These peaks are due to a palm-muted guitar adding an additional percussive element to the arrangement. Ordinarily, this would not be visible when using standard integration based metering but their inclusion in the arrangement does add an additional punch element that should be considered.

The difference between the TSR and TSR+R measures is simply the inclusion of the residual components in the ratio calculation. As such, in a track with little in the way of bleed or excessive reverberation, the two measures would be very similar. In the case of a poorly mixed track where excessive noise is present, whether as a result of the recording process or amp hiss, the measures would give differing results. The TSR+R measure could therefore be used as a qualitative measure.

With the residual extracted, it should be possible to effectively de-noise a piece of audio much in the same way that an image is processed. By examining the residual and suppressing elements that may constitute unwanted noisy components a noise free signal could be recomposed. However, the residual may contain important information that can't be discounted completely, for example, the median filtering approach adopted tends to leave some of the lower level transient tails within the residual. An addition to the model could be employed to re-assign these tails to the transient component block. In addition, distortion may have been added to certain instruments to enhance timbre, these artefacts may well appear in the residual component and therefore may be deemed 'important' as far adding to the overall texture of a music track.

If the mix were such that the steady state components were made much louder, thus masking the transient elements somewhat, the peaks shown in Figure 50(a) & (b) would be expected to reduced accordingly. In the case of a piece of music without a strong percussive element, the transient components will be a result of the note onsets of other instrumentation provided they possess spectral spread.

For a track, where the sound sources had been mixed effectively with minimal masking there should be good transient intensity which will result in a high TSR being achieved, accurately representing greater perceived punch.

With respect to Figure 51 which shows the intensity of the transient (TR) component over time, each peak corresponds with either a kick, snare or palm muted guitar chord. If this measure were utilised for onset detection for drum transcription, the latter palm muted onset could be removed simply by the introduction of an ‘onset detection threshold’.

By utilising the spectral centroid of the transient component, which is shown in Figure 52 (a), fluctuation in peaks correspond to the nature of the audio under test, namely, the pattern KSKSKSK, where K and S represent Kick and Snare respectively. Therefore, unlike the centroid measure of the entire signal, Figure 52 (c), the measure could be useful in discriminating between percussive sources now that the centroid is independent of the steady state and residual colouration. The spectral centroid measure of the steady state component, Figure 52 (b), reveals the ascending nature of the frequency components resulting from the pitch bending guitar part present on every quarter note. Again, this is in contrast to the centroid measure of the entire mix, which reveals very little.

Due to a sub-band approach being adopted it is possible to tune the size of the median filters further to enhance the source separation. As each band has different time and frequency resolution at the sub band level, different values of median filter length should lead to more optimal separation. For example, it was noted with the model adopted, that the median filter applied across the vertical (frequency axis), tended to favour the higher frequencies rather than the lower ones, a larger median filter length improved this. In Dreiger et al (2014), different filter sizes in addition to DFT frame sizes were explored and this should prove very useful in progressing this research.

The inclusion of a soft onset detection mechanism should yield additional components that could be included within the transient data block, however, whether slower onsets would exhibit any correlation to punch perception requires more analysis. The use of the both phase deviation and weighted phase algorithms were explored and whilst effective in detecting the softer transients, they were too susceptible to noise such as that introduced by distorted guitars. A model utilising the Euclidian distance may be more useful in this respect.

The model utilises 4 sub-bands. A more elaborate and natural extension to this could be the implementation of a full auditory filterbank as proposed by Klapuri (1999) whereby TSR analysis could take place close to that of a natural hearing response. Octave band filtering may also yield interesting results, other alternatives could be the use of the Mel frequency-scale and Mel Frequency Cepstral Coefficients (MFCC).

## **7.7 Conclusion**

A hybrid multi-resolution model has been proposed that decomposes a complex musical signal into its transient, steady state and residual components. This allows the transient portion of the musical signal to be analysed independently from or in relation to the other components. The measures proposed are Transient to Steady-state Ratio (TSR) and Transient to Steady-state Ratio + Residual (TSR+R). The signals used in testing were complex audio (polyphonic) musical works consisting of drums, bass and guitars. The method was successful in achieving source decomposition and the associated measures have been shown to have possible implementation in both mixing/mastering and also audio transcription and retrieval. The method explored enables the possibility to perceptually weight the transient and steady state frequency bands and considering that lower spectral centroid components may perceptually exhibit more punch to the listener, the TSR measure could be weighted accordingly.

There is a need to expand the methodology further to incorporate more frequency bands and also to explore the perceptual relevance of onset times detected in each with the punch attribute. The following chapter expands this work by examining the onset time and frequency components of the signal across octave bands. The purpose of which is to extract perceptual weightings that could be utilised in a measurement model. An objective model is then proposed which utilises the weightings obtained from this analysis.

## **Chapter 8 Towards a perceptual model of punch**

This chapter presents a perceptual model for the measurement of ‘punch’ in musical signals. The model combines signal separation and low-level feature measurement to produce a perceptually weighted ‘punch’ score. The parameters explored were the onset time and frequency components of the signal across octave bands. The ‘punch’ score was determined by a weighted sum of these parameters using coefficients derived through a large scale listening test. The chapter concludes by evaluating the perceptual model using a small number of commercially released musical extracts.

Analysis of the results from Chapter 6, indicated the need to measure the effect of onset time independently with respect to listener perception of punch. Independently in this case means across each octave frequency band and in isolation. By doing so, it's possible to map the associated change in punch perception with respect to onset time and the octave frequency and derive weightings for each band. In order to facilitate this, a noise burst test was undertaken.

Pink noise bursts were chosen as the stimuli to facilitate equal energy per octave and a spectrum that is roughly similar to that of a musical signal. The noise was divided into octave bands to roughly correspond with the logarithmic hearing response.

### **8.1 Noise burst listening test**

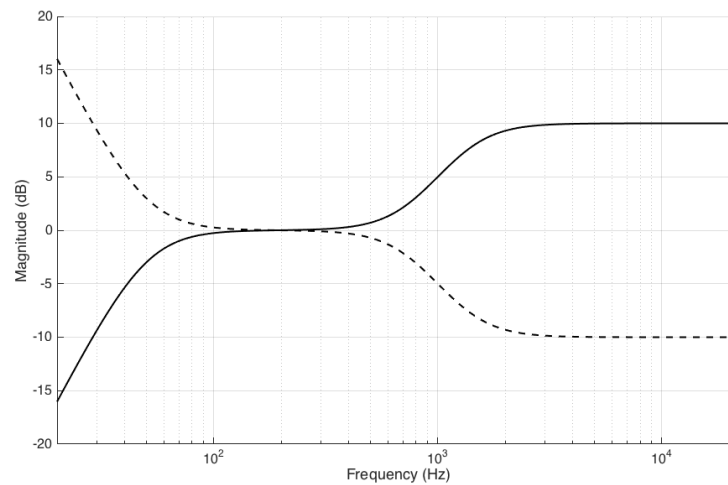
The controlled listening was test undertaken by 11 expert listeners. The test involved each listener listening to and grading the perceived punch of 45 shaped pink noise bursts. Each score was collated and a mean punch score for each onset at each octave frequency was calculated.

The 45 samples were all 100ms in length, 16 bit and presented in mono. The samples were constructed from 9 octave band limited pink noise bursts shaped by five onset times of 0ms, 5ms, 10ms, 20ms and 60ms. The number of onsets times were limited in order to reduce the testing time. All the samples had a fixed offset time of 40ms to help negate any offset effects. They were also loudness normalised using a two-stage filtering/gain algorithm, as described in Section 8.2.

The 100ms duration of the bursts was chosen as it closely approximated to the mean length measured on a collection of kick and snare drum samples. Their lengths being typical of these transient sources found in music. The onsets times were also chosen based on signal measurement data and to allow both a variation within the 100ms timeframe and the facilitation of a 40ms offset.

## 8.2 Loudness normalisation

In order to present the noise burst samples at equal loudness to the listeners, the samples were pre-processed with both spectral and temporal weighting coefficients. The spectral weighting curve was based upon that set out in recommendation ITU-R BS.1770-4 (2015). Modifications to the curve were adopted, namely the gain of the pre-filter was modified to 10dB as opposed to 4dB and the corner frequency of the shelf was adjusted to 1kHz rather than 1.6kHz.



**Figure 53 -Spectral Weighting Filter (and inverse as dotted )**

Figure 53 shows the spectral weighting filters. The inverse filter shown as the dotted line was used to pre-process the noise bursts prior to the temporal weighting being applied. These modifications were based on recommendations made by Pestana et al. (2013) and through the author's perceptual observations and testing.

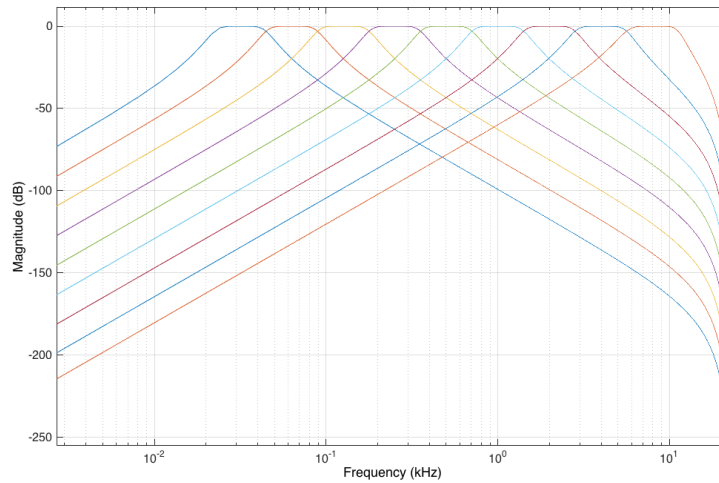
The filter specified in recommendation ITU-R 468-4 (1986) was also tested however it was found that the 2kHz-8kHz octave bands were perceptually significantly louder on playback than the lower bands.

Studies have shown that temporal effects, particularly for signals of less than 100ms in duration, must also be accounted for in order to weight the loudness appropriately (Watson & Gengel, 1969; Verhey & Kollmeier, 2002; Rasch & Rasch, 1981; Zwicker & Fastl, 1999; Moore, 2003).

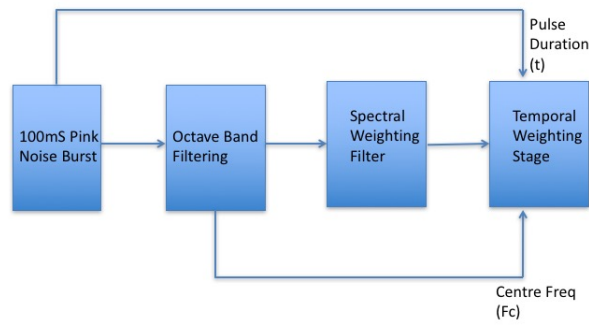
As such, a temporal weighting coefficient (*TWC*) was derived using both the centre frequency of the octave band (*F<sub>c</sub>*) and the signal duration (*t*). The process is shown in Figure 55. The octave band filters are shown in Figure 54. Calculation of Tau ( $\tau$ ) is first derived and this is then used to derive the weighting coefficient for each octave.

$$\tau = [-0.032 * \ln(F_c)] + 0.3095 \quad (24)$$

$$TWC = [1 - e^{-t/\tau}]^{-1} \quad (25)$$



**Figure 54 - Octave Band Filter responses**

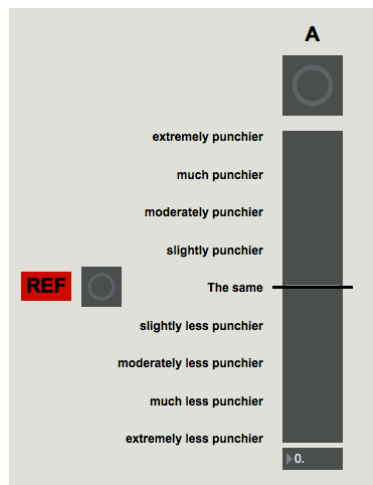


**Figure 55 - Loudness Compensator**

An informal listening test took place prior to the main test to briefly evaluate the effectiveness of the loudness normalisation algorithm. Of the 11 participants that took part, 7 agreed that the loudness of the pulses were presented at roughly equal loudness. The remaining 4 suggested that the differences they were hearing were primarily as a result of timbral differences rather than loudness. Adjustments proposed were recorded but not utilised in the main testing as the majority of changes proposed were of a magnitude of 2dB or less.

Each octave band pulse was measured at the listening position using an SPL meter without weighting. The SPL measured was 76dB.

Following the informal loudness evaluation, the 45 shaped noise burst were presented using a test interface created using HULTI-Gen test suite for Max (Gribben & Lee, 2015). Figure 56 shows the test interface. Each sample was played in a random order and each participant was asked to rank the perceived punch against a 1kHz band reference burst. The reference had a 0ms onset and 40ms offset. Scores were collected using a 100 point scale with 50 and -50 corresponding to the extremely punchier and extremely less punchier limits respectively. A single stimulus experimental interface was preferred over a multiple stimulus due to the number of samples involved.



**Figure 56 - Test Interface - Punch Perception Test**

### 8.3 Noise burst test results

The results in Figure 57 show the noise burst listening test results. The plot represents the punch scores obtained at each octave band and onset. The '0' point on the y-axis indicates the noise burst was perceived to have the same punch level as that of the 1kHz noise burst. A positive value indicates increased punch, negative indicates less. All result plots indicate a maximum and minimum punch scale limit of 1 and -1 respectively.

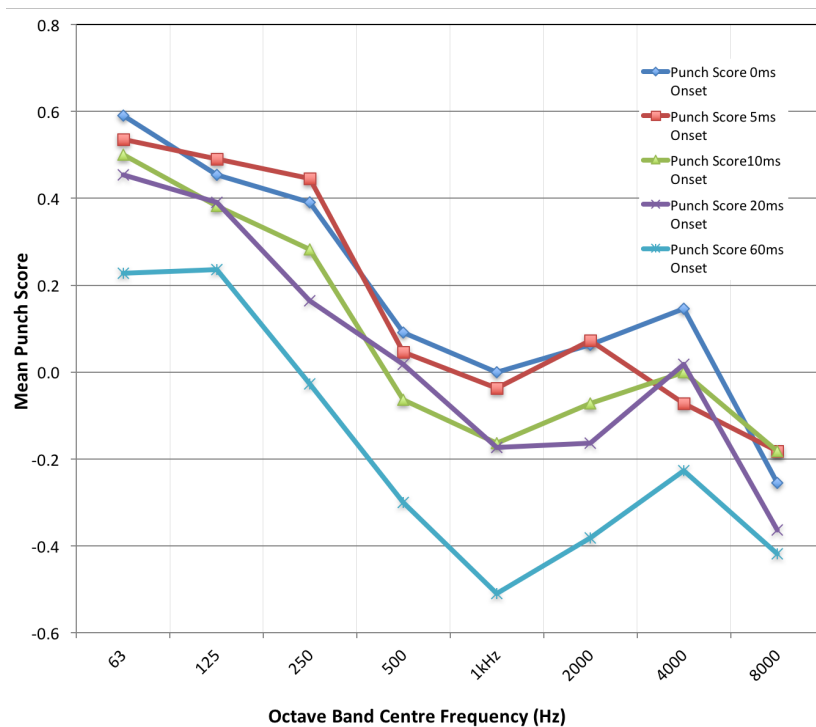


Figure 57 - Mean Punch Scores vs. Octave Band per Onset

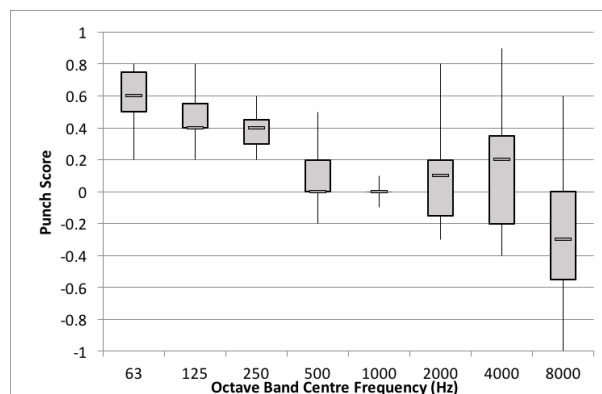
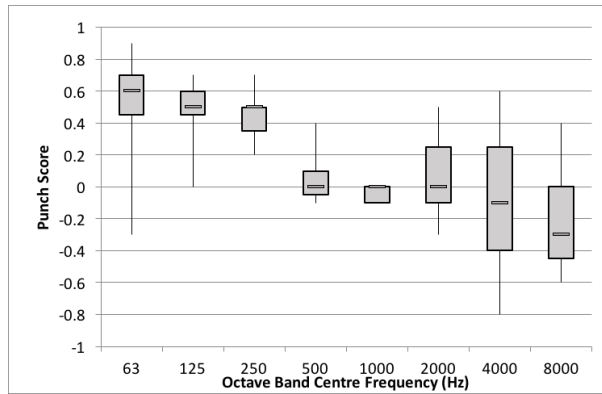
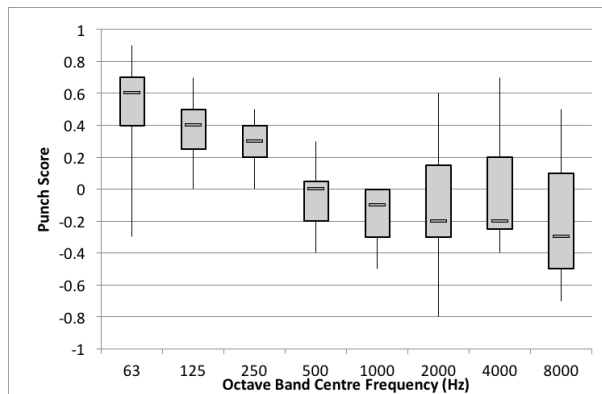


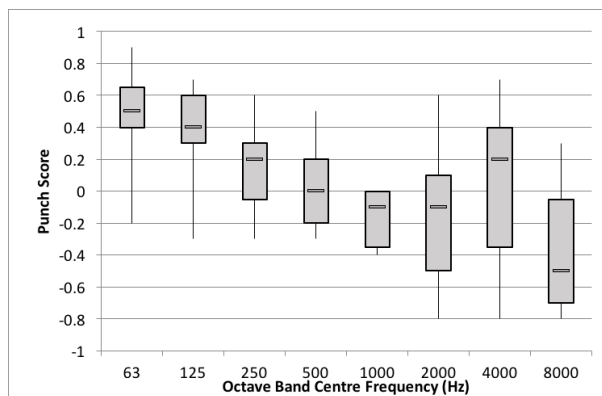
Figure 58 - Punch Score 0ms Onset Noise Burst



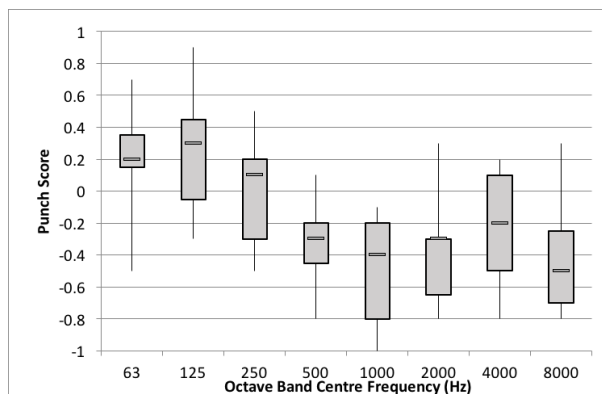
**Figure 59 - Punch Score 5ms Onset Noise Burst**



**Figure 60 - Punch Score 10ms Onset Noise Burst**



**Figure 61 - Punch Score 20ms Onset Noise Burst**



**Figure 62 - Punch Score 60ms Onset Noise Burst**

## 8.4 Test analysis and model parameters

Initial analysis of the results shown in Figure 57 indicated that the mean punch score was related to both the noise burst centre frequency and onset however there appears to be a pivot point around the 1kHz reference band whereby the 2kHz and 4kHz bands have an upward trend. Figure 58 through Figure 62 show the results in box plot format showing the distributional characteristics of the punch scores as well as the levels.

With reference to the inter-quartile ranges, the greatest degree of variation of punch scores can be seen in the 2kHz, 4kHz and 8kHz bands. Again this trend, in addition to the upward trends shown in Figure 57 suggest a greater difference of opinion as to whether the upper bands have a greater punch perception than the 1kHz reference.

Upon interviewing the participants after the experiment, it was found that whilst the upper bands were not necessarily more punchy than the reference, in some cases some had been scored higher as a result of their timbral weight or presence. This is particularly relevant to the 4kHz band.

The results obtained, combined with the comments made by listeners in the informal loudness test, suggest that even with the modifications employed to the hi-shelving gain (Pestana et al, 2013) of the weighting filter, there is still some discrepancy in the 2k to 4kHz region. Some listeners perhaps judging punch directly by perceived loudness level or other timbral attribute. This might also warrant the filter within the current ITU-R BS.1770-4 (2015) loudness standard to be re-visited. The variance observed above the 1kHz reference could suggest that listeners may not even consider punch as relevant to these stimuli.

Multiple linear regression techniques were applied to the results of the noise burst experiment in order to derive model parameters. Due to the pivot point around 1kHz a model utilising only the 64Hz to 1kHz centre frequencies was employed. The dependent variable in the regression analysis was the Punch Score based on the mean of the listener test results and the independent variables were Band and Onset.

<b>R</b>	<b>R<sup>2</sup></b>	<b>Adjusted R<sup>2</sup></b>	<b>Std. Error of Estimation</b>
.952	.906	.898	.09321

**Table 17 - Summary For Punch Score Model**

	<b>SS</b>	<b>df</b>	<b>Mean Squ</b>	<b>F</b>	<b>Sig</b>
Regression	1.846	2	.923	106.22	.000
Residual	.191	22	.009		
Total	2.037	24			

**Table 18 – ANOVA Summary For Punch Score Model**

The Sum of Squares of the residuals is the total deviation of the response values to that of the model prediction values, a value of .191 (see Table 18) shows a very tight fit of the model to the data. The resultant coefficient of determination shown in

Table 17, (denoted by  $R^2$ ) is derived from the sum of squares and further indicates that the model is a good fit. The adjusted  $R^2$  value is not adversely affected by the low number of data points (11 expert listeners and 5 bands used). The standard error predicted by the model is also very small. The effect of the regression is statistically significant, thus the effect on the punch scores is mainly predictable through variation of the independent variables rather than chance.

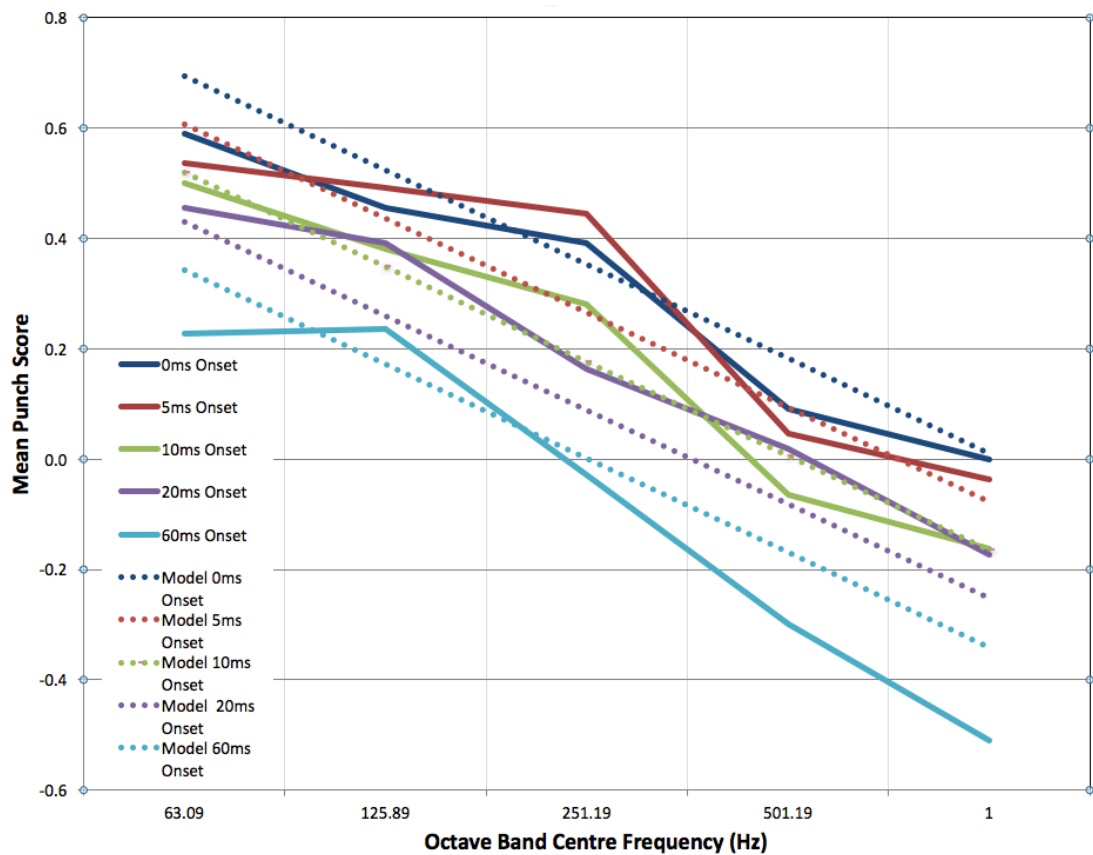
	<b>Unstandard Coefficients</b>		<b>Standard Coefficients</b>		
	<b>B</b>	<b>Standard Error</b>	<b>Beta</b>	<b>T</b>	<b>Sig</b>
<b>Const</b>	.954	.059		16.185	.000
<b>Band</b>	-.171	.013	-.845	-12.941	.000
<b>Onset</b>	-.088	.013	-.438	-6.706	.000

**Table 19 – Coefficients For Punch Score Model**

Table 19, showing the raw coefficient score (partial regression coefficient scores B), indicate that the Band would have both the largest effect and be detrimental to the punch score as it is increased. This trend can be observed in Figure 58 through to Figure 62. The effect of the Onset is somewhat smaller, again having a detrimental effect as it is increased. One might expect, and indeed it can be observed particularly from Figure 62, that the longer onsets have more of an impact in the upper bands with respect to a lowering of the punch score.

The resultant estimated linear regression equation is as follows:

$$\textit{EstPunch} = .954 - 0.171\textit{Band} - 0.088\textit{Onset} \quad (26)$$



**Figure 63 - Model vs. Subjective Results**

Figure 63 shows the model output compared to the subjective results obtained. The results show that the subjective scores have a general trend of punch decrease as the octave band increases. In addition to that, as the onsets are increased within each band, there is a general trend of decreasing punch. 60ms onset across all bands shows the lowest level of punch. 0ms onset shows the most level of punch in bands 1, 4 and 5 and in the case where 5ms is deemed to have greater punch (in bands 2 & 3) the punch mean punch scores given by the listeners is very marginal in difference. This is reflected by the box plots shown in Figure 58 and Figure 59 respectively. It was deemed appropriate to implement a linear model in the first instance however; non-linear modelling could offer the possibility of higher correlation to the subject scores.

## 8.5 Punch model implementation

The derived linear regression equation was incorporated into a punch model. The model employs a multi-stage approach as shown in Figure 64. Firstly, the complex musical signal is separated into its component transient, steady state and residual components. This technique allows the transient portion of the musical signal to be analysed independently of the other components which has significant advantages over approaches that consider the whole signal in its mixed state.

The separation process outlined in Chapter 7 offered both good time and frequency resolution by employing a quadrature mirror filter bank (QMF) and applying a short time Fourier transform (STFT) across each sub-band. Whilst this offered good results in terms of time/frequency the implementation is somewhat processor intensive in that it requires very high order FIR filters in its implementation. A decision was made to implement the separation stage utilising STFT and median filtering alone thus reducing processor overhead and evaluating the model performance with a view to moving towards a real-time implementation.

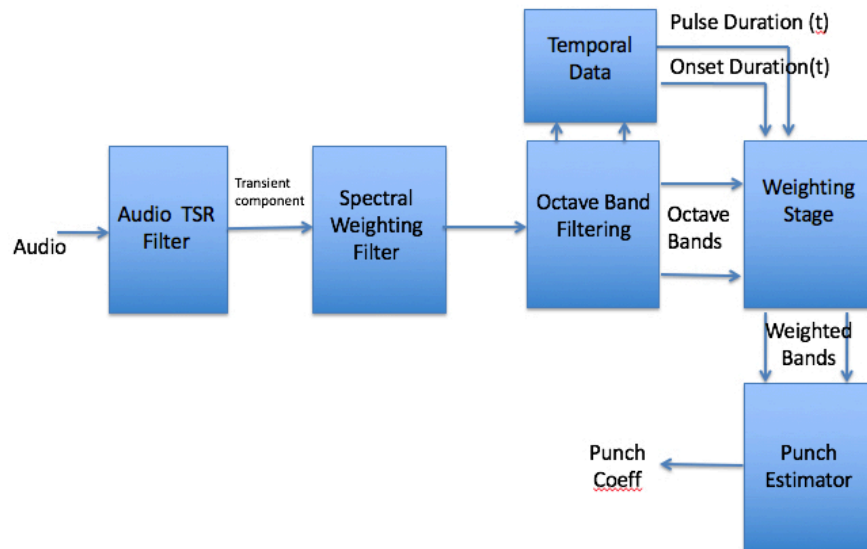


Figure 64 - Punch Model Diagram

The transient component is fed into a spectral weighting filter, as described in Section 8.2, followed by an octave band filter to separate the signal into loudness weighted bands. Each band is fed into a side-chain onset detection stage which extracts onsets and their relative times within each band. Each band is used to produce a punch coefficient based on Equation 26 and the onsets

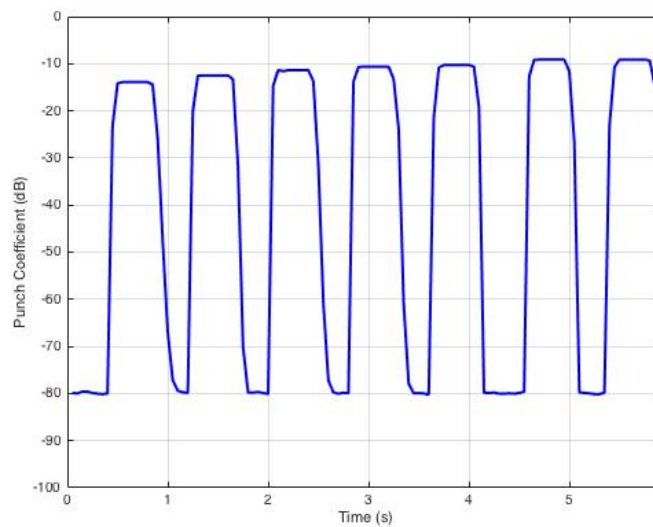
detected within it. Finally, the punch coefficients across bands are summed together to provide an overall punch indicator.

The summing process is based on the block based momentary loudness model as specified in ITU-R (2015) BS.1770-4. However a smaller block size of 100ms with a 50% overlap was incorporated to lower the level of signal integration taking place and also to allow for suitable alignment with the onset detection data (shown as Temporal Data in Figure 64).

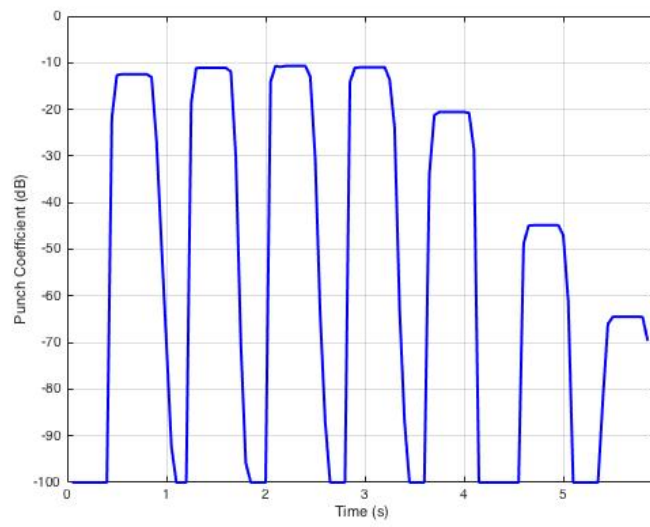
The model was implemented using MATLAB. The onset detection data was created through the utilisation of the *MIRonsets* function provided in MIR Toolbox (Lartillot & Toivainen, 2007).

## 8.6 Model output

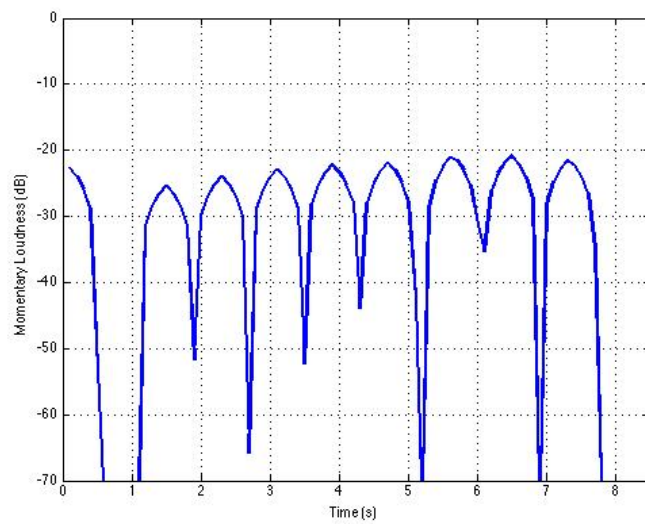
The following plots represent the model output with respect to varying input stimuli.



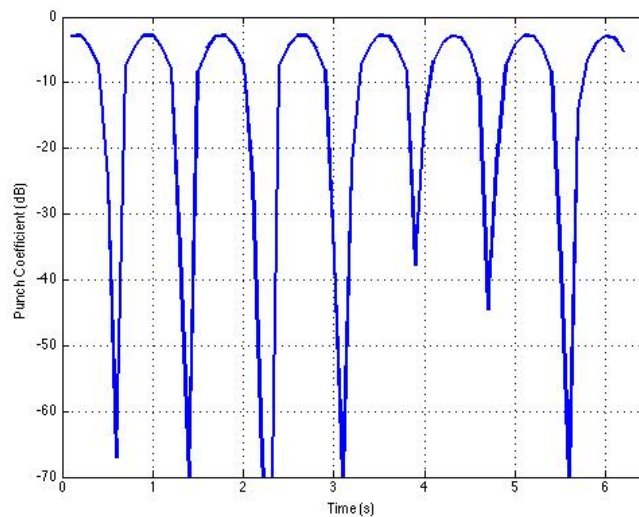
**Figure 65 - Measurement of Noise Bursts (Progressive Octave Bands) 100ms Block Size**



**Figure 66 - Measurement of Noise Bursts (Weighted Progressive Octave Bands) 100ms Block Size**



**Figure 67 - Measurement of Noise Bursts using standard 400ms Momentary Loudness Model**



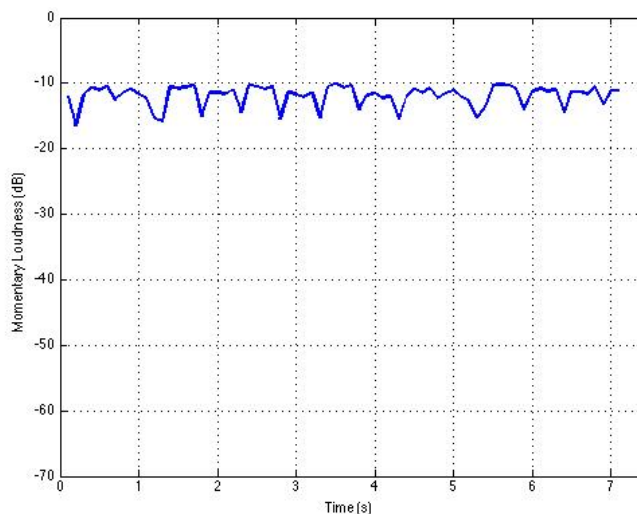
**Figure 68 - Measurement of Full Scale & Full Spectrum Noise Burst using the punch model**

Figure 65 and Figure 66 show the output of the model when presented with the test stimuli used in the subjective tests, these are the octave spaced progressively increasing centre frequency noise bursts with 0ms onsets. Figure 65 simply shows the output without weightings being applied and Figure 66 is with weighting applied. This is to allow comparison with the standard momentary loudness model output shown in Figure 67.

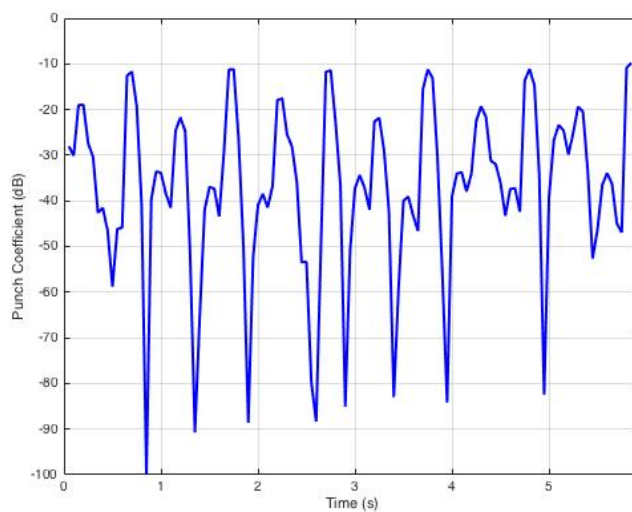
Comparing Figure 65 and Figure 67, the outputs of the models are very similar. However, due to the smaller integration window used in the punch model, the output is able to clearly differentiate the individual noise bursts and associated level. Figure 66 is showing the output of the model with weightings enabled and consideration of the first 5 bands being summed. As the noise burst centre frequency is increased, the output of the model drops due to the weightings employed, this is as expected when compared to Figure 65.

The 0dB point on the figures represents full scale, similar to that of the standard loudness model. That is, if the input stimulus is a full scale digital broad band pink noise burst, the output of the model would be -3dB. This is shown in Figure 68.

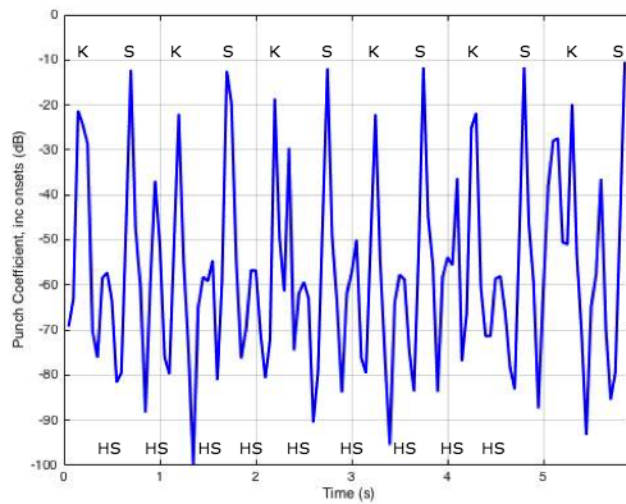
Figure 69, Figure 70 and Figure 71 compare the output of punch model with a standard momentary loudness model using the opening bars Michael Jackson - ‘Billie Jean’. The opening bars consist of kick and snare followed by the introduction of bass and synth melody.



**Figure 69 - Measurement Of Billy Jean Sample, using standard 400ms Momentary Loudness model**



**Figure 70 - Measurement Of Billy Jean Sample, using 100ms Punch Model (weighted)**



**Figure 71 - Measurement Of Billy Jean Sample, using 100ms punch model (weighted with onsets)**

Using loudness as a metric as shown in Figure 69, it can be seen that there is little that can be obtained from his plot in terms of signal dynamic. Comparing this to the model output, Figure 70, the signal dynamic with respect to the transient components is clearly visible. More punch is evident at approximately the 0.7 second point continuing at 1 second intervals starting from 0s,

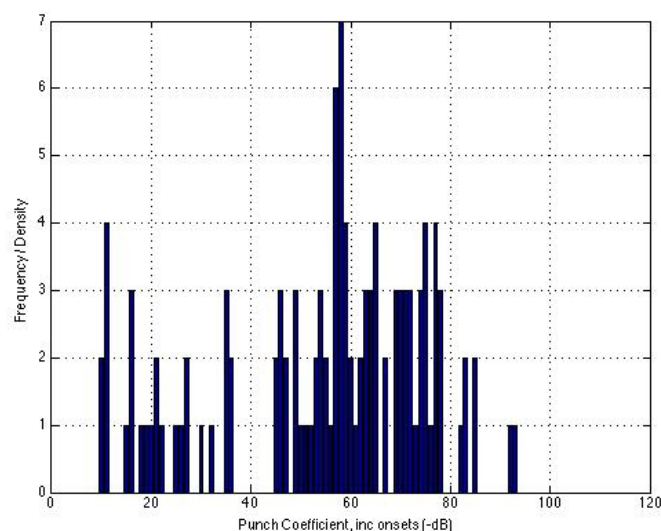
Figure 71 is the output of the model with the onsets active, i.e. the transients are being considered along with their relative rise times. The punch output is much more resolute in identifying the underlying dynamic of the music, with smaller punch peaks being associated with the closed hi-hat and shaker denoted as HS. K & S indicate kick and snare respectively. It can be seen that in some cases the peak level of punch is affected, this is to be expected, as the algorithm is no longer simply summing overall energy within a frame, only weighted energy associated with onsets is summed. In addition, where an onset is not detected within a frame, the frame is ignored from the energy summation therefore resulting in sharper tangents being visible in the plot.

## 8.7 Statistical output

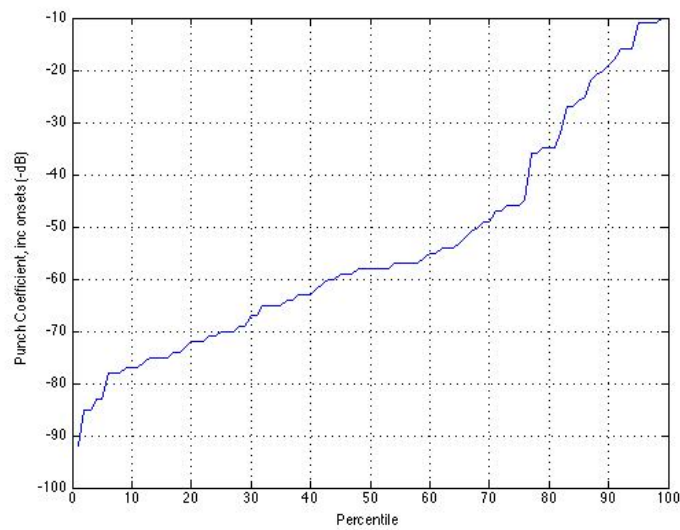
Figure 70 and Figure 71 model outputs, along with suitable metering ballistics, may prove more useful in offering a real time indicator of punch to an engineer. A more useful application of the model output might be for evaluating the punch scores over a period of time, for example in a histogram. A similar approach was adopted with the IBR measure, detailed in Section 5.6. The output variables obtained from the histogram can be analysed, giving statistical data relating to the stimuli under test.

The histogram shown in Figure 72 Figure 72 represents the data extracted from the Billy Jean sample with the onsets active. It shows the frequency of particular punch frame magnitudes within the section of music under test. By examining the data in this way, maximum, minimum, median and standard deviation measures can be extracted.

This data can be shown in the form of a percentile plot, Figure 73, giving an effective insight into the underlying ‘punchiness’ of the music being measured.



**Figure 72 - Histogram of punch scores detected in Billy Jean Sample.**

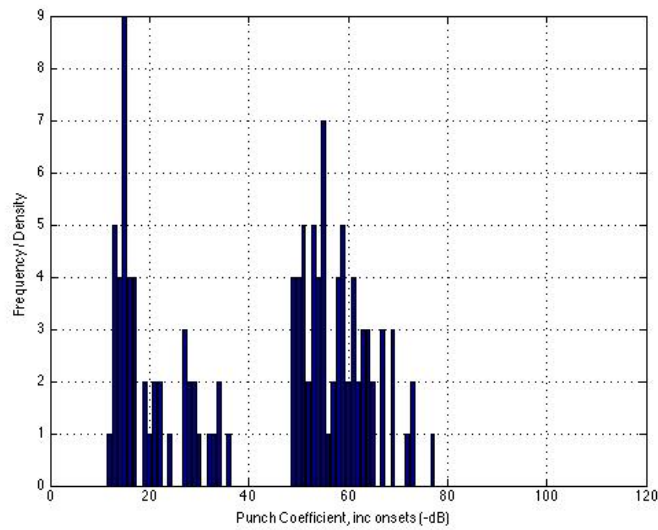


**Figure 73 - Percentile plot of punch scores detected in Billy Jean Sample.**

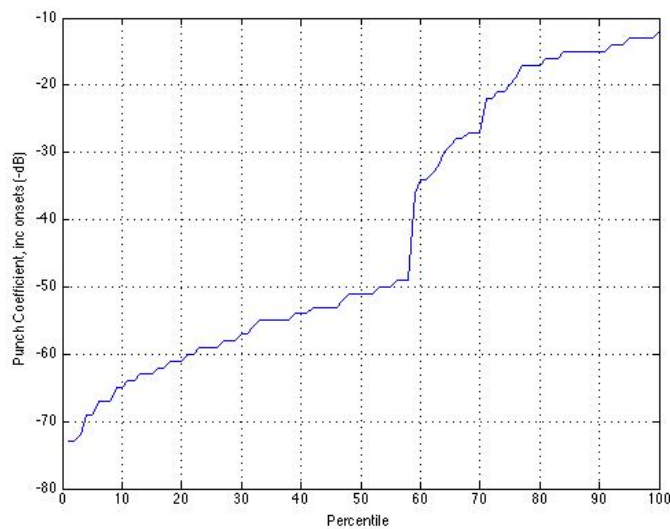
With reference to Figure 73, it can be observed that punch frames of -20dB or more only occur 10% of the time. Conversely, punch frames of -20dB or less occur 90% of the time. Similarly to the current loudness range metering algorithm, ITU-R BS.1770-4 (2015), and its ‘Loudness Range’ measure, upper and lower percentiles could be ignored therefore resulting in a ‘punch range’ measure being extracted. In addition, a peak punch to average may be of use to indicate dynamic variability between audio stimuli. The punch model naming conventions proposed are shown in Table 20.

Model Abbreviation	Description
PM	Punch Model Raw Score
PMx	Punch Model Range Using $x^{\text{th}}$ Percentile
PMxM	Punch Model Range Using $x^{\text{th}}$ Percentile, divided by the Mean

**Table 20 – Punch Model Naming and Description**

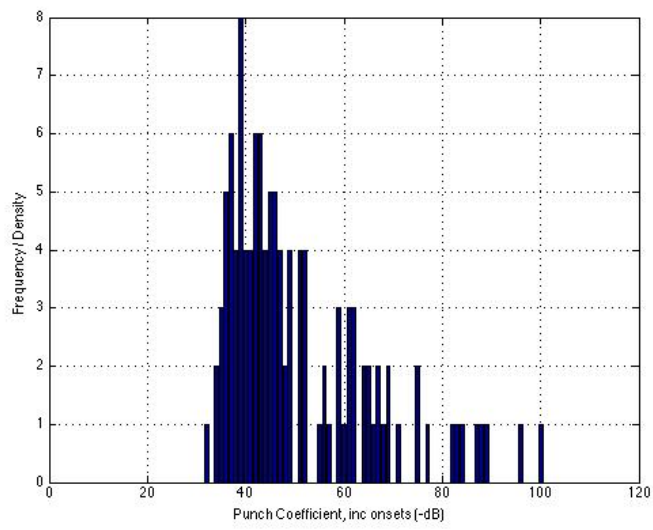


**Figure 74 - Histogram plot of punch scores detected in Rage Against The Machine sample.**

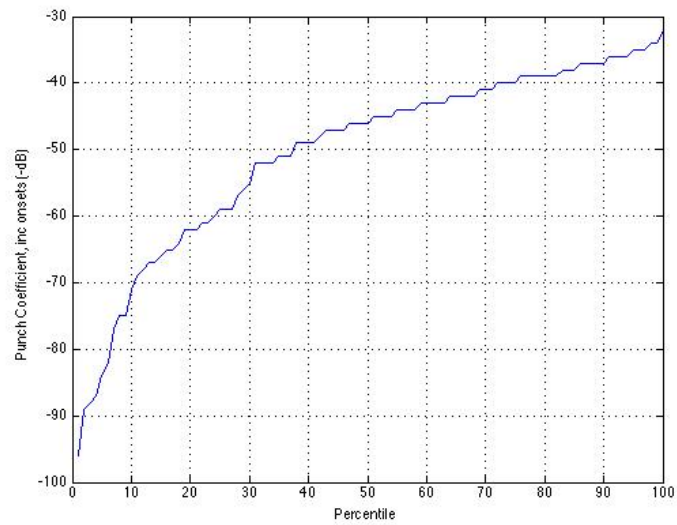


**Figure 75 - Percentile plot of punch scores detected in Rage Against The Machine sample.**

Figure 74 and Figure 75 show the model outputs for an excerpt from “Take the power back” by Rage Against The Machine . In the author’s opinion, perceptually this track is punchier than Billy Jean. Examining the model output in Figure 75 it can be seen that 30% of the frames detected were -27dB or more, this contrasts with the Billy Jean sample whereby 30% of the frames detected had a range of -49dB or more. To put these outputs in perspective, Figure 76 and Figure 77 show the output of an ambient track consisting of synth drones and very little percussive element. No frames are detected that exceed -32dB and the high concentration of frames around the -40dB point indicates a low punch score with very little deviation.



**Figure 76 - Percentile plot of punch scores detected in the ambient sample.**



**Figure 77 - Percentile plot of punch scores detected in ambient sample.**

## 8.8 Conclusions

The model is based around the perceptual weighting of the transient components of an audio signal utilising the onsets detected within octave bands to produce an objective punch output. The raw punch score measure is defined as PM, the statistical variants are PM<sub>x</sub> and PM<sub>xM</sub>, where x is the upper percentile used in the derivation. The output, in its raw form, shows a correlation with the sensation of punch however the model requires validation with a controlled listening test and a large and varied array of input stimuli.

Comparison of the model to other objective models associated with dynamics and punch will now be investigated along with differing statistical variants. CF, IBR, PLR and LDR are included in this testing. This work is undertaken in Chapter 9.

## Chapter 9 Validation of the punch model

A listening test was conducted to measure the perceived punch of a set of musical stimuli of differing genres in order to evaluate the validity of the punch model proposed in Chapter 8. The results of the listening test were then compared to the model output. The goal was to quantify the perceived punch of each stimulus as an attribute of the sound itself thus forming a varied test set that had been perceptually graded. The use of differing genres was to test the validity of the output regardless of genre. The validation test involved the following stages, subjective test and score collation, objective measurement of the stimuli and correlation analysis of various model output variables with respect to the subjective punch scale scores.

Two punch model outputs were compared against the subjective data in order to evaluate the effectiveness of each, these were PM95 and PM95M. Other objective measures were included to evaluate their correlation against the same derived punch scale. The measures were CF, PLR, IBR, IBR\_diff and LDR.

### 9.1 Overview of the objective measures

The following outlines the objective measures compared in the validation test.

PM95 - This is the punch model output indicating the punch range across the stimuli based on the lowest and 95<sup>th</sup> percentile of punch frames measured. The PM95M is the PM95 value divided by the mean punch frame value.

Crest factor (CF) – This indicates the peak to average ratio of the signal. The ‘peak’ is the maximum amplitude level and the ‘average’ is the RMS value (Hartmann, 1998).

Peak-to-loudness ratio (PLR) is similar to CF, except overall loudness level is used instead of its RMS value (ITU-R (2015) BS.1770-4).

Loudness Dynamic Range (LDR) is a measure of microdynamics within a signal. It has been shown to be more robust in this sense that the CF and PLR measures (Skovenborg, 2014). The LDR measure utilised in this testing was based around a 3 second ‘slow’ and 50ms ‘fast’ window size and 95% percentile. This corresponds with the settings giving the most correlation to ‘microdynamic’ perception.

IBR is a measure of dynamic range correlation between frequency bands, the measure was introduced in the early stages of this work in Chapter 3, Chapter 4 and Chapter 5. Section 5.6 described a method of representing the IBR measured frames in statistical format and a measure was proposed that looks at the range of IBR frames within the entire audio excerpt. Hyper-compression in music tends to result in lower IBR scores than uncompressed music. It was concluded in Section 5.6 that the IBR measure shows a stronger correlation to punch/audio quality than a broadband dynamic range measure. The measure utilised in this testing is the IBR\_diff, which is the range measured between the 1<sup>st</sup> and 95<sup>th</sup> Percentile.

## **9.2 Experimental design and listening conditions**

A forced pairwise comparison test was adopted which presented randomised pairs of stimuli to the listeners. The listeners simply had to select the stimuli they thought exhibited the most punch. 12 stimuli were utilised and 11 expert listeners took part in the test. Each listener made a total of 66 comparisons as they compared each stimuli to every other.

This comparison test was chosen over a typical ranking test to reduce biasing and to allow each stimuli to be compared equally. If a straight ranking test had been adopted, then the listeners may have been tempted simply to rank an artist or genre with a preconception of it having the most or least punch. Forced choice was chosen as it was deemed appropriate to reduce the measurement variance of the subjective data for later comparison.

The listeners were given the opportunity to listen to the stimuli prior to the test and were instructed not to base their choices on melody, genre, personal taste or arrangement. This enabled them to adjust to the listening environment and also gauge the range of stimuli they were going to rate. This was an important part of the training phase such that the listener was aware of the full suite of stimuli that was going to be used during testing.

If they rated A as more punchy than B in the pairwise test, then A was awarded a vote of 1 and B was awarded a vote of 0. As this was a forced test, A equals B was not allowed.

Playback was in a near-field setup of Genelec speakers in an ITU-R listening room. The listening level was set and measured to be 76dB(A) and each listener stated this was a general listening level that they were used to.

### **9.3 Stimuli**

The stimuli consisted of 12 excerpts of commercially available music; these are detailed in Table 21. The duration of each was 7s. All sounds were down-mixed to mono, to suppress any spatial effects. Each sound was then loudness normalized according to ITU-R BS.1770-4 (2015) standard, such that the overall loudness of the stimuli would be equal on playback. The level chosen was -23LU. The stimuli were chosen from various genres in order not to bias the test with respect to any particular arrangement or preference.

FILE / ID	ARTIST	COMMENTS	GENRE
Allegro C' Brio / 9	Beethoven	Strong Transients Sparse, Large Dynamics	Classical
Animals / 5	Nickleback	Strong Transients, Heavy Guitars, Vocal	Rock
Beatbox / 6	Roni Size	Vocal Beatboxing No Kick/Snare.	Drum & Bass
Bonfire / 2	Knife Party	Drums, Vocal Samples, Bass, Synths	Dubstep
Frozen Kingdom / 12	Weldroid	Ambient	Electronica
If / 7	Destiny's Child	Rich Vocal Harmonies, Strings, Piano, Sparse Percussion	R&B
Mad World / 4	Tears For Fears	Drums, Percussion, Synth, Bass, (Bridge)	Alternative
Pharaohs / 10	Tears For Fears	Soft Drums, Strong Piano, Vocal Sample, Synths	Alternative
Sheep May Safely Graze / 11	Bach	No percussion, No Strong Transients.	Classical
Sympathy For The Devil / 8	The Rolling Stones	Shaker, Vocals, Bass & Guitar, Percussion	Rock
The Real Slim Shady / 1	Eminem	Drums, Rap, Bass, Synth	Hip-Hop
Titanium / 3	David Guetta feat. Sia	Drums, Loud, Vocals, Synth, Pumping	Pop

**Table 21 - Stimuli Used in The Model Validation Test**

During the training phase, all listeners were asked to confirm that the stimuli were perceived to be playing back at the same loudness levels. All listeners confirmed this to be the case.

## 9.4 Subjective test results

Table 22 shows the raw ranking scores collected from the pairwise test. The table shows the number of times a particular stimulus was chosen as having more punch than another. For example, 9 listeners voted that file ID 1 had more punch than file ID 5. From this data, rank score and empirical probability scores were extracted. Using this data, a scaled response was derived using a Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959). A Matlab script OptiPt.m (Wickelmaier et al. 2004) was utilised to derive the scaled response coefficients.

For ease of interpretation, the table has been arranged in the extracted rank order of preference, i.e. file ID 1 received the most punch votes and file ID 12 received the least. File ID 1 in this case is Eminem.

ID	1	2	3	4	5	6	7	8	9	10	11	12
1	0	7	6	10	9	10	10	11	11	11	11	11
2	4	0	9	8	8	11	10	11	11	11	11	11
3	5	2	0	7	7	9	7	10	11	11	11	11
4	1	3	4	0	5	7	8	11	10	11	11	11
5	2	3	4	6	0	6	8	10	9	11	11	11
6	1	0	2	4	5	0	7	10	9	11	8	11
7	1	1	4	3	3	4	0	5	7	9	10	11
8	0	0	1	0	1	1	6	0	7	11	9	11
9	0	0	0	1	2	2	4	4	0	8	10	11
10	0	0	0	0	0	0	2	0	3	0	6	10
11	0	0	0	0	0	3	1	2	1	5	0	10
12	0	0	0	0	0	0	0	0	0	1	1	0

**Table 22 - Forced-pairwise Test Scores**

## 9.5 Rank score and between sample significance testing

Whilst it's possible to rank the stimuli with respect to the number of votes received, it's important to also establish if the ranking is statistically significant, for example, is the number of subjects that preferred file ID 1 over file ID 2 significantly different to the number of subjects that preferred file ID 2 over file ID 1. To establish this, if we assume a null hypothesis that the listeners were voting randomly, i.e. they could not establish a difference between stimuli, a sample probability threshold of 50% can be assumed. Thus, the alternative hypothesis is if a stimulus is consistently chosen as having more punch a sample probability ( $P_{sig}$ ) of >50% will be achieved. How much higher above the chance level of 50% can be established through the choice of  $z$ -score and consideration of both the population and the assumed chance population percentage (Harris & Holland, 2009). The relationship between  $P_{sig}$  and the sample probability threshold can be summarised in equation 27 as follows:

$$P_{sig} = 1.64 * \sqrt{\frac{Pu(100-Pu)}{n}} + Pu + 0.5 \quad (27)$$

where  $Pu$  = assumed sample probability threshold and  $n$  = sample size, in this case 50% and 11 respectively.

The  $z$ -score in this case was chosen as 1.64, this corresponds to a standard significance level of 5%,  $p=0.05$ , one-tailed. The resulting  $P_{sig}$  probability of 75.22% is relatively high due to the low number of subjects involved in the testing. Using this probability and comparing it against the relative votes each stimuli received, significant differences can be identified between samples. For cases where a stimulus is significantly rated as being punchier than another, the votes have been shaded dark grey in Table 22. Those shaded as light grey are significant based upon a  $p=0.10$ , one tailed  $z$ -score. Inspection of Table 22 reveals that there is a general agreement between the original extracted ranking and the significant difference probabilities. For example file ID 1 is voted as significantly different to all other samples except 2 and 3. In these cases, whilst file ID 1 has received the most votes, it cannot be stated with 95% confidence that it has more punch than file ID 2. On the other hand, we can state that file ID 12 has the least votes and statistically is the least punchy stimuli compared to all others. The goal in this analysis was to extract a punch scale for the stimuli used in the test. Given that in some cases it could be

possible to have the same level of punch perceived between some stimuli, it's possible to remove some of the stimuli utilised in the creation of the scaled output parameters. This process will be detailed in the next Section 9.6.

Whilst the pairwise data can give an indicator of rank score and also between stimuli significance, it's difficult to establish an interval scale of preference that can be used to compare to the punch model output. For example, whilst one stimuli may be ranked in 1st place, how close it is ranked to the second place stimuli? By establishing an interval based scale, it's possible to link this directly to the punch model output and run a correlation test to establish its effectiveness at punch prediction.

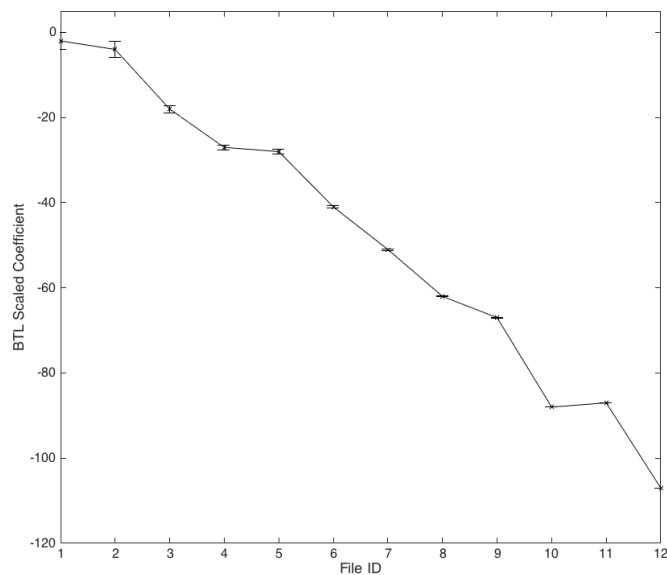
## 9.6 Bradley-Terry-Luce model

The Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) is an established approach that states, given certain testable conditions, preference probabilities may be related to scale values in the following fashion:

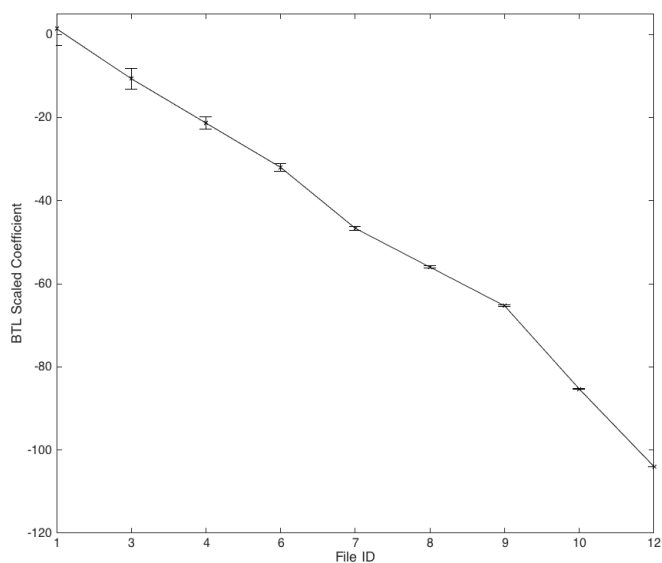
$$P_{ab} = \frac{v(a)}{v(a)+v(b)} \quad (28)$$

Where  $P_{ab}$  denotes the probability that stimuli  $a$  will be preferred over stimuli  $b$ , or on this case is punchier.  $v(a)$  and  $v(b)$  are the number of votes that each stimuli received. A Matlab script OptiPt.m (Wickelmaier et al. 2004) was utilised to derive the scaled response coefficients based on the pairwise data and the output of this is shown in Figure 78. This figure shown is based on all of the stimuli tested and shows the 95% confidence intervals calculated from the covariance matrix returned by the function. In some cases, the confidence intervals are very tight and therefore show significant agreement between listeners on particular stimuli and its relative score. These BTL scale coefficients for each stimuli correspond with extracted ranking based on probability therefore the coefficient value relates to the inter sample significance testing results as described in Section 9.5. A good example of this would be file ID 10 and 11 in that whilst they are deemed to be less punchy than all the other stimuli except 12, they not significantly different from each other, therefore with respect to ranking they lie very close to each other on the BTL scale with small intervals. Likewise, the intervals for file ID 1 and 2 are larger, thus

showing a larger variance in BTL coefficient value, this shows that there isn't 95% or more confidence that they are vastly different in terms of perceived punch.



**Figure 78 - BTL Model output based on pairwise comparison data.**



**Figure 79 - BTL Model output based on 'significant' pairwise comparison data.**

Samples that are deemed to be significantly different have no confidence interval overlap. Where there is overlap, these samples are given roughly the same score by the BTL model as such, it's possible to remove these stimuli from the model fit. Figure 79 shows the model output after the removal of file IDs 2, 5 and 11.

File ID 11 was also removed as it was seemed the same as File ID 10 in terms of the significance testing and in term of BTL coefficient score, was also very close to that of File ID 10. This derives an interval scale that can be tested against the objective punch model.

The testable conditions for a BTL model fit are those of transitivity and goodness of fit. The former can be established by looking at the raw data scores in Table 22.

ID	1	2	3	4	5	6	7	8	9	10	11	12
1	0	7	6	10	9	10	10	11	11	11	11	11
2	4	0	9	8	8	11	10	11	11	11	11	11
3	5	2	0	7	7	9	7	10	11	11	11	11
4	1	3	4	0	5	7	8	11	10	11	11	11
5	2	3	4	6	0	6	8	10	9	11	11	11
6	1	0	2	4	5	0	7	10	9	11	8	11
7	1	1	4	3	3	4	0	5	7	9	10	11
8	0	0	1	0	1	1	6	0	7	11	9	11
9	0	0	0	1	2	2	4	4	0	8	10	11
10	0	0	0	0	0	0	2	0	3	0	6	10
11	0	0	0	0	0	3	1	2	1	5	0	10
12	0	0	0	0	0	0	0	0	0	1	1	0

The general rule of transitivity is described as:

$$\text{If } A \geq B \text{ and } B \geq C \text{ then } A \geq C \quad (29)$$

Examination of the raw data shows that transitivity is not violated when considering the total votes received by each stimulus. The goodness of fit statistic returned by OptiPt.m as a  $\chi^2$  (chi squared) statistic was within bounds of the lower and upper critical values of the chi square distribution,  $p=0.05$ , therefore the BTL model can account for the data.

## 9.7 Model output correlation analysis

As described earlier, different punch model outputs were analysed to establish relative correlation performance with the subjective BTL model output (i.e. the perceptual punch scale). SPSS was utilised to do this analysis. Through an approach detailed in Section 8.7, the statistical output of the punch model was utilised and a punch range measure was investigated. In the statistical punch model, upper and lower bounds of the punch frame distribution form the punch range, similar to that found in the loudness range measure, albeit without any gating mechanism. The lower bound percentile point chosen was the 1<sup>st</sup> percentile, the upper bound was determined by comparison of various values with respect to the relative correlation score achieved.

The upper bound point that resulted in the highest correlation score was the 95<sup>th</sup> percentile measure (PM95). In addition, a measure also showing a high degree of correlation was the 95<sup>th</sup> percentile / mean measure (PM95M). The latter could be considered to be a ratio derivation as found in dynamic range type calculations. Other computational measures were compared against the punch scale data to compare performance in mapping to the punch attribute. As each of the measures chosen offer differing units of measurement (SI), for example, dB, LU or ratio score, correlation analysis was utilised in order to disregard these differences and see how well each mapped to the perceptual punch scale output parameters. For each of the tested objective measures, Table 23 shows the Pearson correlation coefficient ( $r$ ) and the rank correlation (Spearman's  $\rho$ ) between each measure and the punch scale.

If the underlying distribution of the two variables being considered is not a bivariate normal distribution, or if there is a non-linear relationship between the variables, the Pearson correlation ( $r$ ) might not provide an accurate measure of their association. The non-parametric rank correlation ( $\rho$ ) is more robust in that it considers rank order, thus outliers are generally not going to affect the overall correlation statistic. Ideally, both coefficients would be close to each other.

Measure	Corr( $r$ )	Corr( $\rho$ )
CF	-.351	-.483
PLR	.010	-.083
LDR	.442	.333
IBR_diff	.706*	.650
PM95	.849**	.833**
PM95M	.770*	-.750*

**Table 23 - Correlations of the tested measures with the perceptual punch scale**

In addition, and in order to show possible correlations between measures, a correlation matrix as shown in Figure 80, was produced.

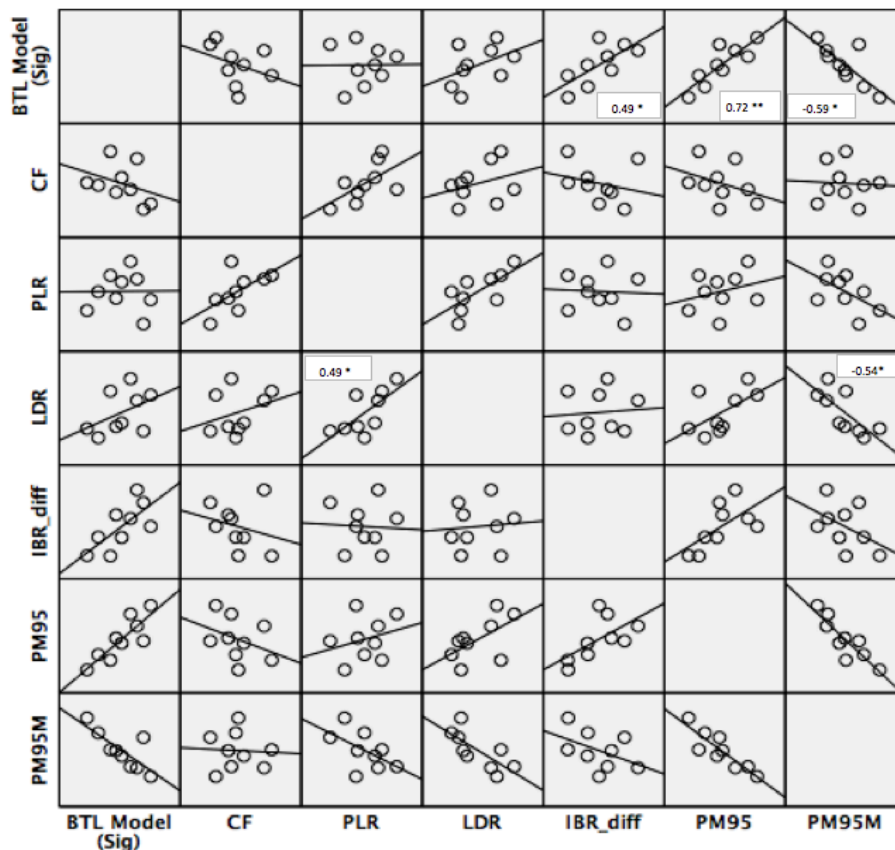


Figure 80 - Correlation matrix of all tested measures

This matrix includes  $r^2$  values where significance is prevalent. In Table 23 and Figure 80, \* and \*\* signifies correlation is significant at the 0.05 and 0.01 level (2-tailed) respectively.

From the results obtained, the PM95 measure showed a ‘very strong’ positive correlation with punch perception. Both  $r$  and  $\rho$  coefficients (0.849 and 0.833) being significant at the 0.01 level (2-tailed). The PM95M measure, which is the PM95 measure divided by the mean value of punch frames also correlated very well with the perceptual punch scale.

The PLR and CF measures showed the least correlation with punch perception. One might assume that a reduced CF or PLR may correlate well with punch due to the use of compression

in music production attempting to maximise loudness and possibly punch in the process. In the stimuli tested, with the loudness having been normalised between stimuli, punch perception did not correlate with the measures. Only the CF measure showed a ‘weak’ negative correlation with an  $r$  coefficient of -.351. This reinforces the conclusion stated in Chapter 6 that the use of compression from a stem perspective, e.g. on a kick or snare drum with a view of increasing punch is more likely to correlate to the temporal envelope modification that occurs as part of the process rather than changes in dynamic range.

The IBR\_diff measure showed a ‘strong’ correlation with punch perception, albeit less than the PM95 and PM95M measures. An  $r$  coefficient of .706 was observed with a p-value of 0.034 (2 tailed). This measure is based upon the relationship between dynamic ranges measured across frequency bands. Causes of correlation between bands could be caused by application of compression during mastering, noise like stimuli or stimuli with no percussive based content. Higher IBR\_diff values corresponded with higher perception of punch in stimuli tested.

The LDR measure, a measure of loudness dynamic range, is proposed as a measure of microdynamics (Skovenborg, 2014). Whilst it isn’t a measure of how the underlying dynamic content is perceived it was included to ascertain any correlation with the punch attribute. An  $r$  coefficient of .442 indicates ‘moderate’ correlation strength. This correlation could be an indicator of the ‘dynamic’ content contained within the stimuli which is in turn linked to the punch level. Stimuli with little dynamic content are likely to have a low punch score, however, even with high dynamic content (in which case LDR would return a high level) this may not always result in high levels of punch, an example of this would be a series of hi-hat hits.

It can be seen in Figure 80 that the LDR has a ‘strong’ correlation with PLR, with an  $r^2$  coefficient of 0.49 significance being less than the 0.05 level (2-tailed). This similarity can be explained due to the method employed in the LDR algorithm. The measure is based on deriving the maximum difference between a ‘fast’ and a ‘slow’ loudness levels. The peak utilised in the PLR measure may roughly correspond to the ‘fast’ loudness level calculated, whilst the average loudness will be that of the ‘slow’ loudness integration employed. If for example, the ‘fast’ integration window were made to be 1 sample in length, it’s likely that the LDR/PLR correlation coefficient would approach 1.

Both punch output models and the LDR measure implement an integrative process in the form of windowing, the LDR measure tested uses 3s ‘slow’ and 50ms ‘fast’ windows whilst the punch model employs 100ms in its frame calculations. With reference to the LDR algorithm, one might expect larger LDR values as a result of a decrease in the length of the ‘fast’ window size, particularly in perceptually dynamic material. Indeed, this was the case with the stimuli tested when compared to the use of a 100ms window size in the LDR model.

The LDR algorithm is also based upon the ITU-R BS.1770-4 (2015) model of loudness measurement and given that it is a maximal difference type measure, it also has parallels with the PM95M measurement. The PM95M measurement utilises the 95% percentile punch frame level along with the ‘mean’ of the punch frames to formulate its output. One could say this is equivalent to somewhat of a ‘peak’ to ‘rms’ punch frame measure (or ‘fast’ to ‘slow’ ratio). The PM95M model employs frequency weighting in its algorithm (see Chapter 8) unlike the LDR that utilises the K-weighting of the loudness model only.

In general, higher values of LDR did correspond with higher levels of punch perception. As an indicator of ‘microdynamics’ within stimuli, one might expect this to be the case, for example if drums or percussion are present or not. This was certainly the case with the stimuli tested, whereby ‘Beatbox by Roni Size’ was measured with the highest LDR value. This particular stimulus was noted by the listeners as being the most dynamic. The PM95 and PM95M models on the other hand showed a stronger correlation to the punch perceived in the stimuli tested than the LDR measure, this may be due to the combination of both onset detection and frequency band weighting employed.

## 9.8 Model validation conclusion

The punch model outlined in Chapter 8 was evaluated against subjective scores obtained through a forced pairwise comparison test. 12 stimuli were utilised and 11 expert listeners took part in the test. Each listener made a total of 66 comparisons.

For comparison four additional types of objective measures relating to signal dynamics and loudness perception were evaluated against the same derived punch scale. The PLR and CF measures showed the least correlation whilst the PM95 and PM95M model outputs showed the best correlations.

From the results obtained, the PM95 measure can be considered to have a ‘very strong’ positive correlation to the punch scale.

The IBR\_diff measure showed a ‘strong’ correlation with punch perception, albeit less than the PM95 and PM95M measures.

The LDR measure correlated well with dynamic levels perceived in the stimuli and correlation to the punch attribute was ‘moderate’. This yields the possibility of using a combination of the LDR and PM95 models to give both an indication of underlying dynamics and the ‘*punchiness*’ of those dynamics.

The model proposed yields the possibility to perceptually weight the transient components of an audio signal. In doing so, output relative to the perception of punch in the signal is possible. The model could be of use in both mixing and mastering as well as audio transcription and retrieval.

## Chapter 10 Conclusions and future work

This aim of the thesis was to explore new measures with respect to audio dynamics and develop an objective model that predicts punch in musical signals. The literature review revealed that whilst there has been a great deal of work in the area of semantic description low-level analysis with respect to the perception of punch remained largely unexplored.

In light of this, the work was iterative and a number of methods were investigated which included multi-resolution signal separation, analysis of statistical model outputs, sub-band filtering and detailed low-level feature extraction. In total, four new measures have been proposed within this body of work along with a formalisation of the punch attribute with respect to low-level features of the audio under test.

This final chapter will highlight the conclusions drawn from this work and the measurements proposed. The main findings based on experimentation with each measure are shown as bullet points. Finally, possible future work will be discussed.

### 10.1 Main research findings

- The punch model (PM95) presented offers the ability to measure a perceptual parameter that was previously only able to be described subjectively by listeners. It shows a very strong correlation to the perceptual attribute.
- The attack onsets of the audio across all octave bands affects the punch perceived by the listener. However; greater weight is evident in the lower octaves and there is little variance between 0ms and 5ms onset times.
- The total energy summation of the onsets, across the 2nd, 3rd, 4th, 5th and 6th octave bands, shows a very strong correlation to the punch perceived by the listener.
- The current loudness model K-filter should be re-visited to address anomalies in its weightings around and above the 1kHz range.

- Additional parameters identified showing a degree of correlation with the punch attribute when taken in isolation were spectral skew, spectral spread and rhythm strength.
- With respect to dynamic range compression, no singular control setting is responsible for punch modification. The important factor lies not in the process involved in audio modification but rather the final signal and low-level features that result.
- The Inter-Band-Ratio (IBR) measure presented shows a stronger correlation to perception of audio quality than existing dynamic range measures. In addition, the statistical output of this measure is shown have a moderate correlation to the perception of punch as graded by a panel of expert listeners.
- A reduction in overall dynamic range of a piece of audio does not necessarily result in a perception of low overall audio quality by the listener when compared to the uncompressed version. Rather, the relationship between the dynamic ranges across frequency bands has been shown to correlate to this score. As such, the proposed IBR measure is more effective than the CF measure when used to assess audio quality with respect to audio dynamics.
- Transient content and dynamic range de-correlation between frequency bands relate to higher subjective scores being given by the listeners with respect to grading of punch and clarity.
- The statistical IBR output, both in terms of percentile and histogram representation is an improvement on the integrative-based method. It affords more insight into the underlying dynamic contour of the sample under test than current dynamic range based measures and shows a moderate correlation to the punch attribute.

- The Transient to Steady-state Ratio (TSR) and Transient to Steady-state Ratio+Residual (TSR+R) measurements presented can be utilised to indicate the perceptual dynamics and possible masking artefacts present within a piece of audio.
- The separation of an audio signal into its transient, steady-state and residual components enables new measures that are perceptually relevant to punch perception. For example, TSR and TSR+R offer the possibility of a ‘rhythm to background’ measurement.
- For a track, where the sound sources had been mixed effectively with minimal masking, there should be good transient intensity which will result in a high TSR being achieved.
- The method of separation explored enables the possibility to perceptually weight the transient and steady state frequency bands.
- The TSR+R measure, by incorporating the residual components in the ratio calculation, can be used to indicate the presence of noise in a track. It could therefore be used to indicate perceived clarity or quality differences.

In formalising the definition of punch the following is proposed:

- *Punch can be described as a short period of significant change in power in a piece of music or performance. The magnitude of change is associated and proportional to the signal dynamics that are present and thus, productions that do not possess any transient or dynamic attributes cannot possess punch. The onset of the transient present across octave bands affects the listener perception of punch, with the lowest octave attributing the most punch as the onset is decreased and vice-versa. Punch is therefore related to transient change and the energy density (summation across frequency bands) occurring at a moment in time and duration.*

## 10.2 Further work

Further work has been identified which may prove beneficial should this work be continued. This is summarised as follows:

- *In order to investigate the effect on both processing speed and possible improvement in correlation to listener perception, further testing of the PM95 punch model could be applied with differing window sizes and percentile parameters*
- *Temporal weightings could be applied in the PM95 model. Whilst these were applied to the noise burst stimuli, signal duration of the detected onsets wasn't incorporated into the P95M model that has been described.*
- *Additional noise burst testing would be beneficial to explore the perception of pink noise burst above the 1kHz octave band. These bands are currently omitted in the model. Indeed, exploration into noise burst perception would be beneficial to the wider research community particularly in relation to the current ITU-R BS.1770-4 (2015) loudness model.*
- *The inclusion of a 'steady state' component into the PM95 model would enable the transient to steady state ratio relevance to be evaluated. Currently, poorly mixed audio may still give the same objective punch output as well mixed audio as the transient to steady state ratios are ignored.*
- *The TSR and TSR+R measures could be tested with respect to establish if they correlate with listener perceived clarity amongst other quality based features.*
- *The IBR measure could be extended to include more perceptually relevant frequency bands*
- *A real-time implementation of the punch model (and other measures outlined in this thesis) could be possible.*



## References

Aarts, R.M. (1992). A Comparison of Some Loudness Measures for Loudspeaker Listening Tests. In *JAES*, vol.40:3, pp.142-146.

AES (2006) 6id: AES information document for digital audio – Personal computer audio quality measurements, Audio Engineering Society.

AES Tech Doc 1004.1.15-10, Recommendations for Loudness of Audio Streaming and Network File Playback, October 2015.

ANSI (1994) *American National Standard: Acoustical Terminology*, ANSI S1.1-1994, New York: Acoustical Society of America / American National Standards Institute.

Avendano, C. & Goodwin, M. (2004). Enhancement of Audio Signals Based on Modulation Spectrum Processing. In, *AES Convention Paper 6259*.

Ballou, G. (2005). Handbook for sound engineers, 3<sup>rd</sup> Edition, Gulf Professional Publishing.

Barry, D., Fitzgerald, D., Coyle, E. & Lawlor, B. (2005). Single Channel Source Separation using Short-time Independent Component Analysis. In *AES 119th conference*.

Bech, S. & Zacharov, N. (2006). *Perceptual audio evaluation, theory, method and application*, J.Wiley, Chichester.

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M, & Sandler, M. (2005). A Tutorial on Onset Detection in Music Signals. In *IEEE Transactions on Speech and Audio Processing*. 13, 1035–1047.

Boley, J., Lester, M., & Danner, C. (2010). Measuring Dynamics: Comparing and Contrasting Algorithms for the Computation of Dynamic Range. In *Audio Engineering Society 129th Convention, Paper 8178, Audio Engineering Society*, San Francisco, CA, USA

Bradley, R.A & Terry, M.E, (1952). Rank Analysis of incomplete block designs, The method of pair comparisons. In *Biometrika*, 324-345

Bullock, J. (2007). Libxtract: A lightweight library for audio feature extraction. In *Proceedings of the International Computer Music Conference*, volume 43.

Cannam, C., Landone, C & Sandler, M. (2010). Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files. In *Proceedings of the ACM Multimedia International Conference*.

Casey, M.A., Veltkamp, R. Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. In *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696

Collins, N. (2005). A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychacoustically Motivated Detection Functions. In *AES Convention Paper 6363*.

Croghan, N. B. H., Arehart K.H, & Kates J.M. (2012). Quality and loudness judgments for music subjected to compression limiting. In *JASA.*, vol. 132, no. 2, pp. 1177–1188.

Daudet, L. (2005). A Review Of Techniques For The Extraction Of Transients in Musical Signals. In *Computer Music Modelling and Retrieval Conference*, Italy.

De Man, B., Leonard, B., King, R. & Reiss, J. (2014). An analysis and evaluation of audio features for multitrack music mixtures. In *15th International Society for Music Information Retrieval Conference (ISMIR)*, October.

Deruty, E. & Tardieu, D. (2014). About Dynamic Processing in Mainstream Music. In *JAES Volume 62*, Issue 1/2, pp.42-45.

Driedger, J., Muller, M. & Disch, S. (2014). Extending Harmonic-Percussive Separation Of Audio Signals. In *ISMIR*.

EBU Tech Doc 3342, Loudness Range: A descriptor to supplement loudness normalisation in accordance with EBU R 128, August.

Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing.

Every, M. (2008). Discriminating Between Pitched Sources. In *Music Audio, IEEE Transactions*.

Feaster, P (2008). Edouard-Leon Scott de Martinville's – Principes De Phonoautographie (1857), Retrieved from <http://Firstsounds.org>

Fenton, S & Lee, H. (2015). Towards a Perceptual Model Of 'Punch' In Musical Signals. In *Audio Engineering Society Convention 139*, New York, USA.

Fenton, S. & Wakefield, J. (2012). Objective profiling of perceived punch and clarity in produced music. In *Audio Engineering Society Convention 132*, Berlin.

Fenton, S., Fazenda, B., & Wakefield, J. (2009). Objective quality measurement of audio using multiband dynamic range analysis. In *Institute of Acoustics Conference*, Brighton.

Fenton, S., Fazenda, B., & Wakefield, J. (2011). Objective Measurement Of Music Quality using Inter-Band Relationship Analysis. In *Audio Engineering Society Convention 130*, London.

Fenton, S., Lee, H. & Wakefield, J. (2015). Hybrid Multiresolution Analysis Of Punch In Musical Signals. In *Audio Engineering Society Convention 138*, Warsaw.

Ferguson, S., Cabrera, D. & Schubert, E. (2010). Comparing continuous subjective loudness responses and computational models of loudness for temporally varying sounds, In *Audio Engineering Society Convention 129*, San Francisco, CA, USA.

Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. in *Proc. of the DAFx-10*, Graz, Austria, Sept.

Fletcher, H. and Munson, W.A. (1933). Loudness, its definition, measurement and calculation, In *Journal of the Acoustic Society of America* 5, 82-108

Freed, D.J. (1990). ‘Auditory Correlates of Perceived Mallet Hardness For a Set of Recorded Percussive Sound Events’. In *J.Acoustical Society of America*, Am.87.

Glasberg, B.R. & Moore, B.C.J. (2002). A Model of Loudness Applicable to Time-Varying Sounds. In *Journal of the Audio Engineering Society*, vol.50:5, pp.331-342.

Gonzalez, E. & Reiss, J.D. (2009). Automatic equalization of multi-channel audio using cross-adaptive methods. In *Proceedings of the 127th AES Convention*, New York.

Goodwin, M. & Avendano, C. (2004). Enhancement of Audio Signals Using Transient Detection and Modification. In *AES Convention Paper 6255*.

Gordon, J. W. (1987). The perceptual attack time of musical tones. In *Journal of the Acoustic Society of America*, 82(1), 88–105

Grey, J.M. (1977). Multidimensional Perceptual Scaling of Musical Timbres. In *JAES*, vol 61.

Grey, J.M. & Gordon, J.W. (1978). Perceptual effects of spectral modifications on musical timbres. In *The Journal of the Acoustical Society of America* . Vol 63 (1493)

Gribben, C. & Lee, H. (2015). Hulti-Gen - Huddersfield Universal Listening Test Interface Generator. Retrieved from <http://www.hud.ac.uk/research/researchcentres/mtp rg/projects/apl/>

Hainsworth, S. & Macleod, M. (2003). Onset detection in musical audio signals. In *Proc. Int. Computer Music Conference*, pages 163– 6.

Harris, L.E & Holland, K.R (2009). Using statistics to analyse listening test data: some sources and advice for non-statisticians. In *Proceedings of the Institute of Acoustics*.

Hartmann, W.M. (1998). Signals, Sound, and Sensation. New York: Springer.  
IEC 60651, Sound level meters, International Electrotechnical Commission, 1979.

International Standards Organisation. (2016). *ISO Standards*. Retrieved from [http://www.iso.org/iso/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=63077](http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=63077).

Iragaray, I & Biscainho, L.W.P. (2013). Transient and Steady State Component Extraction Using Non-Linear Filtering. In *Congreso Internacional de Ciencia y Tecnología Musical – CICTeM*.

ISO 226, Acoustics - Normal Equal- Loudness Level Contours, Geneva: International Organization for Standardization, 1987

ISO 532, Acoustics. Method for calculating loudness level. International Standard, International Organisation for Standardisation, 1975.

ITU-R. (1986) 468-4: Measurement of audio-frequency noise voltage level in sound broadcasting, International Telecommunications Union, Geneva, Switzerland.

ITU-R. (1994) BS.1116: Methods for the subjective assessment of audio systems including multichannel sound systems. Geneva, Switzerland

ITU-R. (1998) BS.1387-1: Method for objective measurement of perceived audio quality. Geneva, Switzerland

ITU-R. (2003) BS.1284-1: General methods for the subjective assessment of sound quality International Telecommunications Union, Geneva, Switzerland

ITU-R. (2003) BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems. Geneva, Switzerland

ITU-R. BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level, International Telecommunications Union, Geneva, Switzerland, 2015

ITU-T P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications, International Telecommunications Union, Geneva, Switzerland, 2004

ITU-T. (1990) P800: Methods for subjective determination of transmission quality, Geneva, Switzerland

Katz, B. (2002). *Mastering Audio: The Art and the Science*, Oxford: Focal Press.

Klapuri, A. (1999). Sound Onset Detection by Applying Psychoacoustic Knowledge. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 115–118.

Koelsch, S. & Siebel, W. A. (2005). Towards a neural basis of music perception. In *Trends in Cognitive Sciences*. 9 (12): 578–84.

Lagrange, M., Raspaud, M., Badeau, R., & Richard, G. (2009). Explicit Modeling of Temporal Dynamics within Musical Signals for Acoustical Unit Similarity.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. In *Perceptual Psychophysics* 26, p1426.

Lartillot, O. & Toivianen, P. (2007). MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio. International Conference on Music Information Retrieval, Vienna.

Laurenti, N., De Poli, G., & Montagner, D. (2007). A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds. In *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb.

Learned, R. & Willsky, A. (1995). A Wavelet Packet Approach To Transient Signal Classification. In *Applied and Computational Harmonic Analysis* 2, pp. 265-278.

Levitin, D.J (2006). This is Your Brain on Music : The Science of Human Obsession. Penguin Group, 375 Hudson Street, New York, New York.

Ljudtekniska, S (2013). “MasVis - Allpassed crest factor,” Retrieved September 10<sup>th</sup>, 2013, from: <http://www.lts.a.se/lts/manual>.

Loudness Wars. (n.d.). In *Wikipedia*. Retrieved September 24th, 2009, from [http://en.wikipedia.org/wiki/Loudness\\_war](http://en.wikipedia.org/wiki/Loudness_war)

Lu, L., Liu, D., Zhang, H.J. (2006). Automatic Mood Detection and Tracking of Music Audio Signals. In *IEEE Transactions on Audio Speech and Language Processing* 14.

- Lund, T. (2011). ITU-R BS.1770 Revisited. In NAB Engineering Conference, 2011.
- Maempel, H.J. & Gawlik, F. (2009). The influence of sound processing on listeners' program choice in radio broadcasting. In *Audio Engineering Society Convention 126*, Munich, Germany (May 7–9), Paper 7785.
- Marston, D & Mason, A. (1994). Cascaded audio coding In *EBU Technical Review 304*, Geneva, Switzerland
- MInterface (2016). Experimental Design. Retrieved from <https://iosr.uk/projects/quality/design.php>.
- Moore, B. (2004). An Introduction To The Psychology of Hearing, pp.138-145, 5th Edition, Elsevier.
- Moore, B. C. J., Glasberg, B. R., and Baer, T. (1997). A model for the prediction of thresholds, loudness and partial loudness. In *JAES*. 45, 224-240.
- Moore, B.C.J., Glasberg, B.R. & Stone, M.A. (2003). Why Are Commercials so Loud? Perception and Modeling of the Loudness of Amplitude-Compressed Speech, In *JAES*, vol.51:12, 1123-1132.
- MPEG 7 – ISO/IEC. (2001) 15938: Information Technology – Multimedia Content Description Interface - Part 4 – Audio
- Nielsen, S.H. & Lund, T. (2000). 0dBFS+ Levels in Digital Mastering, In *Audio Engineering Society Convention 109*.
- Ono, N., Miyamoto, K., Le Roux, J., Kameoka, H., & Sagayama, S. (2008). Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In *Proc. of the EU- SIPCO 2008*, Lausanne, Switzerland.

Oriol, R.P, Hector, P.R, Dara, D., Hiroshi, T., Wataru, H., Koji, O. & Xavier, S. (2015). A Real-Time System for Measuring Sound Goodness in Instrumental Sounds. In *Audio Engineering Society Convention 138, Convention paper 9350*, Audio Engineering Society, Berlin, Germany.

Pederson, T.H, & Zacharov, N., (2015). The development of a Sound Wheel for Reproduced Sound. In, *Audio Engineering Society Convention 138*, Paper 9310, Warsaw.

Peeters, G. (2004). A Large Set of Audio Features for Sound Description (similarity and classification) in the CUIDADO project. Retrieved from <http://www.ircam.fr/>.

Pestana, P., Reiss, J. & Barbosa, A. (2013). Loudness Measurement of Multitrack Audio Content using Modifications of ITU-R BS. In *Audio Engineering Society Convention 134, Rome*.

R.D Luce, (1959). Individual Choice Behavior. New York, Wiley.

Rasch, J.V. & Rasch, R.A. (1981). The Perceptual Onset of Musical Tone. In, *Perception and Psychophys*, vol 29.

Robinson, D.W, & Dadson, M.A. (1956). A redetermination of the equal-loudness relations for pure tones. In *British Journal of Applied Physics*, Vol 7, P166-181

Ronan, M., Sazdov, R., & Ward, N. (2014). Factors influencing listener preference for dynamic range compression. In *Audio Engineering Society Convention 137, Los Angeles, USA*

Salomons, E, M. & Janssen, S, A. (2011). Practical Ranges of Loudness Levels of Various Types of Environmental Noise, Including Traffic Noise, Aircraft Noise, and Industrial Noise. In *Int. J. Environ. Res. Public Health* , 8(6), 1847-1864

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: a survey. In *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 133–141

Scheirer, E. & Slaney, M.(1997). Construction and evaluation of a robust multi-feature speech/music discriminator. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1331–1334, Munich, Germany.

Scheuren, J. (2014). ISO 532 – Living and working with alternative loudness standards, In *Inter.noise Conference*, Melbourne, Australia.

Schroeder, M., Atal, B. & Hall, J. (1979). Optimizing Digital Speech Codecs by Exploiting Masking Properties of the Human Ear. In *JASA*, vol. 66, 1647-1652.

Skovenborg, E. (2012). Loudness Range (LRA) – Design and Evaluation. In *AES Convention paper 8616*.

Skovenborg, E. & Lund, T. (2008). Loudness descriptors to characterize programs and music tracks. In *AES Convention paper 7514*.

Skovenborg, E. & Nielsen, S.H. (2004). Evaluation of Different Loudness Models with Music and Speech Material, In *Audio Engineering Society Convention 117*, San Francisco, CA, USA.

Soulodre, G.A. (2004). Evaluation of Objective Loudness Meters, In *Audio Engineering Society Convention 116*, Berlin, 6161.

Soulodre, G.A. & Norcross, S.G. (2003). Objective Measures of Loudness, In *Audio Engineering Society Convention 115*.

SQAM (2005), Test CD, Sound Quality Assessment Material, Recordings for subjective tests – Cat. No. 422 204-2, EBU 1988.

Stables, R., Enderby, S., De Man, B., Fazekas, G., & Reiss, J.D. (2014). SAFE: A system for the extraction and retrieval of semantic audio descriptors. In *The International Society for Music Information Retrieval (ISMIR)*.

Stepanek, J. (2006). Musical Sound Timbre: Verbal Descriptions and Dimension', DAFX Conference.

Stevens, S.S. (1957). On the psychophysical law, In *Psychol.Rev.*, vol.64, pp.153-181.

Stoll, G. & Kozamernik, F. (2000). EBU listening tests on internet audio codecs. In *EBU Technical Review*, Geneva, Switzerland

Taylor, R.W. & Martens, M.L. (2014). Hyper- Compression in Music Production: Listener Preferences on Dynamic Range Reduction. In *Audio Engineering Society Convention 136*, 2014

Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J., Colomes, C., Keyhl, M., Stoll, G., Brandenburg, K. & Feiten, B. (2000). PEAQ – The ITU Standard for Objective Measurement of Perceived Audio Quality. In *JAES*, Vol. 48, No. 1/2, January/February.

Tollerton, R. (2008). In *pfpf: An Experimental Estimator of Dynamic Range in Music*, Retrieved January 2010, from [http://audiamorous.blogspot.co.uk/2008/01/pfpf-experimental-estimator-of-dynamic\\_13.html](http://audiamorous.blogspot.co.uk/2008/01/pfpf-experimental-estimator-of-dynamic_13.html)

Toole F. E. (1985). Subjective Measurements of Loudspeaker Sound Quality and Listener Performance. In *JAES*, Vol.33.

TT Dynamics Meter (n.d). In *Pleasurize Music Foundation*. Retrieved September 10<sup>th</sup>, 2011, from <http://www.pleasurizemusic.com>

TurnMeUp!. (2009). *Turn Me Up! Bringing dynamics back to music*. Retrieved from <http://www.turnmeup.org>.

Tzanetakis, G. (2012). *Music Data Mining, Audio Feature Extraction*. CRC Press.

Tzanetakis, G. & Cook, P. (1999). MARSYAS : a framework for audio analysis. In *Organized Sound*, vol. 4, no. 3.

Verhey, J. & Kollmeier, B. (2002). Spectral Loudness Summation as a Function of Duration. In *JASA, Vol. 111, No 3*.

Vickers, E. (2001). Automatic Long-term Loudness and Dynamics Matching, In *Audio Engineering Society Convention 111*.

Vickers, E. (2011). The loudness war: Do louder, hypercompressed recordings sell better?, In *JAES. 59, 346–351*.

Vincent, E., (2005). MUSHRAM 1.0, Centre for Digital Music, Queens Mary, University of London, November 2005.

VisLM (2012). In NuGen Audio. Retrieved February 2012, from [http://www.nugenaudio.com/visLM\\_loudness-meter\\_VST\\_AU\\_RTAS.php](http://www.nugenaudio.com/visLM_loudness-meter_VST_AU_RTAS.php)

Vos, J. & Rasch, R. (1981). The Perceptual Onset of Musical Tones. In *Attention, Perception, and Psychophysics* 29(4):323–335.

Walsh, M., Stein, E. & Jot, J.M. (2011). Adaptive Dynamics Enhancement. In *AES Convention Paper 8343*.

Wang, E. & Tan, B.T.G. (2008). Application of Wavelets to Onset Transients and Inharmonicity of Piano Tones. In *JAES, Vol 56, No.5*.

Watson, C. & Gengel, R. (1969). Signal Duration and Signal Frequency in Relation to Auditory Sensitivity. In *JASA*, Vol. 46, No 4 (Part 2).

Waves (2008). In Waves. Retrieved February 2008, from <http://www.waves.com>

Wickelmaier, F. & Schmid, C. (2004). A Matlab Function to Estimate Choice-model parameters from Pairwise-comparison data. In *Behaviour Research Methods, Instruments & Computers*.

Wilson, A., & Fazenda, B. (2013). Perception & Evaluation of Audio Quality in Music Production. In *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*

Wilson, T, R., Fenton, S. & Stephenson, M. (2015). A Semantically Motivated Gestural Interface for the Control of a Dynamic Range Compressor. In *Audio Engineering Society Convention 138*, 7th-10th May 2015, Warsaw, Poland.

Zaunschirm, M., Reiss, J. & Klapuri, A. (2012). A High Quality Sub-Band Approach to Musical Transient Modification. In, *Computer Music Journal*, Volume 36, Number 2, pp. 23-36

Zielinski, S. & Rumsey, F. (2008). On some biases encountered in modern audio quality listening tests- A Review. In *JAES*, Vol 56, No 6.

Zwicker, E. (1960). Ein Verfahren zur Berechnung der Lautstärke (A procedure for calculating loudness), In *Acustica*, vol.10, pp.304-308.

Zwicker, E. & Fastl, H. (1999). *Psychoacoustics: Facts and Models*". Berlin, Springer.