Fenton, Steven, Lee, Hyunkook and Wakefield, Jonathan P.

Evaluation of a Perceptually-Based Model of "Punch" with Music Material

Audio Engineering Society

Convention Paper

Presented at the 141st Convention
2016 September 29–October 2 Los Angeles, USA

Evaluation of a Perceptually Based Model Of 'Punch' with Music Material

Steven Fenton, Hyunkook Lee and Jonathan Wakefield

School of Computing &Engineering, The University of Huddersfield, Huddersfield, UK.

Correspondence should be addressed to Steven Fenton (s.m.fenton@hud.ac.uk)

ABSTRACT

This paper evaluates a perceptually motivated objective model for the measurement of 'punch' in musical signals. Punch is a perceptual attribute that is often used to characterise music that conveys a sense of dynamic power or weight. A methodology is employed that combines signal separation, onset detection and low level parameter measurement to produce a perceptually weighted 'punch' score. The model is evaluated against subjective scores derived through a forced pairwise comparison listening test using a wide variety of musical stimuli. The model output indicates a high degree of correlation with the subjective scores. Correlation results are also compared to other objective models such as Crest Factor, Inter-Band-Ratio (IBR), Peak-to-Loudness Ratio (PLR) and Loudness Dynamic Range (LDR).

1 Introduction

Music classification and information retrieval (MIR) is an area that benefits from the extraction of low level features to determine such things as, but not limited to, genre, BPM and musical key. Different approaches to obtain the features are utilized, some of which involve time and frequency domain transforms to achieve this. In the multimedia content description standard (MPEG-7) a number of audio based descriptors have been defined and their underlying algorithms have been formalised. This standard, and associated tools developed from it, allow measurement and categorisation of audio content which allows fast and efficient searching for material that is of interest to the user

Automatic loudness normalisation is being used by broadcasters and 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 correctly use this normalisation 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 artist desire of releasing music that possesses both dynamic quality and spaciousness, all of which can be somewhat destroyed through 'target' driven mastering. Additional metrics, or low level parameters, available to both artist and engineers would be beneficial in highlighting perceptual attributes affected, or not, during this process.

A characteristic that is related to dynamics is known as 'punch'. Punch can be defined as a short period of significant change in power in a piece of music or performance [1]. In essence, productions that do not possess any transient information cannot possess punch. Thus, punch is both related to transient change and the energy density at a particular moment in time and duration. Further to the above hypothesis, dynamic change in particular frequency bands contribute to the overall perception of punch perceived by the listener and the overall average loudness level inherently affects this at that time [2].

Previous work by the authors [3] explored and proposed a perceptually based model that combined signal decomposition and low level parameter measurement to produce a perceptually weighted 'punch' score. Within that publication, links between the stimuli under test and the perception of the punch attribute based on the changes made to the stimuli were explored. The reasoning behind this was that subjects might differ in their choice of low-level attribute when describing their perception of 'punch'. Even if they agree on the attribute, they may give a punch rating based on different ranges of values within that attribute.

The model output indicated a possible correlation with the perceptual sensation of punch but a formal listening test and evaluation was not undertaken to evaluate the various model outputs.

This work begins by briefly outlining the model and its outputs. Following this a controlled listening test is employed to elicit subjective perceptual punch scores for a wide selection of musical stimuli. The goal was to quantify the perceived punch of each stimulus as an attribute of the sound itself. The use of differing genres was to test the validity of the output regardless of genre. The experiment did not aim to establish a good or bad level of punch. The validation test involved various stages, these were 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.

As punch could be considered as both a momentary and/or long-term perceptual attribute, statistical metrics are also explored. These can be useful in quantifying an overall song attribute.

2 Model Outline

The model proposed by Fenton and Lee [3] consists of a signal separation stage, weighting stage and onset detection mechanism. Its outline can be seen in Figure 3. The model is perceptually weighted through the use of coefficients derived through the analysis of subjective perception relating to pink noise stimuli having differing onset time and octave band centre frequencies.

The audio signal is first separated into transient, steady state and residual components. The transient components are then fed through a spectral weighting filter. Measurements are then made across the octave bands of both signal energy and onset time. Onsets and transient durations are extracted in the temporal data stage. This data is utilised to control the weighting factors of each frequency band before the final punch output is summed together.

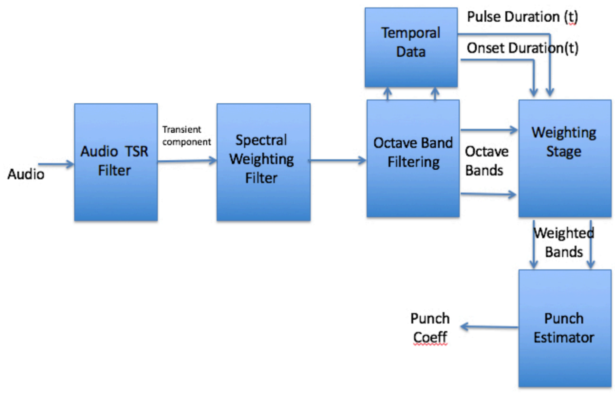

Figure 1. Block diagram of the punch-prediction model.

3 Model Output

The model output range is in dB(FS), with 0dB representing the largest punch output possible by the model. For comparison, Figures 2 and 3 show the standard momentary loudness model [4] and the punch model respectively for the opening bars of Michael Jackson's track - 'Billy Jean'.

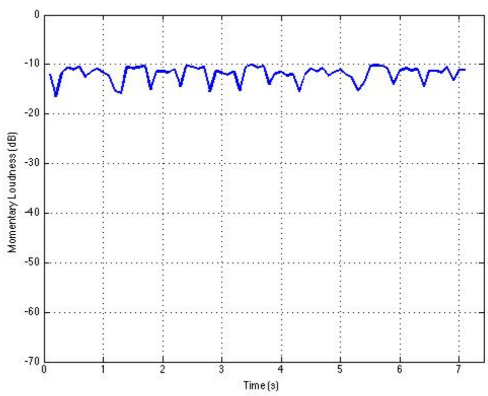

Figure 2. 'Billy Jean' – standard momentary loudness model output.

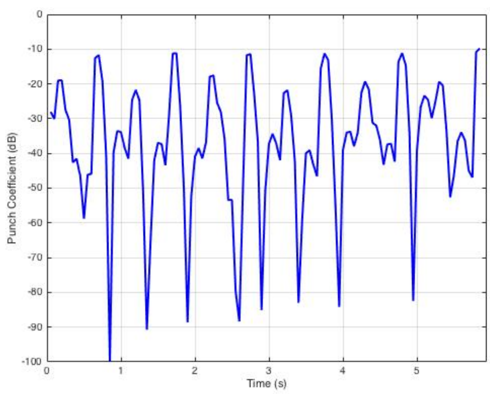

Figure 3. 'Billy Jean' – punch model output.

Using loudness as a metric, it can be seen that there is little that can be obtained from Figure 2 in terms of signal dynamics. Comparing this to the model output, Figure 3, the signal dynamic with respect to the transient components is clearly visible. More punch is evident at roughly 1-second intervals starting from 0s, less punch is evident starting at approximately 0.75s using the same interval.

With suitable metering ballistics, the raw output in dB(FS) of the model could prove useful in offering a real time indicator of punch to an engineer. A more useful application of the model output can be considered by evaluating the punch scores over a period of time, for example in a histogram. The output variables obtained from the histogram can be analysed, giving statistical data relating to the stimuli under test.

The histogram shown in Figure 4 represents the data extracted from the Billy Jean sample. It shows the magnitude of a particular punch frame and its frequency within the section of music under test. By examining the data in this way, maximum, minimum, range, median and standard deviation type measures can be extracted.

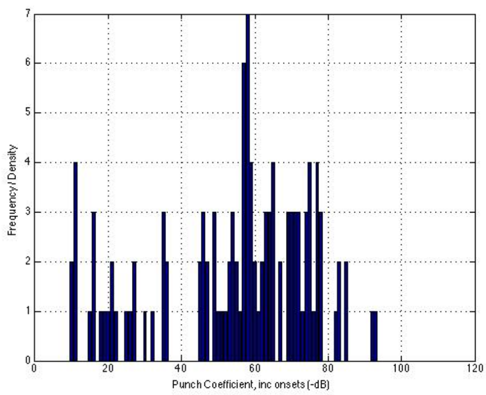

Figure 4. 'Billy Jean' - Histogram of punch scores.

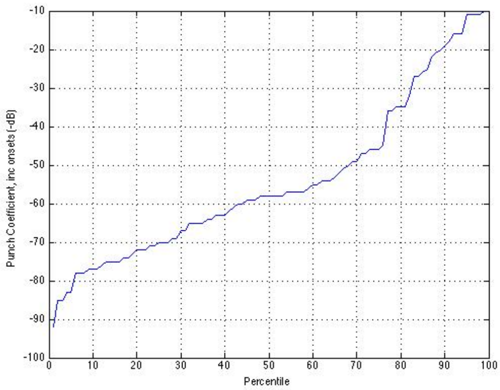

Figure 5. 'Billy Jean' – percentile plot of punch scores.

This punch model output can also be shown in the form of a percentile plot, Figure 5, showing an effective insight into the underlying '*punchiness*' of the music being measured. Use of such percentile data offers the possibility to compare entire pieces of music or sections therein.

With reference to Figure 5, it can be stated that punch frames of -20dB or more only occur approximately 10% of the time. Conversely, punch frames of -20dB or less occur approximately 90% of the time. Similarly to the current loudness range metering algorithm [4] 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.

4 Subjective Perception of Punch

4.1 Experimental design

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

The subjects 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.

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

The subjects listened in a near-field setup with Genelec speakers in an ITU-R BS. 1116-compliant listening room. The listening level was set and measured to be 76dB(A) and each subject stated this was a general listening level that they were used to.

4.2 Stimuli

The stimuli consisted of 12 excerpts of commercially available music; these are detailed in Table 1. The duration of each was 7s. All excerpts were down-mixed to mono, to suppress any spatial effects.

Each excerpt was then loudness normalized according to ITU BS.1770-3 [4], such that the overall loudness of the stimuli would be equal on playback. The level chosen was -23LU.

All the stimuli were perceived by the listeners to be playing back at the same loudness levels. The stimuli were chosen from various genres in order to not 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 Transient,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	Ambience	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 1. Stimuli used in the model validation test.

4.3 Subjective test results

Table 2 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 [8, 9]. A Matlab script OptiPt.m [10] 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 2. Forced-pairwise test scores.

4.4 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 of >50% will be achieved. How much higher above the chance level of 50% can be established through the calculation of z-score and consideration of both the population and the assumed chance population percentage [11]. The relationship can be summarised in equation 1 as follows:

$$Psig = 1.64 * \sqrt{\frac{Pu(100-Pu)}{n}} + Pu + 0.5 \qquad (1)$$

P_u = assumed population percentage 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 $Psig$ 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 5. Those shaded as light grey are significant based upon a p=0.10, one tailed z-score.

By inspection of Table 2, it can be seen that there is a general trend with respect to 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 this case, whilst file ID 1 has received the most votes, it cannot be stated with 95% confidence that it has more punch than files 2 & 3. We can state however that file ID 2 has more punch than file ID 3. 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 is 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 stimuli, it's possible to remove some of the stimuli utilised in the creation of the scaled output parameters. This is the method adopted and is detailed in section 4.5.

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. To do this, a Bradley-Terry-Luce model approach was applied.

4.5 Bradley-Terry-Luce model

The Bradley-Terry-Luce (BTL) model [8, 9] is a well-founded approach that states, given certain testable conditions, preference probabilities may be related to scale values in the following fashion:

$$Pab = \frac{v(a)}{v(a)+v(b)} \qquad (2)$$

Where Pab denotes the probability that stimuli a will be preferred over stimuli b, or on this case punchier. $v(a)$ and $v(b)$ are the number of votes that each stimuli received. A Matlab script OptiPt.m [10] was utilised to derive the scaled response coefficients based on the pairwise data and the output of this is shown in Figure 6. This figure is based on all of the stimuli tested and shows the 95% confidence intervals calculated from the covariance matrix returned by the function. These intervals correspond with the inter sample significance testing results described in section 4.4 The BTL model output shows the samples with significantly different rankings as having no confidence interval overlap. Likewise, 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. The important goal is to derive an interval scale that can be tested against the objective punch model. Figure 7 shows the model output after the removal of file IDs 2, 5 and 11. Note that the x-axis is replaced with the label ‘Samples’ denoting file order with respect to punch rating.

The testable conditions for a BTL model fit are those of transivity and goodness of fit. The former can be established by looking at the raw data scores in Table 2. The general rule of transivity is described as:

$$If\ A \geq B\ and\ B \geq C\ then\ A \geq C \quad (3)$$

Examination of the raw data shows that transivity is not violated when considering the total votes received by each stimuli. The goodness of fit statistic returned by OptiPt.m as a χ^2(chi squared) statistic is 22.84 (28). This value is 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.

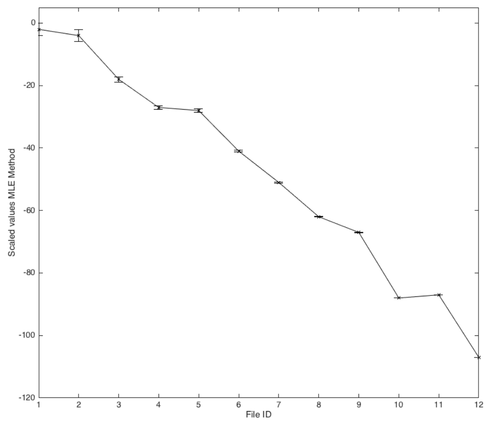

Figure 6. BTL Model output based on pairwise comparison data.

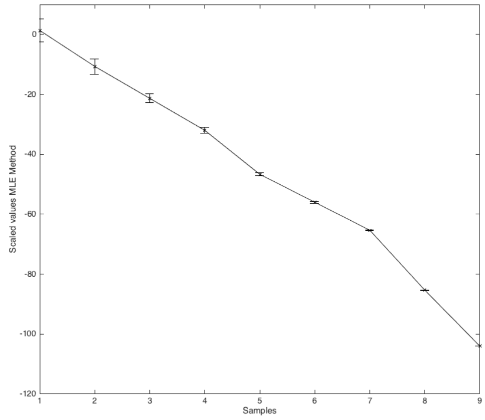

Figure 7. BTL Model output using only significantly different scoring files.

4.6 Model output correlation analysis

Two punch model outputs were compared against the subjective punch scores. These were the 95% percentile value (PM95) and the 95% Percentile value / Mean (PM95M). The latter could be considered to be a ratio derivation as found in dynamic range type calculations. Other computational measures were included to evaluate their correlation against the same derived punch scales, these were Crest Factor (CF) [5], Peak-to-Loudness ratio (PLR) [4], Inter-band-Ratio (IBR) [6] and Loudness Dynamic Range (LDR) [7].

The IBR measurement adopted was given the name IBR_diff as it was the range of values measured up to the 95th percentile. The LDR measure adopted utilised the 10th and 95th percentile in its calculation with a 3s and 50ms integration time for slow and fast loudness level calculation respectively.

As each of the measures chosen offers differing unit (SI) outputs, 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 3 shows the Pearson correlation coefficient (r) and the rank correlation (Spearman's *rho*) between each measure and the subjective punch scale.

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 3. Correlations of the tested measures with the perceptual punch scale.

In addition, and in order to show possible correlations between measures, a correlation matrix, shown in Figure 8, was produced. This table includes r2 values where significance is prevalent. In Table 3 and Figure 8, * and ** signifies correlation is significant at the 0.05 and 0.01 level (2-tailed) respectively.

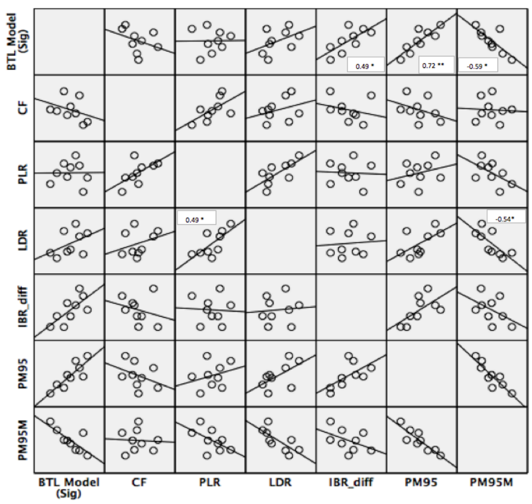

Figure 8. Correlation matrix of all tested measures.

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 respectively) being significant at the 0.01 level (2-tailed). The p-values for both these were .004 and .005 respectively (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 argument 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 rather than the perceptual increase in loudness afforded, which in turn would result in a change the PLR and CF [1].

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 [1,6]. 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 levels of perception of punch in stimuli tested.

The LDR measure, a measure of loudness dynamic range, is proposed as a measure of microdynamics [7]. It was included to ascertain any correlation with the punch attribute given the importance of dynamics within audio stimuli. An *r* coefficient of .442 indicates 'moderate' correlation strength, however a p-value of .234 (2-tailed) was observed. It can be seen in Figure 8 that the LDR has a 'strong' correlation with both the PLR and PM95M measures, with *r2* coefficients of 0.49 and -0.54 respectively, both are significant 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.

The LDR algorithm is also based upon the BS 1770 model of loudness measurement [4]. Its maximal difference type measure 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). In addition the PM95M model employs a frequency weighted derivation in its algorithm based on coefficients derived from a noise burst test [3], unlike the LDR which 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 by the listeners. This may be due to the combination of both onset detection and frequency band weighting employed in the punch model.

Both punch and LDR models implement an integrative process in the form of windowing. The LDR measure tested uses 3s 'slow' and 50ms 'fast' windows whilst the punch model employed 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.

5 Conclusions

A perceptually motivated model of punch 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.

The model utilised combines signal separation, onset detection and low-level parameter measurement to produce a perceptually weighted 'punch' score. The model coefficients were derived through a series of noise burst tests [3].

Four additional types of objective measures related to signal dynamics and loudness perception were evaluated against the same derived punch scale.

The PLR and CF measures showed the least correlation to the punch attribute, whilst the PM95 and PM95M model outputs showed the best correlations with *r* coefficients of 0.849 and 0.770 respectively.

From the results obtained, the PM95 measure can be consider 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. An *r* coefficient of .706 was observed with a p-value of 0.034 (2 tailed). Higher IBR_diff values corresponded with higher levels of punch perception in the stimuli tested.

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.

Further testing of the punch model could be undertaken in order to investigate the effect of its parameters on both processing speed and possible improvement in correlation with the punch scale derived for the stimuli. Parameters such as the aforementioned ‘100ms’ window size could be modified. In addition, temporal weightings could be applied as whilst temporal based weightings were applied to the noise burst stimuli to normalise their perceptual loudness, signal duration of the detected onsets wasn't incorporated into the model that has been described.

A wider selection of audio stimuli and greater number of subjects taking part in the test would also be beneficial.

6 Acknowledgements

The first author would like to thank his colleagues and expert listeners that took part in the many listening tests. Particular thanks go to Esben Skovenborg of TC Electronics for his useful comments received both at conferences and via email. Also, for his input in providing the LDR measures utilised in the testing process.

References

[1] S. Fenton, H. Lee and J. Wakefield, “Elicitation and Objective Grading of ‘Punch’ Within Produced Music,” 136th Audio Engineering Society Convention, Berlin (2014).

[2] B. Moore, “An Introduction to the Psychology of Hearing,” Great Britain: Academic Pres. (2003).

[3] S. Fenton and H. Lee, “Towards a Perceptual Model of 'Punch' in Musical Signals”. 139th International AES Convention, October, New York, USA (2015).

[4] ITU-R BS.1770-3, Algorithms to measure audio programme loudness and true-peak audio level, International Telecommunications Union, Geneva, Switzerland (2012).

[5] W. M. Hartmann, *Signals, Sound, and Sensation*. New York: Springer, (1998).

[6] S. Fenton, B. Fazenda and J. Wakefield, “Objective Measurement of Music Quality Using Inter-Band Relationship Analysis”, 130th Audio Engineering Society Convention, London, May (2011).

[7] E. Skoveborg, “Measures Of Microdynamics”, AES 137 Convention Paper, Los Angeles, October (2014).

[8] R. A. Bradley and M. E. Terry, “Rank Analysis of incomplete block designs,” The method of pair comparisons. Biometrika, 324-345 (1952).

[9] R. D Luce, “Individual Choice Behavior”. New York, Wiley, (1959).

[10] F. Wickelmaier, C. Schmid, “A Matlab Function to Estimate Choice-model parameters from Paire-comparison data”, Behavior Research Methods, Instruments & Computers, (2004).

[11] L. E. Harris and K. R. Holland, “Using statistics to analyse listening test data: some sources and advice for non-statisticians,” Proceedings of the Institute of Acoustics, (2009).