University of Huddersfield Repository

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

Elicitation and Objective Grading of 'Punch' Within Produced Music

Original Citation

Fenton, Steven, Lee, Hyunkook and Wakefield, Jonathan P. (2014) Elicitation and Objective Grading of 'Punch' Within Produced Music. In: 136th International Audio Engineering Society Convention, 26th-29th April 2014, Berlin, Germany.

This version is available at http://eprints.hud.ac.uk/id/eprint/20141/

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.

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

Audio Engineering Society

Convention Paper

Presented at the 136th Convention
2014 April 26–29 Berlin, Germany

This Convention paper was selected based on a submitted abstract and 750-word precis that have been peer reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This convention paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Elicitation and Objective Grading of 'Punch' Within Produced Music

Steven Fenton[1], Hyunkook Lee[2], and Jonathan Wakefield[3]

[1] *School Of Computing & Engineering (MTPRG), University of Huddersfield, Huddersfield, UK*
s.m.fenton@hud.ac.uk

[2] *School Of Computing & Engineering (MTPRG), University of Huddersfield, Huddersfield, UK*
h.k.lee@hud.ac.uk

[3] *School Of Computing & Engineering (MTPRG), University of Huddersfield, Huddersfield, UK*
j.p.wakefield@hud.ac.uk

ABSTRACT

This paper details the results from an investigation into the objective grading of punch within a complex musical signal. The term punch is a subjective term, which is often used to characterize music or sound sources that exhibit a sense of dynamic power or weight to the listener. In a novel reverse elicitation process, experts were asked to create audio samples which they perceived as having punch using a multi-band wave shaping process. Expert 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.

1. INTRODUCTION

Whilst much work has been performed and standards have been defined with respect to audio encoding and subsequent codec quality measurement, the measurement of perceived quality in music production and mastering within the music industry is still in its infancy.

It is important to clarify what is meant by 'produced music' in the context of this paper. In general, a completed piece of produced music will be the sum (or mix) of products of a number of discrete processes and audio stems resulting in a stereo or multichannel audio file. This file would then undergo a further level of processing, referred to as mastering. This final process is generally utilised to normalise loudness levels (in the case where a collection of songs are to be presented) or

to address any spectral anomalies that may have occurred during the mixing process.

Objective measurement of both the completed file and/or stems would be useful in the music production process. Such measures would enable engineers to evaluate the success of their production techniques in achieving certain perceptual requirements even if the listening environment and/or equipment being used for playback was not ideal.

An example of an objective measure that has been successfully implemented is that of loudness and the resultant EBU R-128 standard [1][4] based upon the ITU-BS.1770-3 specification [3]. Despite the standard being directed primarily at the broadcast industry and the regulation and monitoring of broadcast sound levels, metering tools supporting the R-128 model are being implemented as plugins [2] that can be used as tools in the music creation process. It is the hope of many, including the authors, that use of this particular objective measure will continue to proliferate in this field and lead to the majority of music being pre-normalised to the required level, prior to delivery to the broadcaster. Thus, finally ending the loudness war.

Furthermore objective measures would help to streamline the music production process and help to reduce the need to rework mixes or masters of songs.

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.

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 and creating additional nuisance artefacts or annoyance.

This forms the motivation for this work wherein we describe a method for the reverse elicitation of parameters pertaining to the sensation of punch.

In section 2, we provide an outline description of musical transients and dynamic range and their relationship with music attribute perception. In section 3, we outline the testing strategy adopted to ascertain the parameters relating to punch within a musical signal, the results of the tests are presented in section 4. In section 5, we discuss the results and propose some signals measures that may be useful in the mapping of the punch perceptual attribute.

2. AUDIO PERCEPTION & TRANSIENTS

Music can be considered to be a collection of complex tones, with complex tone consistsing of a number of differing 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 [5][6].

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. Detection of transients can be useful in such applications as note detection, signal enhancement, dynamic range control and musical transcription [7][8][9][10]. Various methods of transient detection can be employed with varying degrees of success depending on genre and application [7][11].

Almost all genres of music have significant transient content throughout as a result of differing tone onsets. Modification of the transient portion of a sound source has been shown to modify the perception of the source by the listener [7][11][12]. A wide vocabulary of terms are used in the music industry and wider circles alike to describe perceptual attributes; for example, *warm, bright, soft* or *heavy*.

Work to establish verbal description and dimensions for some of these perceptual attributes has been extensively explored in previously published papers [5][13][16]. Early work by Freed [19] and others, focusing on the perception of mallet hardness and also noted that whilst the musical importance of the attack portion of a signal is well known, most studies have focuses on steady state sounds. Freed concluded that the mean spectral centroid is a strong predictor for the mallet hardness. Feature extraction of audio, based on a set of low level descriptors is defined by the MPEG 7 standard [20]. Spectral Centroid, amongst other measures is defined in

this standard but further experimentation in mapping the measures to perceptual attributes is required.

The authors have found very little literature on measuring the perceptual attribute 'punch' and indeed its definition. This is surprising given that, as stated earlier, music is often characterised by listeners as being punchy or not.

Goodwin et al. [11] refer to punch as a legitimate perceptual attribute and their work 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 therefore, 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. [12] also refer to the 'punchiness' as a perceptual attribute of a mix and their paper goes on to test various transient detection techniques along with a sub-band approach to musical transient modification. Both of the works cited don't measure punch objectively however, they do show results that imply the perception of punch is altered by the modification of the transient. The latter stating that although the perception of punch was greater in all modified cases than the hidden references, there was no significant difference between the use of different transient detection models.

As stated by the authors in earlier research [17][18], punch, can be described as a particular moment in a production where there is a degree of change in power in the music. In essence, productions that do not possess any transient information cannot posses 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 may contribute to the perception of punch perceived by the listener. 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.

3. DESCRIPTION OF TESTING

3.1. Stimuli Elicitation

Twelve expert listeners took part in an 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 sound source could be carefully controlled resulting in a known reference that did not suffer from room coloration and was independent of drum skin type and batter. 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.

The kick drum source was synthesised using a T Bridge oscillator type model commonly found in the TR-909 synthesiser. It was then fed through a 3-band linear phase filter, their respective cut-off frequencies and Q settings are shown in Table 1.

Table 1: Filter Corner Frequencies

Filter Type	Fc (Hz)	Fc (Hz)	Q
Low Pass LF	947	-	6.5
Band Pass MF	947	3186	6.5
High Pass HF	-	3186	6.5

These frequencies were chosen as they approximate the 1st, 2nd and 3rd set of eight critical bands in the auditory system and also formed the subbands used in previous testing by the authors [17]. Each sub-band was then fed into a temporal shaper, the outline of which is shown in Figure 1. Pr1, Pr2 and Pr3 are identical shapers. As can be seen, the test interface the test subjects were asked to use to modify the audio was intentionally left unlabelled and was merely a collection of control knobs in a random arrangement, therefore any pre-conceptions of typical audio wave shaping controls or production preference biasing effects were avoided. The experts 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 to until they thought they had achieved a maximum punch attribute. The exercise took place using headphones to eliminate room coloration and

speaker influences. Playback levels were set to 76dB(A).

Figure 1: Test Interface Wave Shaper

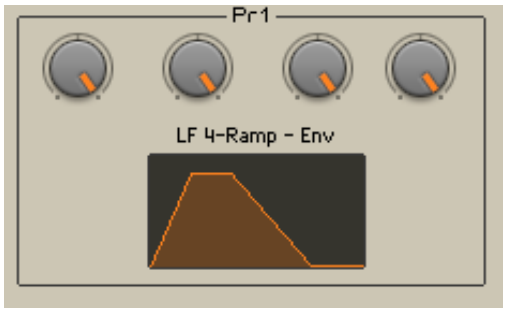

Figure 2: Example waveshape setting; the left to right controls refer to the attack time, the peak level, the peak hold time and finally the release time respectively.

During the waveshaping 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 a control marked MU adjust on the interface. The level was monitored using a NuGen Audio Loudness Meter [2] and was set to a level of -32 LUFS.

The reciprocal of the waveshaper output was used to shape the respective sub-band. An example waveshape is shown in Figure 2. The rationale behind waveshaping by envelope rather than modelling of a specific audio function (eg. equaliser or compressor) compressor type was to reduce the number of experimental variables and prevent 'equipment signatures' being considered during the process by the experts.

Each expert 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 second sample. In total 24 samples were created which when referenced to the original sound source, exhibited an increased perception of punch to the expert who created it. All samples, including the sources were 44.1kHz, 16bit, Mono WAV format.

3.2. 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 using headphones to eliminate room coloration and the playback level was fixed at 76dB(A).

A modified MUSHRA formed the basis of the listening test. The test being modified to allow the listeners to rate the samples as being less punchy than the hidden reference, in this case the unprocessed sample. As the samples were created without a reference it could hold true that a listener may perceive a sample as being less punchy than the reference itself due to the waveshaping chosen to create that sample. The scale chosen ranged from 0 to 140, with samples rated as 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 3, the full interface consisted of all 14 samples visible across the screen.

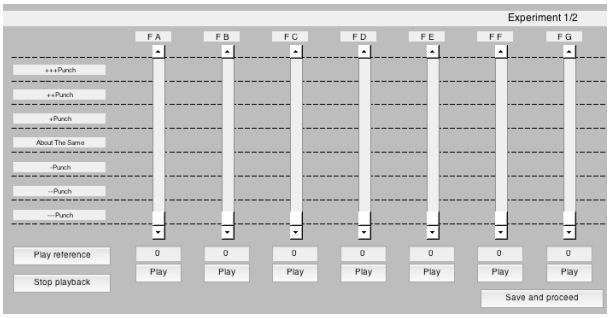

Figure 3: Modified MUSHRA interface portion.

The individual scores were collated and then Z-score normalised. From this a Mean Punch Score (MPS) profile for each sample was obtained and 95% confidence intervals were calculated for each. In addition the listeners were asked to describe what they perceived the punch attribute to be, and these were collected as set of verbal punch descriptors.

3.3. Signal Processing and Perceptual Mapping

A number of parameters were analysed using the best and worst samples based on normalised MPS achieved. 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 [20]. The signals were analysed using a combination of Matlab scripts using an N-point SFFT with a variable step size and Sonic Visualiser [21].

Parameters measured were *Temporal Crest Factor, Spectral Centroid, Log Attack Time, Signal Intensity, Intensity Ratio* and *Rhythm Strength*. A description of these measures is omited from this paper, however, explanations for all of them can be found in the MPEG-7 specification [20] and in the plug in documentation for Sonic Visualiser [21]

4. EXPERIMENTAL RESULTS

The following charts (Figures 4-5) show the normalised MPS along with 95% confidence intervals. The x-axis shows each waveshaped file. File 1 (F1) is the unprocessed reference. 3 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.

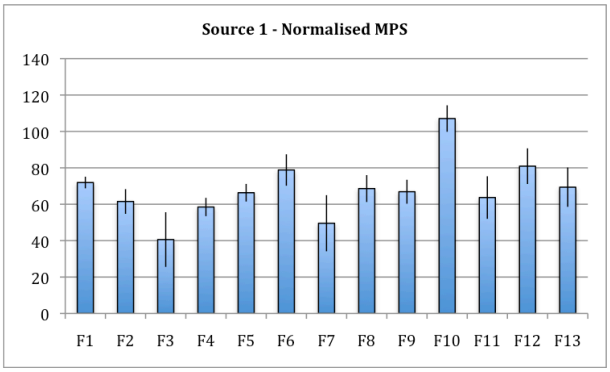

Figure 4: Source 1 – Normalised MPS per file.

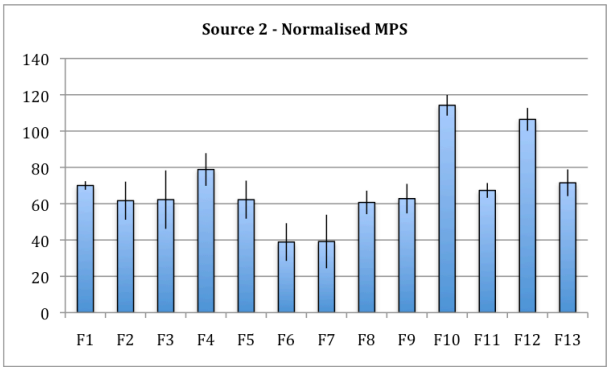

Figure 5: Source 2 – Normalised MPS per file.

4.1. 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"

They are included as a general guide for further elicitation research into this subject area.

4.2. 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 ($p = 0.676$, $F = 0.190$).

4.3. Objective Feature Extraction

The subjective experimental results were examined and through post statistical analysis of the data best and worst samples were identified. The naming conventions used in the objective feature extraction results relate to the subjective test result files as follows.

Referring to Figure 4:
- F1: Reference 1
- F10: Ex 1 – Best
- F3: Ex 1 – Worst

Referring to figure 5:
- F1: Reference 2
- F10: Ex 1 – Best
- F6: Ex 2 - Worst

A large number of measures were taken. The following represents a selection of those measures that were

considered the most significant. The authors are undertaking further analysis, the results of which will be presented in a future paper.

Table 2: Spectral Centroid Measures (1024-point FFT)

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

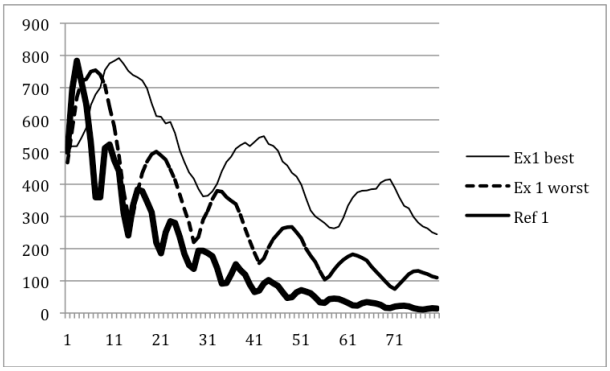

Figure 7: Signal Intensity over time

Table 3: Rhythm Strength

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

Table 4: Typical Intensity Ratio

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

5. DISCUSSION OF RESULTS

Figures 4 and 5, 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 Figues 8 and 9. The samples are shown in rank order from left to right with the highest first for each source tested.

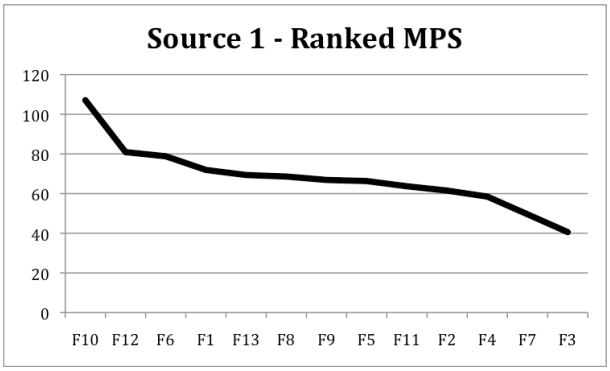

Figure 8: Rank Scored Samples vs. MPS– Source 1

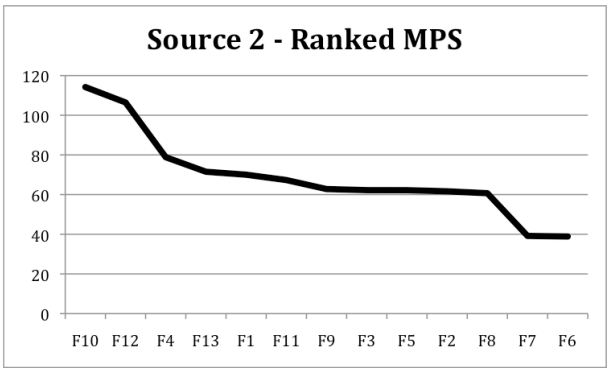

Figure 9. Rank Scored Samples vs. MPS– Source 2

Table 2 shows the Spectral Centroid measures for the reference sources and best and worst samples for each experiment. One can observe that the highest MPS ranking sample obtained a spectral centroid value of 1263.11Hz. Contrast this with 575.14Hz of the worst case sample and one can see there is a significant difference between the two. With reference to a typical percussive instrument timbre [20], a spectral centroid of approximately 1217.34 Hz would be expected. All the samples that achieved a high mean punch score have centroid measures around this figure.

The spectral centroid measures in Table 2 were obtained by analysing the full temporal response of the kick samples. An additional analysis was run to measure the spectral centroid at the point of maximum intensity and a mean value of 41.9Hz between all samples was obtained. 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 modeled using a simple T-Bridge arrangement, which is fundamentally a modulated sine wave without complex modeling of the membrane, beater, drum shell or dampening, the spectral centroid measure outlined above would be expected. Further testing to establish the variation in punch perception on modification of the centroid at the point of maximum onset would be beneficial. If the spectrum of the onset was made more complex, one might expect the maximum onset centroid to change.

Intensity ratio is an indication as to which band of frequencies constitutes the majority of a signals power. From the Intensity Ratio measures taken, the first subband 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 4. As previously outlined, that the samples were synthesised with a fundamental tuning of 47Hz, this is expected.

Table 3 shows the rhythm strength measured for sample selection. Both the best MPS samples measurements were greater than both the reference samples and worst MPS rated samples. Rhythm strength is the sum of the magnitudes of the power spectrum in the onset of the signal. This could be intuitively linked the rating of the perceived punch of the sample. The log attack time measured for the three samples shown in Figure 7 (best, worst and reference) is -0.856, -1.092 and -1.468. this measure is based on the logarithm of the onset time in seconds.

The best scoring sample has a longer onset but larger rhythm strength than the other two. Hence, more spectral energy in the onset portion of the sample. If one examines Figure 7 which shows the intensity of the signal over time for source 1, 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. Crest factors for the best, worst and reference in this plot are 8.358, 10.559 and 12.428 respectively. With reference to the MPscores, the reference scored better than the worst case sample with source 1, thus suggesting that crest factor alone may not be the key contributor towards punch perception but a consideration of the onset must also be taken into account.

As highlighted earlier, the spectral centroid of the whole envelope needs to be considered to determine whether the timbre of samples under test fall within the bounds of an expected percussive sound. The spectral centroid measures shown in Table 2 showed significant variation between the samples. Given the initial sample creation exercise involved only temporal waveshaping and no direct modification of frequency spectrum (i.e. use of equalisation) took place, the resulting change in centroid was a direct result of the envelopes used. This is expected as the temporal modification would result in additional harmonic components appearing in the frequency domain.

6. CONCLUSIONS

The authors have begun to identify a possible correlation between the perceived punch attribute and the measures of rhythm strength and crest factor. In the case of the samples used in this paper (kick drum), the overall spectral centroid is important in establishing the timbre at least lies within the boundaries expected of a percussive instrument. A perception of low punch is synonymous with kick drum samples with a low overall spectral centroid score.

7. FURTHER WORK

Further testing to establish the variation in punch perception on modification of the centroid at the point of maximum onset would be beneficial.

Examination of the individual temporal envelopes chosen for each sample, with respect to its associated perceived punch level will be undertaken and presented in a future paper by the authors. This work could identify weighting factors that could be applied to a tuned transient modifier aimed at increasing punch in an audio signal but with a reduced level of artifacts in the audio processing algorithms.

Additional testing is required to further support the correlation between key objective measures outlined and the perceived punch score.

8. REFERENCES

[1] EBU-R128, Loudness normalisation and permitted maximum level of audio signals, EBU PLoud Group, August 2011

[2] Vis-LM-H, NuGen Audio Loudness Meter with history, http://www.nugenaudio.com/visLM_loudness-meter_VST_AU_RTAS.php [accessed 2nd February, 2012]

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

[4] EBU – Tech Doc 3343 Practical guidelines for Production and Implementation in accordance with EBU R 128, August 2011

[5] J.M.Grey "Multidimensional Perceptual Scaling of Musical Timbres", JAES, vol 61, 1977.

[6] J.V and R.A. Rasch, "The Perceptual Onset of Musical Tones", Perception and Psychphys, vol 29, 1981.

[7] N.Collins, A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychacoustically Motivated Detection Functions, AES Convention Paper 6363, May 2005

[8] C.Avendano and M.Goodwin, Enhancement of Audio Signals Based on Modulation Spectrum Processing, AES Convention Paper 6259, October 2004

[9] M.Walsh, E.Stein and Jean-Marc Jot, Adaptive Dynamics Enhancement, AES Convention Paper 8343, May 2011

[10] E.Wang and B.T.G.Tan, Application of Wavelets to Onset Transients and Inharmonicity of Piano Tones, JAES, Vol 56, No.5, May 2008

[11] M.Zaunschirm, J.Reiss and A.Klapuri, A High Quality Sub-Band Approach to Musical Transient Modification, Computer Music Journal, Volume 36, Number 2, Summer 2012, pp. 23-36

[12] M.Goodwin and C.Avendano, Enhancement of Audio Signals Using Transient Detection and Modification, AES Convention Paper 6255, October 2004

[13] J.Stepanek, Musical Sound Timbre: Verbal Descriptions and Dimension', DAFX Conference, 2006

[14] S.Hainsworth and M.Macleod. Onset detection in musical audio signals. In Proc. Int. Computer Music Conference, pages 163–6, 2003

[15] JP Bello, L Daudet, S Abdallah, C Duxbury, M Davies, M Sandler, A Tutorial on Onset Detection in Music Signals. IEEE Transactions on Speech and Audio Processing. 13, 1035–1047, Sept. 2005

[16] S Lakatos, A common perceptual space for harmonic and percussive timbres Perceptual Psychophysics 26, p1426 2000

[17] S.Fenton., B.Fazenda, and J.Wakefield, 'Objective quality measurement of audio using multiband dynamic range analysis'. Institute of Acoustics (IOA) Conference- November 2009

[18] S.Fenton and J.Wakefield, Objective profiling of perceived punch and clarity in produced music. 132nd Audio Engineering Society Convention, April 2012

[19] D.J.Freed, Auditory Correlates of Perceived Mallet Hardness For a Set of Recorded Percussive Sound Events, J.Acoustical Society of America, Am.87, 1990

[20] MPEG 7 – ISO/IEC-15938 Information Technology – Multimedia Content Description Interface - Part 4 – Audio – October 2001

[21] C.Cannam, C.Landone, and M.Sandler, Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files, in Proceedings of the ACM Multimedia 2010 International Conference