View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by University of Huddersfield Repository

University of Huddersfield Repository

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

Hybrid Multiresolution Analysis Of 'Punch' In Musical Signals

Original Citation

Fenton, Steven, Lee, Hyunkook and Wakefield, Jonathan P. (2015) Hybrid Multiresolution Analysis Of 'Punch' In Musical Signals. In: 138th Annual Audio Engineering Society AES Convention, 7th-10th May 2015, Warsaw, Poland.

This version is available at http://eprints.hud.ac.uk/id/eprint/24358/

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 138th Convention
2015 May 7–10 Warsaw, Poland

This paper was peer-reviewed as a complete manuscript for presentation at this Convention. 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.

Hybrid Multiresolution Analysis Of 'Punch' In Musical Signals

Steven Fenton, Hyunkook Lee, and Jonathan Wakefield

School Of Computing & Engineering (MTPRG), University of Huddersfield, Huddersfield, UK
s.m.fenton@hud.ac.uk, h.k.lee@hud.ac.uk, j.p.wakefield@hud.ac.uk

ABSTRACT

This paper presents a hybrid multi-resolution technique for the extraction and measurement of attributes contained within a musical signal. Decomposing music into simpler percussive, harmonic and noise components is useful when detailed extraction of signal attributes is required. The key parameter of interest in this paper is that of punch. A methodology is explored that decomposes the musical signal using a critically sampled constant-Q filterbank of quadrature mirror filters (QMF) before adaptive windowed short term Fourier transforms (STFT). The proposed hybrid method offers accuracy in both the time and frequency domains. Following the decomposition transform process, attributes are analyzed. It is shown that analysis of these components may yield parameters that would be of use in both mixing/mastering and also audio transcription and retrieval.

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. If the perceptual 'punch' attribute can be extracted as a feature, it can be utilized as an additional search criterion in MIR in addition to being a reliable normalization metric in music production.

Previous work by the authors [4] explored the reverse elicitation of parameters pertaining to the sensation of punch. From this work subjectively graded punch samples were obtained. Regression analysis of the high-level control settings chosen by the expert listeners revealed no significant correlation between any singular control setting and the resulting punch score. Therefore, further detailed analysis of the resultant signals is required.

This paper describes a hybrid multi-resolution technique that initially decomposes the musical signal using a quadrature mirror filter bank (QMF) before applying a short time Fourier transform (STFT) to each band. 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 such advantage is a high degree of resolution achieved in both the time and frequency domains, this is explained in section 6. Following the initial transform process, transient, steady state and residual components (TSR) are extracted. The method of separation presented 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 analyzed 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 part.

New measures may then be utilized to model and objectively evaluate punch in produced music and be an additional metric used in MIR. The paper concludes with an example of how the measure can be applied in a music production and/or mastering environment to both measure dynamics in addition to identifying source elements within the music itself.

2. BACKGROUND

In music production, metering tools are often used to signify signal presence, level and in the case of an audio mastering or broadcast scenario signal loudness. It's well known that the past two decades have seen a gradual decrease in dynamic range across a wide range of formats, particularly on the CD.

Automatic loudness normalization 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 artist 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.

A characteristic related to dynamics is known as 'punch'. A hypothesis stated by the authors in earlier research [4] is that punch can be described as a short period of significant change in power in a piece of music or performance. 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. Furthermore dynamic change in particular frequency bands contribute to the overall perception of punch perceived by the listener and this is inherently affected by the overall average loudness level at that time [21].

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 a reliance 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. Some of those parameters are examined later in this paper.

3. EXISTING METERING

The current ITU loudness standard measurement algorithm [13] incorporates individual audio channels, which are independently filtered to simulate the sensitivity of the human ear and head diffraction effects.

The power in each channel is summed to obtain the power in the entire signal. This power is averaged over the entire program to obtain a single number metric for the program loudness. In addition, Loudness Range, Short-term Loudness and Momentary Loudness are all offered to indicate 'dynamics' within the program material. This approach, along with other general dynamic range measures consider 'macrodynamics'.

These are largely based on an integrative approach thus can't really be utilized to quantify attributes such as punch, exists in the 'microdynamic' scale of the signal. Peak to Loudness ratio (PLR), also specified in [13] could be utilised to measure a degree of microdynamics however, the peak measure obtained in the case of an entire track, may not be attributable to the track as a whole.

Fine time-scale approaches have been developed [14][11] however these approaches still consider the signal as whole when calculating Peak and average levels loudness levels at different resolutions. By whole, we mean the complete complex mix of transient, steady state and residual components.

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 utilizing signal separation the true dynamics of the signal as a whole can analyzed whilst considering the effect of each component within the signal. What we're considering 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.

4. SINES, TRANSIENTS & RESIDUAL

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.

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].

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 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][10]. We outline some of these approaches in section 5.

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. Modification of the transient portion of a sound source has been shown to modify the perception of the source by the listener [7][10][12].

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. [1][2].

The steady state components of the signal are usually related to pitched instrumentation. It's shown in section 7 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 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 qualitative efficiency in audio compression algorithms.

5. COMPONENT EXTRACTION

To precisely discriminate transient, steady-state and residual components is not an easy task. Much work has been performed in this area, excellent reviews and tutorials on the subject are given in [15][16]. It's concluded within this work 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 a minimum processing and latency load when considering audio metering applications. However, when the onsets are softer, the complexity of the algorithm increases, a combination of differing techniques is suggested as being the most effective means of detection of all transients.

The complexity of the algorithm should be chosen to best fit the application. This papers focus is that of audio metering therefore processing speed needs to be minimal and latency reduced to a minimum. For this paper, we do not present the detection of soft onsets however work on this more complex model is ongoing.

5.1. Fast Onset Detection

Fitzgerald [1] proposed an efficient method of transient and steady state separation that utilised median filtering. This approach, inspired by Ono et al.[17] considers that transient components will be broad-band in nature with highly concentrated energy in time, whereas steady-state sources are taken as discrete narrow-band components with smooth magnitude temporal behaviour. These components can be seen in spectrogram as vertical and horizontal ridges, respectively.

Further investigation utilising this method was performed Iragaray et al.[2]. Their work incorporated the use of a Wiener filter stage and a Stochastic Spectrum Estimation (SSE) method proposed by Laurenti et al. [17] 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, Fitzgeralds approach [1] was adopted for this paper to detect fast onsets. However the separation algorithm was modified to reduce spill between components. This modification, proposed by Driedger et al.[19] incorporated separation factors which allow for the tightening or reduction of steady-state or transient bleed.

6. PROPOSED METHOD

The chosen analysis model is shown in figure 1. 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 psycho-acoustically tuned to the auditory response.

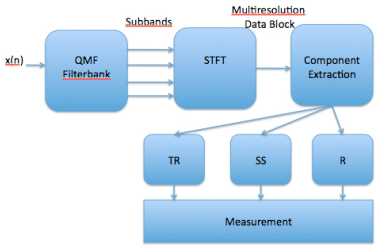

Figure 1. Analysis Model.

The choice of decomposition 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 [20]. 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 Fs, where Fs 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 Fs/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 [20], 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 our method, a full packet tree decomposition was deemed unnecessary. This was due to the fact that a model based on auditory response requires lower resolution at higher frequencies so 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 0.5π.

Downsampling by a factor of 2 is employed after each QMF filter stage, thus reducing data redundancy. In order to keep processing overhead to a minimum, 3 level decomposition into 4 bands took place as shown in figure 2.

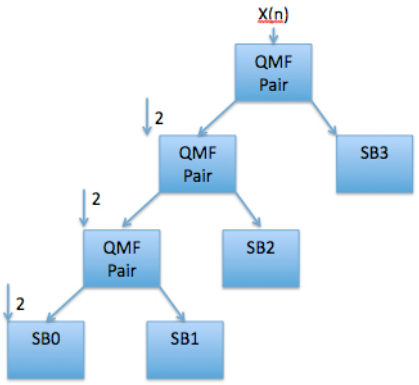

Figure 2 Multiresolution filterbank of cascaded QMF filters.

This filtering results in four sub-bands, as shown in table 1.

Table 1. Measurement Frequency Bands

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

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

$$S(t,k) = \sum_{n=-\infty}^{\infty} w(n)s(n+tH)e^{-j\omega kn/N} \qquad (1)$$

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.

6.1. 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 outlined in section 4, strong percussive onsets tend to spread across the spectrum in a broadband nature, this spread tend to narrow in time in the upper frequency bands. 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 infact able to analyse the signal in a multi-resolution in time basis. 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 psycho-acoustic 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 resulting STFT coefficients are then re-combined into an overall multi-resolution data block, a spectrogram example of which is shown in Figure 3 before being passed through the median filters.

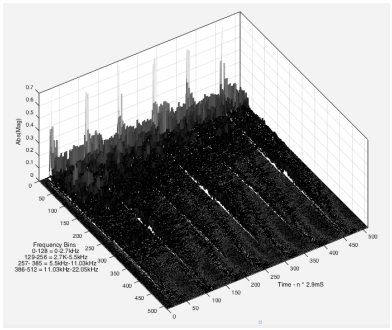

Figure 3. Multiresolution Spectrogram of 'Animal' example.

6.2. Separation Of Components

Median filtering performed across the time axis results in a steady-state enhanced spectrogram, in addition transient outliers are suppressed. Likewise, filtering across the frequency axis tends towards suppressing the steady state components and enhancing the transients. This can be seen in figure 4.

Following the proposal [19] outlined in section 5.1, separation factors of 3 and 2.5 were chosen for βt and βss respectively. These gave good separation when tested on a variety of sources.

The binary masks utilised for component separation are defined as

$$M_{SS}(t,k) = (S_{SS}(t,k) / S_t(t,k) + e) > \beta ss \quad (2)$$
$$M_{tr}(t,k) = (S_t(t,k) / S_{ss}(t,k) + e) \geq \beta t \quad (3)$$

Where S_t & S_{ss} are the median filtered STFT data blocks. Separation is achieved by applying the masks to the overall multi-resolution data block which results in transient and steady state data blocks.

$$TR(t,k) = S(t,k) * M_{tr}(t,k) \quad (4)$$
$$SS(t,k) = S(t,k) * M_{ss}(t,k) \quad (5)$$

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

$$R_m(t,k) = 1 - [M_{tr}(t,k) \mid M_{ss}(t,k)] \quad (6)$$

The residual components are then extracted as

$$R(t,k) = R_m * S(t,k) \quad (7)$$

An example of a file that has been separated is shown in figure 4.

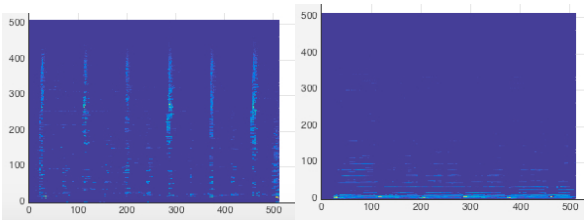

Figure 4. (a) Transient; (b) Steady State enhanced by median filtering

6.3. Analysis Parameters

As outlined in section 3, common metering tools used during the mixing and mastering consider the signal as a whole. Whilst being accurate for a 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 peak and RMS calculations being the sum of all the signal components, transient, steady state and residual.

The primary use of integration in dynamic range measures is to stabilize 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 from a psychoacoustic perspective. E.g. During moments of true dynamic activity, one would expect a measure based on the components that relate solely to this activity.

Through 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.
Our first measure of interest is the Transient to Steady State ratio. Considering the hypothesis outlined in section 2 that punch perception is related to transient change, the energy density at a particular moment in time and the overall loudness, this measure considers all three.

In this paper, the component power of each component is calculated as a summation of each frequency bin for every STFT hop.

Should the steady state component power 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 prior to the detected transient, thus determining the potential for masking.

The parameter is given as

$$TSR(t) = 10 * \log[TR(t) / SS(t)] \quad (8)$$

where TR & SS are the sum of the k magnitude bins of the transient and steady state components respectively.
An additional parameter is also measured which takes into account the residual component, as follows

$$[TSR + R](t) = 10 * \log[TR(t) / [SS(t) \mid R(t)]] \quad (9)$$

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 8 shows just the transient component centroid measure, where *f(n)* is the bin centre frequency.

$$SCtr(t,k) = \frac{\sum_{n=-\infty}^{\infty} f(n)TR(n,k)}{\sum_{n=-\infty}^{\infty} TR(n,k)} \qquad (10)$$

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 utilized are summarized in table 2.

Table 2. Measurements proposed

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

A raised-cosine (Half Hanning) filter was applied 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 that have this filter applied are shown as 'Smoothed'.

7. PRELIMINARY RESULTS

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

Following the separation and filtering processes, the measures described in section 6.3 were obtained. The key measures of interest are included here. All plots show the 'n * 2.9mS' timeframe along the 'x' axis where n is the STFT timeframe block.

With respect to Figure 6, which shows the power summation of the transient components over time, one can clearly see the effectiveness of the transient separation. Each peak corresponds with either a kick, snare or palm muted guitar chord. If this measure were utilized for onset detection for drum transcription, the latter palm muted onset could be removed simply by the introduction of an 'onset detection threshold'.

A further enhancement to this can be obtained by utilizing the spectral centroid of the transient component which is shown in Figure 7(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 7(b), 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 7(c), 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.

With reference to Figure 5 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 peaks average around -5 to 10 dB whereas when the residual is considered the associated levels fall to between -12 to -22 dB.

One can clearly see the dynamics of the signal, created by the transient components. Of note is the addition of peaks present at 75 block intervals. These are as a result of the small power peaks in Figure 6 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 to the arrangement does add an additional punch element that should be considered.

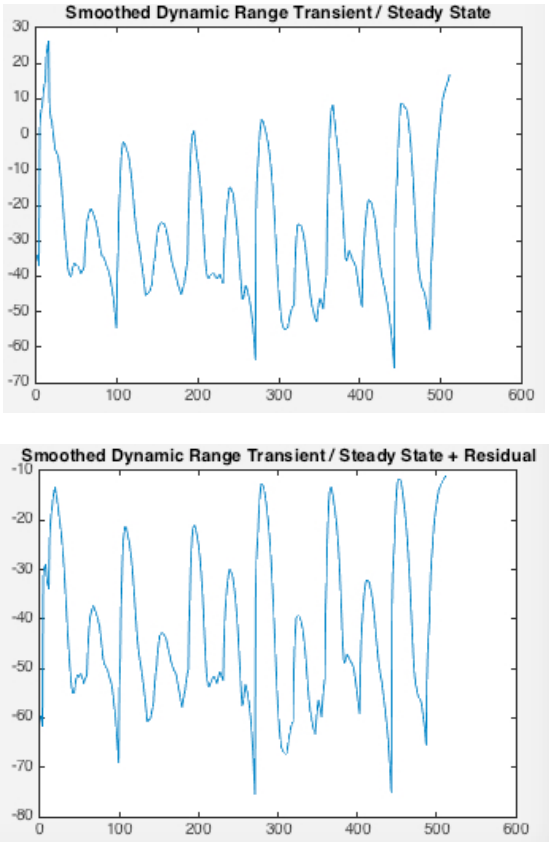

Figure 5 (a) TSR and (b) TSR+R vs. Time.

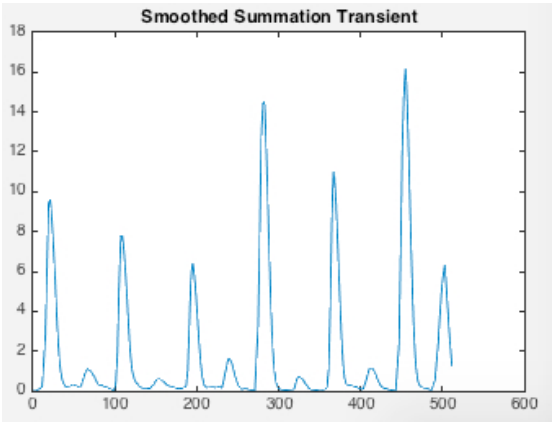

Figure 6 Transient Power Summation vs. Time.

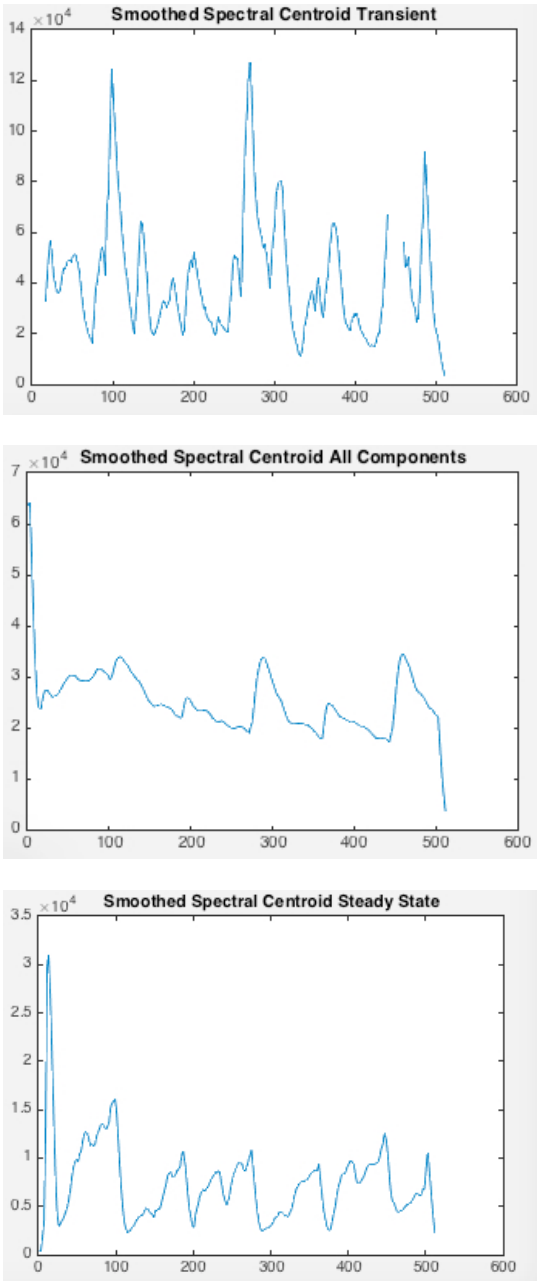

Figure 7 Spectral Centroid Measures of (a) Transient (b) Overall and (c) Steady State component.

If the mix were such that the steady state components were made much louder, thus masking the transient elements somewhat, it could be expected to see the peaks shown in Figure 5(a) & (b) reducing 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. The measure of TSR+R would be an applicable measure in this instance.

Considering a well mixed track, whereby the sound sources had been mixed effectively with minimal masking should reveal good transient power that will allow a high TSR to be achieved, and thus the perceived punch to be greater.

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 the resultant could be then utilized to recompose a noise free signal. However, the residual may contain important information that can't be discounted completely, e.g. 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 artifacts may well appear in the residual component and therefore may be deemed 'important' as far adding to the overall texture of a music track.

8. FURTHER WORK

The inclusion of a soft onset detection mechanism should yield additional components that would be included with in the transient data block. The authors have explored the use of the both phase deviation and weighted phase algorithms however, whilst effective in detecting the softer transients, they were too susceptible to noise, noise such as that introduced by distorted guitars being one such issue. A model utilizing the Euclidian distance is currently being explored.

The model utilizes 4 sub-bands. A more elaborate and natural extension to this could be the implementation of a full auditory filterbank as proposed in [22] whereby TSR analysis could take place close to that of a natural hearing response.

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 [19], different filter sizes in addition to DFT frame sizes are explored and this should prove very useful in progression of this research.

The model yields the possibility to perceptually weight the transient and steady state frequency bands. Considering that lower centroid components may perceptually exhibit more punch to the listener, the TSR measure could be weighted accordingly.

Subjective tests are planned with a panel of expert listeners. The listening test will attempt to evaluate the effectiveness of the TSR measure against punch perception across differing audio samples.

9. CONCLUSION

A hybrid multiresolution model has been proposed that decomposes an audio signal into its component parts. It's shown that analysis of these components may yield parameters that could be of use in both mixing/mastering and also audio transcription and retrieval.

10. REFERENCES

[1] D. Fitzgerald, "Harmonic/percussive separation using median filtering." Proc. of the DAFx-10, Graz, Austria, Sept. 2010.

[2] I. Iragaray and L.W.P Biscainho, "Transient and Steady State Component Extraction Using Non-Linear Filtering," Congreso Internacional de Ciencia y Tecnología Musical – CICTeM, 2013.

[3] C.Duxbury, M.Davies and M.Sandler, "Separation of Transient Information in Musical Audio Using Multiresolution Analysis Techniques" Proc. of the DAFx-01, Limerick, Ireland, Dec. 2001.

[4] S.Fenton, H.K.Lee and J.Wakefield, "Elicitation and Objective Grading of 'Punch' Within Produced Music. 136th Audio Engineering Society Convention, Berlin, April 2014.

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

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

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

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

[9] 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

[10] 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

[11] S.Fenton, B.Fazenda and J.Wakefield, "Objective Measurement Of Music Quality using Inter-Band Relationship Analysis", AES 130 Convention Paper, London, May 2011.

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

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

[14] E.Skoveborg, "Measures Of Microdynamics", AES 137 Convention Paper, Los Angeles, October 2014.

[15] L.Daudet, "A Review Of Techniques For The Extraction Of Transients in Musical Signals", Computer Music Modelling and Retrieval Conference, Italy, 2005.

[16] J.Bello et al. "A Tutorial On Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing, 2005.

[17] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complemen- tary diffusion on spectrogram," in *Proc. of the EU- SIPCO 2008*, Lausanne, Switzerland, Aug. 2008.

[18] N. Laurenti, G. De Poli, and D. Montagner, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb. 2007.

[19] J.Driedger, M.Muller and S.Disch, "Extending Harmonic-Percussive Separation Of Audio Signals" ISMIR, 2014.

[20] R.Learned and A.Willsky, "A Wavelet Packet Approach To Transient Signal Classification", Applied and Computational Harmonic Analysis 2, pp. 265-278, 1995.

[21] B.Moore, "An Introduction To The Psychology of Hearing", pp.138-145, 5th Edition, Elsevier, 2004.

[22] A.Klapuri, "Sound Onset Detection By Applying Psychoacoustic Knowledge", Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP-99, page 115-118. Phoenix, AZ, 1999.