AN AUTOMATED EVENT ANALYSIS SYSTEM WITH COMPOSITIONAL APPLICATIONS

Nick Collins
Centre for Music and Science
Faculty of Music, University of Cambridge, 11 West Road, Cambridge, CB3 9DP, UK
http://www.cus.cam.ac.uk/ nc272/

ABSTRACT

A modular system for event analysis is described which utilises psychoacoustically motivated onset detection to segment a musical audio signal. The target events have duration above the grain level in the 100-500mS range. Captured events are further analysed for features of pitch, integrated loudness and perceptual attack time. As a pragmatic approach to polyphonic audio, heuristics are specified to select and reject events meeting certain criteria based in statistical moments of instantaneous loudness designed to eradicate double hits and other unbalanced sound events. This system has been applied in both non-realtime composition within the MATLAB environment, and in a real-time form for interactive music via extension UGens and classes for SuperCollider.

Keywords: Event Segmentation, Onset Detection, Psychoacoustic Feature Extraction

1. INTRODUCTION

Event detection and analysis has exciting applications in composition, both in the non-realtime (NRT) case where a database of sound events can be automatically generated to form the source material [21, 23, 9], and in the realtime case where information is extracted on-the-fly [4, 2].

The pertinent time scale of the events sought in this paper has been called the *sound object*[19], *continuation* [27] and *note/phone* [21]. Such rhythmic rate (1-15Hz) events, typically of duration 70-500mS, are a step up from Roads' standard grain durations of 10-100mS, in that they should allow a true temporal integration of their energy rather than an impulse-like percept.

It would be most useful to extract events perceived as fundamental units by a human listener, and so, psychoacoustic features will inform the detection functions. Their percept as singular entities may depend on stable pitch and timbre percepts and a loudness envelope of natural shape. Slow modulations in frequency (vibrato) or amplitude (tremolo), however, may be factored out as event boundary cues [21].

As Scheirer has noted[22] a human observer may understand a signal without an explicit segmentation. Whilst marking the presence of perceptually detectable events would be compatible with this view, the physical extraction and reuse of events is a novel application of technology beyond traditional auditory cognition. There is no guarantee that a perfect solution exists; the best segmentation against which to test this might be defined as the compromise selected by a human user of a sound editing program. A monophonic source should be amenable to segmentation, though even here there are problems caused by the flow of vowels into consonants[12], and of the smooth concatenation of musical events in a legato phrase[20]. In polyphonic audio, events from distinct instrumental sources will overlap. A pragmatic approach to tackle this situation is followed in this paper. Where an extracted event contains obvious rhythmic content within its scope due to 'double hits' heuristics can weed out this event as unsuitable for addition to the discovered events database, or in need of further processing.

Labelling of audio based on some set of features and the use of a database of such information under certain compositional constraints of continuity forms the basis of 'audio mosaicing' or 'concatenative synthesis'[23, 13, 25]. Whilst concatenative systems at present usually deal with individual FFT frames (thus operating at a granular level) and the database of these frames is searched for closest matches to a target feature set, the event chunks can of course be much larger. The database construction and access in NRT composition or live performance may also provide useful techniques for those taking the database part of this work further. NRT MATLab implementations of concatenative sound synthesis have been made by Schwarz and Sturm[23, 25]. Jehan[9] demonstrates a general system with event segmentation capabilities based on loudness, pitch and timbre from a Bark scale frequency band frontend. Lazier presents a real-time model[13].

2. EVENT DETECTION

Sound events are tagged using some form of onset detection; the exact algorithm may be selected for different compositional needs. It is common in research to separate a detection function which gives some measure of evidence for likely onset positions, from the peak picking function which selects the onsets themselves[1].

2.1. A psychoacoustically motivated detection function

Klapuri[10] introduced a detection function based on intensity change discrimination, utilising the log difference of spectral power in bands. Variations of this method along with other detection functions from the literature were compared with respect to onset detection performance on a large expert annotated database; the full details are given in an earlier paper[5]. The best performing detection function used Klapuri's difference of log power on ERB scale bands, where power was converted to decibels and weighted by ISO2003 equal loudness contours. Contour corrected ERB scale band signals also form an input to the loudness estimation and perceptual attack time models presented in subsequent sections.

Whilst this onset detection function is optimised for percussive transients, its performance on slow attacks and especially pitch percept dominated note events, for instance, legato cello

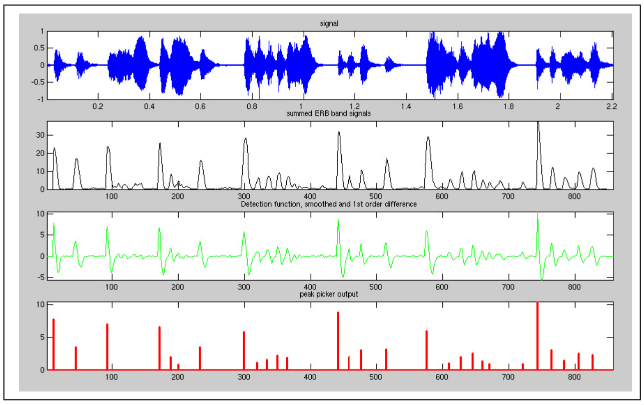

Figure 1. Original signal, ERB band detection function, smoothed and differenced, peak picker output

playing, is non-optimal[5]. It remains however an accurate and efficient solution for its intended domain.

2.2. Peak picking the detection function

The critical effect of the peak picking algorithm operating on a given onset detection function has been highlighted by Juan Bello and co-authors[1], who introduced an adaptive threshold peak picker. Some alternative peak pickers were assessed empirically, and whilst none outperformed the adaptive threshold peak picker on the evaluation test set, competitive performance was obtained using a simpler fixed threshold algorithm.

The steps in this algorithm are enumerated below and an example appears in figure 1.

1. Smooth the detection function with the FIR filter given by coefficents [0.378, 0.27, 0.162, 0.108, 0.054, 0.028];
2. Take the first order difference
3. For all frames i=1 to N
   (a) Require maxima: If ((df(i) $<$ df(i-1) OR (df(i) $<$ df(i+1))) df(i)=0;
   (b) Ignore values below some noise floor: If (df(i) $<$ 0.8) df(i)=0;
   (c) If there is a greater value of df within 10 frames either side, df(i)=0

The peak picker works online with a delay of 10 frames to account for the scoring with respect to previous and future detection function values. Thus it is not suitable to an 'as fast as possible' onset detection but works with a cognitively plausible processing delay for event spotting.

2.3. Event Extraction

Given an onset detection procedure, offsets can be selected based on the criterion that an event be in a required duration range, that the loudness does not fall below some threshold relative to the peak, and that no new onset is detected. Any fail of these conditions signifies an offset position. This is the natural procedure suggested by Smith a decade ago[24]. A zero crossing correction is applied to minimise clicks. A further small envelope at onset and offset may be applied as a further precaution against clicks in resynthesis of the events, though this was often unnecessary in practise. Note that only one event is extracted at a time, and whilst the event boundaries could be allowed to overlap slightly, true polyphonic extraction is not attempted. The great difficulties in resynthesising independent sound streams from ambiguously overlapping spectra should be apparent.

3. FEATURES DETECTED

Three main perceptual features of an event are extracted, and form primary attributes to catalogue events in the database. Additional features to these were explored and included Parncutt's notion of salience[17] and the statistical features detailed in section 4 below.

3.1. Loudness percept

The 40 ERB scale band phon loudnesses can be summed to form an instantaneous loudness function. This may be integrated over time (frames) to make a loudness percept. Jehan[9] dealt with this in terms of spectral and forwards temporal masking. Whilst more complex loudness models have also been implemented, a useful first approximation feature for the comparison of events was found by considering the loudness during the attack stage of the event as a weighted sum of the first 17 frames of instantaneous loudness:

$$\text{loudness}(n) = 10\log_{10}\left(\sum_{j=0}^{39} 10^{0.1*E_n(j)}\right) \tag{1}$$

$$\text{attack percept} = \frac{\sum_{n=1}^{17}(18-n)*\text{loudness}(n)}{153} \tag{2}$$

Where the event starts at frame 1. The calculation of the attack percept weights earlier frames more than later with an additive series, favouring fast build-ups of energy. The number 17 corresponds to a 200mS integration limit for the chosen FFT (44100/512= 86.1328 frames per second. 200mS corresponds to 0.2*86 or about 17 frames), consistent with psychoacoustic models of loudness[16, 7].

3.2. Pitch percept

Just as many onset detection models can be selected for the segmentation, so too many published pitch detection algorithms can be imported. Whilst this attribute is most easily managed for monophonic instrument tones, primary pitches in polyphonic audio may be extractable, for instance by a spectral component analysis[11].

In prototyping, various models were implemented including Klapuri's aforementioned work[11], autocorrelation methods[6], and a related FFT of FFT transform[14]. The most successful model, however, and the one adopted, was the Brown/Puckette constant Q transform on a quartertone scale with phase corrected frequency analysis[3]. Figure 2 demonstrates the output of this pitch detection, showing tracks for the basic quartertone scale detection by spectral template, and the fine tuning of the instantaneous frequency correction. A power envelope was used to turn the pitch detector on or off for near silences, to avoid wild estimates during such times.

3.3. Perceptual attack time

Not all events are percussive. Slow attack envelopes may shift the perceived onset time later into the physical event. Even with a percussive transient attack, it takes time for the temporal integration of the signal energy to trigger a cognitive detection of the event. This leads one to suspect that the perceptual onset

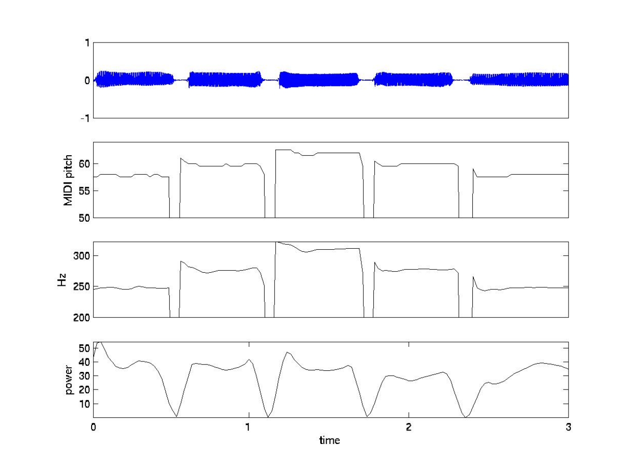

Figure 2. Pitch detection, with the second line showing the curve for a quartertone resolution recognition based on a constant Q transform, the third line an instantaneous frequency correction. The fourth is the power envelope used to turn the pitch detection off for near silences.

rather than the physical may provide a useful feature of signals, and in particular assist with accurate scheduling of a sequence of events, with regard to synchonising onset times within the stream, and with respect to external time points. The study of *p-center* grew out of work on automatic analysis of prosody in the speech processing literature and has been termed *perceptual onset time*[26] or *perceptual attack time*(PAT)[8] in experimental work on instrumental tones.

Predicting the PAT allows the early scheduling of the playback of events so as to 'sound' at a desired time point. Particularly for slow rising tones, naive scheduling may lead to the perception of the tone occuring after a desired entry point. Knowledge of the attack portion of the perceptual envelope also allows a further parameter for the classification of events in our database.

For this project a number of models of PAT were investigated. Vos and Rasch[26] promote a model where the PAT is determined by some proportion between initial power and peak power in the attack envelope. Gordon [8] explores a greater variety of options, discarding the Vos and Rasch *percent of max* model, with the most successful model on his data being one which takes into account the time the power envelope slope is above a threshold, which he calls the *rise time*.

Finding difficulties with these musical tone models, particularly for polyphonic audio, a more complicated model due to Pompino-Marschall from the speech processing literature was adapted, which leverages the ERB scale band energy signals[18]. A bandwise approach considers the peaks in the envelopes as influencing the position of the perceptual attack point. The full algorithm is not reproduced here. In the adaptation, the band wise signals are the ERB band loudness values calculated as above.

In practise, there was some difficulty in getting any model of PAT to work on general polyphonic audio signals. A constant 20mS rule for the PAT was useful as a first approximation, especially for percussive sources. Psychoacoustic experiments are underway to investigate this factor further.

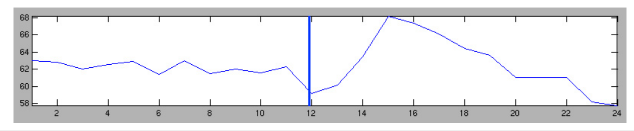

Figure 3. Loudness curve in phons against frame of a 'bad' event, expectation marked as a vertical line

4. HEURISTICS FOR EVENT SELECTION

In order to assess the usability of captured events, a number of heuristics were devised. These utilise the first four statistical moments of the loudness envelope (1), being expectation, variance, skewness and kurtosis, and are passed to the database as attributes of the event. The loudness curve over the frames of the event is normalised into a discrete distribution by subtracting the minimum value and dividing by the sum of all values. This normalisation step is convenient for comparing the envelope of different events varying in dynamic.

Four heuristic rules to determine 'good' or 'well-behaved' events were established empirically. Flags allow them to be turned on or off in a particular application, and the constants mentioned in the rules are really parameters, set here at effective values found in trials. The rules are expressed as conditions which if passed, mark an event 'misbehaved'.

1. (LENGTH) event length not within 100 mS to 1500 mS range
2. (HEAVY WEIGHTING) expectation $>$ 11.2 frames (130 mS)
3. (SKEW) skewness < 0
4. (SECONDPEAK) Find maximum loudness FIRSTMAX in first 45% of the event Find a second maximum SECONDMAX at least 60mS after the first up to the end of the event If SECONDMAX exists and the difference FIRSTMAX- SECONDMAX < 1 (phon)

Failure of length is just a simple test condition to make sure onsets don't occur too often. Some experimentation has taken place with changing the threshold of the onset detection based on feedback from the frequency and duration of events detected, but in practise the fixed parameter onset detection above has been good enough for compositional purposes. The tests on expectation and skewness consider cases where the loudness envelope is not a standard attack then longer decay shape. The HEAVY WEIGHTING expectation test penalises events that have too much of a bias to significant loudness later in the event. The SKEWNESS test looks for loudness curves asymmetrically slanted to the left rather than the right. This corresponds to a 'reverse' sound shape, with a large proportion of time spent in attack rather than decay.

The SECONDPEAK test considers problems of double strikes. These occur not only in polyphonic audio, but also with incorrect onset detections on fast event sequences in monophonic instrumental music.

In figure 3, the skewness was -0.2541, obviously skewed to the left. The expectation is 11.9166. This event failed the HEAVY WEIGHTING, SKEW and SECONDPEAK tests. It probably corresponds to a misdetection by the onset detector where a double hit has not been segmented.

5. COMPOSITIONAL APPLICATIONS

The analysis system was developed in MATLAB, in a modular design to test alternative feature algorithms and optimise for given compositional applications. Ready application is found in the automated production of databases for composition, allowing the composer to spend more time on composing with events rather than preparing them. On a given source, MATLab code produced a database in the form of an output text file annotated with event locations in the source soundfiles, pitch and loudness contours, perceptual attack time and loudness attack rating, loudness statistics and salience. This output text file could be loaded into a composition language like SuperCollider[15], in which the actual algorithmic composition work exploiting the discovered events took place. The system was tested in particular on a library of multi-stopping recordings made with an old school violin; tweaking for specific cases is straight forward.

These analysis procedures have also been ported into a real-time system. SuperCollider was the real-time environment chosen, for its efficiency, and the ease of extending it through the writing of new classes of the SuperCollider language, and new C Unit Generator plug-ins for the signal processing. The integration of algorithmic composition with sound synthesis code in the language greatly facilitated use of the database in live performance, where events from a live audio input can be captured and catalogued on-the-fly with a necessary latency of the duration of an event (to determine its boundaries and features). This empowers a number of novel compositional effects, including delay lines that are event sensitive, event based time stretching and order manipulation[9], on-the-fly categorisation effects[4] and any algorithmic reuse of events recorded from a human performer by the database creating computer.

6. FURTHER WORK

Work is ongoing to move as much as possible of the database system to a real-time footing for electronic music performance. There are great dividends to the automatic extraction of 'interesting' events from an acoustic performer, as the source material for a 'sensitive' accompanying computer music part.

There are many more possible features described in the literature, and variants to the heuristics and assumptions on the underlying signal are all directions of further exploration. Events may also be characterised in terms of timbre (perhaps with timbral descriptors found in an unsupervised sense). Whilst instrument recognition remains a topic for extensive future research, a real-time prototype of a drum sound categoriser based on the spectral centroid was presented in an earlier paper[4] and used in performance to analyse a vocal beat boxing input.

The most significant area of exploration is that of computational scene analysis. Further event heuristics could be devised to consider a bandwise loudness envelope view of multiple strikes, where events can be assessed for the likely overlap of multiple timbres within their duration. The combined detection function has abandoned spectral information that might give clues to such activity.

7. CONCLUSIONS

An event analysis system has been described which segments audio based on a psychoacoustically motivated function. Further features were extracted for captured events including pitch, perceptual attack time and loudness statistics. A number of heuristics were introduced to aid pragmatic use of the event capture technology in a compositional setting for the construction of event databases in both non-realtime and live causal performance. The system described in this paper is relatively general and modular, and new onset detection, pitch detection and PAT algorithms can be substituted for those described here. The same model has been successfully applied to gathering events from pop music, monophonic and polyphonic acoustic instruments.

8. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and S. B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 2004.

[2] Paul Brossier, Juan P. Bello, and Mark D. Plumbley. Real-time temporal segmentation of note objects in music signals. In *Proc. Int. Computer Music Conference*, 2004.

[3] Judith C. Brown and Miller S. Puckette. A high-resolution fundamental frequency determination based on phase changes of the Fourier transform. *J. Acoust. Soc. Am.*, 94(2):662–7, 1993.

[4] Nick Collins. On onsets on-the-fly: Real-time event segmentation and categorisation as a compositional effect. In *Sound and Music Computing (SMC04)*, pages 219–24, IRCAM, Paris, October 20-24 2004.

[5] Nick Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *AES Convention 118*, Barcelona, May 28-31 2005.

[6] Patricio de la Cuadra, Aaron Master, and Craig Sapp. Efficient pitch detection techniques for interactive music. In *Proc. Int. Computer Music Conference*, Havana, Cuba, September 2001.

[7] David A. Eddins and David M. Green. Temporal integration and temporal resolution. In Brian C. J. Moore, editor, *Hearing*, pages 207–42. Academic Press, San Diego, CA, 1995.

[8] John W. Gordon. The perceptual attack time of musical tones. *J. Acoust. Soc. Am.*, 82(1):88–105, July 1987.

[9] Tristan Jehan. Event-synchronous music analysis/synthesis. In *Proc. Digital Audio Effects Workshop (DAFx)*, Naples, Italy, October 2004.

[10] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, pages 3089–92, 1999.

[11] Anssi Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6), November 2003.

[12] Reinier W. L. Kortekaas, Dik J. Hermes, and Georg F. Meyer. Vowel-onset detection by vowel-strength measurement, cochlear-nucleus simulation, and multilayer perceptron. *J. Acoust. Soc. Am.*, 99(2):1185–99, February 1996.

[13] Ari Lazier and Perry Cook. Mosievius: Feature driven interactive audio mosaicing. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2003.

[14] Sylvain Marchand. An efficient pitch tracking algorithm using a combination of Fourier transforms. In *Proc. Digital Audio Effects Workshop (DAFx)*, Limerick, Ireland, December 2001.

[15] James McCartney. Rethinking the computer music language: SuperCollider. *Computer Music Journal*, 26(4), 2002.

[16] Brian C. J. Moore, Brian R. Glasberg, and Thomas Baer. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc.*, 45(4):224–40, April 1997.

[17] Richard Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409–64, 1994.

[18] Bernd Pompino-Marschall. On the psychoacoustic nature of the p-center phenomenon. *Journal of Phonetics*, 17:175–92, 1989.

[19] Curtis Roads. *Microsound*. MIT Press, Camb, MA, 2001.

[20] Xavier Rodet. Synthesis and processing of the singing voice. In *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, November 2002.

[21] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle. Automatic characterisation of musical signals: Feature extraction and temporal segmentation. *Journal of New Music Research*, 28(4):281–95, 1999.

[22] Eric D. Scheirer. Towards music understanding without separation: Segmenting music with correlogram comodulation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999.

[23] Diemo Schwarz. New developments in data-driven concatenative sound synthesis. In *Proc. Int. Computer Music Conference*, 2003.

[24] Leslie S. Smith. Sound segmentation using onsets and offsets. *Journal of New Music Research*, 23:11–23, 1994.

[25] Bob Sturm. Concatenative sound synthesis for sound design and electroacoustic composition. In *Proc. Digital Audio Effects Workshop (DAFx)*, 2004.

[26] Joos Vos and Rudolf Rasch. The perceptual onset of musical tones. *Perception and Psychophysics*, 29(4):323–35, 1981.

[27] Trevor Wishart. *Audible Design*. Orpheus the Pantomime Ltd, York, 1994.