An auditory model for the detection of perceptual onsets and beat tracking in singing.

Martin Coath and Susan L. Denham.
Centre for Theoretical and
Computational Neuroscience.
University of Plymouth. UK
mcoath@plymouth.ac.uk

Leigh Smith and Henkjan Honing.
ILLC,
Music Cognition Group.
Universiteit van Amsterdam. Netherlands.
lsmith@science.uva.nl

Amaury Hazan, Piotr Holonowicz and Hendrik Purwins.
Music Technology Group.
Universitat Pompeu Fabra - Barcelona, Spain.
ahazan@iua.upf.edu

Abstract

We describe a biophysically motivated model of auditory salience and present results which show that the derived measure of salience can be used to successfully identify the position of perceptual onsets in a musical stimulus. We evaluate the method using a corpus of unaccompanied freely sung stimuli. We briefly show that perceptual onsets detected by the model are in good agreement with those identified by a combination of state-of-the-art algorithms and manual correction. We show that this continuous measure of salience can be used to track and predict rhythmic structure on the basis of its periodicity, thus avoiding the necessity for *ad hoc* decisions as to if, or when, an event has occurred.

1 Introduction

When listening to auditory stimuli, particularly music, we tend to find certain events perceptually salient; if this were not the case then it is difficult to see how a sense of rhythm could emerge. Such events are often referred to as *perceptual onsets*. This begs the question ‘what is meant by perceptual onsets?’ and challenges us to propose a method by which they might be identified. There are, of course a number of candidate features that may contribute to the greater perceptual salience of one part of the stimulus compared to another; abrupt changes in energy and spectral distribution for example might seem like plausible candidates. Even sophisticated algorithms based on such hypotheses enjoy mixed success, particularly with difficult stimuli such as unaccompanied singing [1]. For research purposes perceptual onsets are annotated manually and it is these judgements that are compared to those of any candidate algorithm.

One thing that emerges from the temporal pattern of perceptual onsets in human listeners is a complex judgement of rhythmic structure at many levels. One of the most important percepts to emerge is the *tactus* which can be thought of as the rhythm with which one would be tempted to ‘clap along with the tune’. In order to quantify the performance of any model that identifies and tracks the tactus the ‘clapping positions’ or *beat markers* also need to be manually annotated.

In previous work [2, 3, 4] we have proposed that the perceptual salience of events might arise out of the population response of an ensemble of spectro-temporal filters, such as those used to describe cortical responses *in vivo*, which are related to properties of formative stimuli during early experience, in particular speech. Here we take the first steps in investigating whether this novel, biophysically motivated approach is suitable for the field of automatic musical processing. The response of the ‘cortical’ part of our model does, we have shown, identify events in speech which form temporally sparse markers, and the pattern of responses within these events can be used to classify the stimuli in a variety of behaviourly relevant ways. Here we employ the same algorithms with some additional processing to show that this approach successfully identifies positions in musical samples which agree well with positions of annotated perceptual onsets.

However, the usefulness of discrete event markers notwithstanding, what emerges from the model is a *continuous* measure of salience. This can be used to mark the position of events only by using *ad hoc* criteria such as a threshold. However we show, using a wavelet decomposition model, that the continous salience variable can be used to derive the tactus without using the perceptual onset positions, and that models of emerging rhythmic perception can be built without any recourse to complex discussions about what constitutes an event or perceptual onset. This approach is similar to, but distinct from, approaches which are based on the envelope of the signal; for example [5].

2 Methods.

2.1 Salience.

Our model of auditory processing was developed in order to investigate the representation and classification of complex sounds [2, 3, 4]. In this section we briefly outline the key aspects of the existing model before describing how it can be applied to the problem of detecting perceptual onsets in music.

Cochlear model.

We start with the waveform of the stimulus, an example is shown in Figure 1(a). The first processing stage is a linear gammatone filter bank and this is followed by half wave rectification and low pass filtering, with cut off frequency 1000 Hz, to simulate the phase locking characteristics of auditory nerve firing. We use 30 filters with centre frequencies (CFs) ranging from 50 to 8000 Hz equally spaced on the ERB scale [6]. The resulting cochleographic representation is illustrated in Figure 1(b). The cochlear model is down-sampled to 1000 Hz.

Transient enhancement.

The second stage simulates the transient responses prevalent in the central auditory system. Transients are calculated as short term increases or decreases in energy independently within each of the 30 channels of the cochlear model using the third order moment of the amplitude distribution within a sliding window [3, 4]. The duration of the window varies with the centre frequency (CF) of the channel and is $\min(0.01, 8/\text{CF}_i)$ seconds where i is the channel number. Therefore, in order to calculate the transient response, a variable duration short term memory of up to 160 ms is required. Both onset and offset transients are found in this way but only the onset transients are used in further processing, see Figure 1(c). To further reduce the computational overhead the output of the transient module is down-sampled to 200 Hz.

Cortical model.

The third stage consists of convolving the onset transient activity with a set of kernels representing cortical filters [3]. These filters are a set of fragments of stimuli chosen to maximize information with respect to a set of formative sounds, in this case speech. The detailed derivation of the cortical filters is beyond the scope of this paper, and is fully described in [2, 3, 4]. In the model there are 303 cortical filters, the properties of which are very similar to the spectro-temporal receptive fields estimated for neurons in primary auditory cortex [7]. An important characteristic of the responses of these filters is that they generate a set of punctate bursts of activity which mark salient events in the ongoing sound, see Figure 1(d). As the results presented herein indicate, these correlate closely with the positions of perceptual onsets annotated in the stimulus corpora. A short term memory of 100 ms is required for the convolution, to match the maximum temporal extent of the kernels that describe the cortical filters.

Finally the rises and falls in energy of the summed cortical response are detected using the third order moment of the amplitude distribution within a sliding window in a similar way to that described above; this constitutes the salience measure and is illustrated in Figure 1(e)(solid line). If the output is to be interpreted as discrete, the threshold is set as a function of the first peak in the salience response output and adjusted dynamically to track the changing saliency levels.

2.2 HFC- based onset detection.

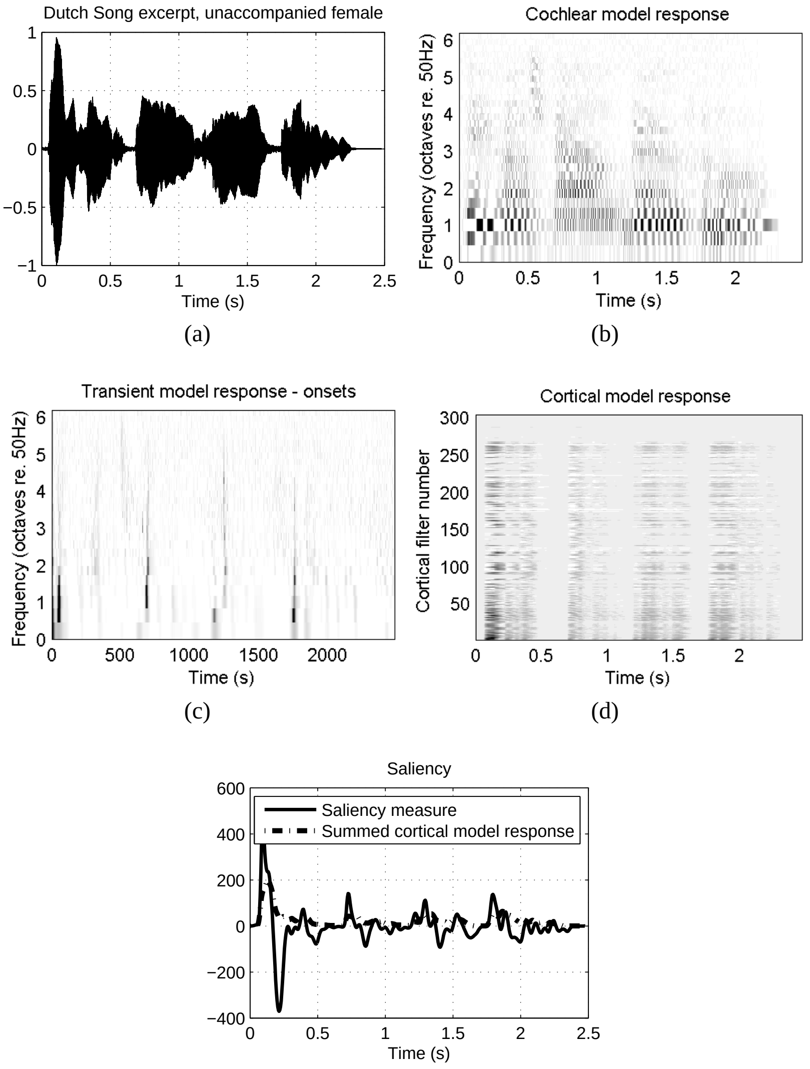
The other acoustic model we use for comparison is the Aubio algorithm proposed by Brossier [8] which is described in detail elsewhere. This implements a number of onset detection algorithms including that used here; the High-Frequency Content (HFC) [9], which is a linear weighting corresponding to bin frequency of the Short Time Fourier Transform (STFT) frame to emphasise the high frequency energy within the signal. As pointed out by Brossier [8], the HFC detection function emphasises the higher part of the spectrum, usually associated with percussive sounds, while it is less responsive to soft onsets such as bowed sounds or flute attacks.

2.3 The Wavelet Transform.

Multi-resolution representations of rhythm have been demonstrated to reveal periodicities in the temporal structure of onsets [10, 11, 12]. The continuous wavelet transform (CWT) [13] decomposes a time varying signal using scaled and translated versions of a *mother-wavelet*. The geometric scaling gives the wavelet transform a ‘zooming’ capability over a logarithmic frequency range, such that high frequencies are localized by the window over short time scales, and low frequencies are localized over longer time scales. For a discrete implementation, each wavelet is a scaled and translated instance from a bank of constant relative bandwidth filters. A sufficient density of scales or ‘voices’ per octave is required (16 in this application) for discrimination of expressive timing. Morlet and Grossmann’s mother-wavelet [14] is used in this application, being a scaled complex Gabor function,

$$g(t) = e^{-t^2/2} \cdot e^{i2\pi\omega_0 t} \tag{1}$$

where ω_0 is the frequency of the mother-wavelet before it is scaled, $\omega_0 = 6.2$ in this application. The Gaussian envelope over the complex exponential provides the best possible simultaneous time/frequency localization [14]. This enables short term periodicities contained in the rhythm to be represented in the analysis.



(a)

(b)

(c)

(d)

(e)

Figure 1: (a) sound wave from an excerpt of Dutch folk song; (b) cochlear model response; (c) onset transients; (d) results of convolution with cortical filters; (e) the summed response from (d) (broken line), and the salience measure derived from the summed cortical response (solid line). For explanation see Section 2.1.

The wavelet coefficients at each time and scale can be computed as separate magnitude (*scaleogram*) and phase components, with spectral energy across time defined as time-frequency ridges. When applied to musical rhythm, a time-frequency ridge is an oscillation at a rhythmic frequency, over a period of time, incorporating rubato. Ridges in the scaleogram function as beat periods that are prominent and can, for example, serve as the rate that listeners tap or otherwise attend to a musical rhythm, ie the tactus.

A CWT analysis is used here to identify time varying periodicities in the salience signal described above. While the thresholded discrete perceptual onsets can also be analysed with the CWT, the continuous salience measure also captures rhythmic information expressed other than simply in the event onset, such as the rate of vibrato. The CWT scaleogram is weighted for absolute tempo preference by a Gaussian envelope with a mean matching the spontaneous tempo rate of 0.6 seconds [15] and a standard deviation of one rhythmic octave (i.e. doubling or halving the beat rate). An integrating auditory store amasses evidence as to the most prominent rhythmic ridge corresponding to the tactus. The frequency of this ridge will vary over time as the rhythm unfolds. In combination with the analysis phase, a rhythmic oscillator can be computed which can be used to clap to accompany the original rhythm at the tactus rate [16]. The clapping is locked to the phase of the first large peak in the salience.

2.4 Evaluations against annotated corpora.

Unaccompanied and freely sung stimuli are not those which would typically be chosen for evaluating systems that identify perceptual onsets or beats. This is one of the hard problems in automatic processing of music. Although such stimuli do not contain percussive elements they do have a rhythmic structure that is clear to listeners even if it is not marked by large, or abrupt changes in amplitude or spectral contour but is marked out by salient parts of the stimuli.

The first corpus used for evaluation consisted of 94 melodies sung, without words, to imitate the sound of a saxophone [17]. These stimuli are referred to here as the SUNG-SAX corpus. Each melody occurs twice, in slow and fast style. The perceptual onsets in these recordings were automatically annotated and then checked and adjusted manually.

We also report results from a small corpus of six freely sung unaccompanied folk songs referred to here as the SUNG-FOLK corpus. These represent a yet greater challenge as some exhibit beat omissions and inconstant tempi. Three are Austrian folk songs from the Essen collection recorded at the MTG. The remaining three are Dutch folk songs from the collection of the Meertens institute in Amsterdam. These stimuli have been annotated by one or more of the authors who are experienced musicians. Each stimulus requires two sets of annotations; one for onset detection and one for beat positions. An example from the second corpus is shown in Figure 3, the waveform is overlaid with the cortical salience response and stem markers showing the annotated positions (diamonds) and the positions of the events identified by thresholding the salience (stars).

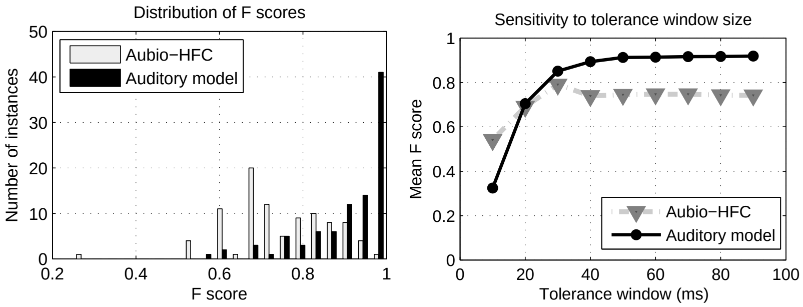
To assess the performance, detected onsets and tactus beat positions from the salience signal and the CWT model are compared with the annotated values. If a detected onset or beat falls within some tolerance window of an annotated event, typically ±50 ms, then it is considered to be correct, otherwise it is considered to be an error. A distinction is made between precision P, the number of correct detections as a proportion of all detections, and recall R, the number of annotated onsets which were correctly detected. Clearly it is desirable to maximize both, and so a combined measure the F-score is computed, where $F = (2 \times P \times R)/(P + R)$ [18].

3 Results.

3.1 Perceptual onsets using singing samples.

The first set of results are obtained from the SUNG-SAX corpus; Figure 2 summarizes the performance. Figure 2 shows the distribution of F-scores over the entire corpus and, for comparison, the results from the same corpus using the onsets derived from the Aubio-HFC algorithm.

The results in Figure 2 show that the salience signal outperforms the Aubio-HFC algorithm



(a) (b)

Figure 2: (a) distribution of F-scores over the entire SUNG-SAX corpus using 50 ms tolerance window, shown for both methods outlined in the text (Aubio with HFC onset detector, threshold 0.5). (b) Graphs showing the change in the F-scores with a range of tolerance window sizes for both methods.

for this group of stimuli in a 50 ms tolerance window. Results from the two methods are similar for windows of 20-30 ms and the Aubio-HFC algorithm outperforms the auditory salience model with windows less than 20 ms.

The second set of results is derived from the SUNG-FOLK corpus; the F-scores are shown in Table 1. The F-scores are comparable to those obtained using the SUNG-SAX corpus for the Austrian folk songs (a,b,c) and lower (particularly song-d) for the songs from the Netherlands which are sung in a less precise, less formal style.

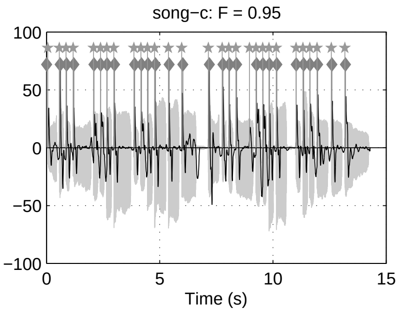


Figure 3: Events in an example stimulus. The waveform is plotted in grey and the salience (cortical response) in black. Peaks in this trace are used to identify perceptual onsets (stars). Annotated onsets are marked with diamonds. The height of the onset stem markers is unrelated to saliency and simply chosen to make the correspondence clear.

Table 1: The F-scores for onset position detection in the six folk songs.

F-scores	song-a	song-b	song-c	song-d	song-e	song-f
Auditory model	0.921	1.000	0.950	0.343	0.500	0.609
Aubio-HFC	0.909	0.915	0.778	0.417	0.772	0.598

mean F-scores	songs a,b,c	songs d,e,f
Auditory model	0.957	0.484
Aubio-HFC	0.867	0.596

3.2 Beat marking in singing samples.

In this section we present a number of results obtained from the continuous salience output. This is used as input to the CWT algorithm that extracts periodicity and makes predictions of the tactus. For each stimulus in the corpus of folk songs the continuous salience measure was derived as detailed in Section 2.1 and the positions of the tactus beat markers derived using the methods described in Section 2.3. These were then compared to the annotated beat times using the method outlined above. The results are presented here in Table 2 but are best appreciated by listening to the sound files of the original stimuli overlaid with synthesized percussion at the times identified by the model. Representative samples of these results can be downloaded from `http://emcap.iua.upf.es/` the EmCAP project website.

Table 2: The results of the F scores for the tactus events, our annotations against the times from the CWT model.

	song-a	song-b	song-c	song-d	song-e	song-f
F-scores	0.500	0.639	0.394	0.750	0.457	0.141
Precision	0.375	0.479	0.292	0.750	0.444	0.104
Recall	0.750	0.958	0.609	0.750	0.471	0.217

Stimuli a-c show recall scores well above precision scores indicating that the model predicts the tactus beat positions well, but places other beats between these positions. In other words it tends to ‘clap too fast’; for example in song-f which is in 6/8 time the model’s tempo weighting prevents the selection of the correct tactus and instead a quaver beat is selected rather than the more likely dotted crotchet beat. The tuning and application of tempo weighting is a current research task. As previously mentioned, these stimuli (d-f) are sung in a very loose style which would be challenging even for a human listener to clap to on first hearing.

4 Conclusions and Discussion.

This method is novel in that it is inspired by models of the auditory system and tasks such as beat marking in auditory stimuli are not routinely addressed using a biophysically inspired approach. The results presented here are closely related to previous work which has demonstrated that transient peaks in the model cortical response mark events that are salient; ie the pattern of responses in these peaks can be used to classify stimuli in a number of behaviourly important ways [3, 4]. However the emergence of rhythmic structure as a high level percept certainly *is* behaviourly important, at least in humans. We have shown that multiscale models that reveal periodicities in stimuli, from which rhythmic structure might emerge, can be informed by models of cortical processing and need not be complicated by considerations of when a salient event might, or might not have occurred. However, if the position of salient events is required for some other processing, then these too can be identified from the model cortical response.

A key feature of the derivation of the cortical filters, on which these results are based, is that they were derived from speech, and selected on the basis of their responses to speech. In additional work, not reported here, we have observed that these filters perform less well in the detection of perceptual onsets in non-sung musical stimuli. The implication is, clearly, that the population characteristics of the ensemble of filters used can be optimized with respect to certain classes of stimuli. This possibility is currently being explored.

One aspect not explored here is how the onset detection is affected when the stimulus is degraded by noise. In previous work we have shown that the response of the auditory periphery model is itself robust to interference by noise [2, 19] so there is good reason to be optimistic that the salience measure will also be useful in situations where the signal is noisy.

The current version of the auditory pre-processing is fully causal and suitable for real time applications. The beat marking algorithm, which is not formulated in a causal way in the current version, is nonetheless suitable for causal implementation. This modification to a fully causal system would enable not just beat *marking*, but beat *prediction*; that is the system would 'know' when it should next clap before the clap was due.

Acknowledgements

This work is funded by EU Open FET IST-FP6-013123 (EmCAP). Thanks are due to the Meertens Institute for supplying the Dutch folk songs from the 'Onder de groene linde' collection and to Frans Wiering for his help in obtaining them.

References

[1] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions On Audio Speech And Language Processing*, 14(5):1832–44, 2006.

[2] M. Coath. *A Computational Model of Auditory Feature Extraction and Sound Classification.* PhD thesis, Centre for Theoretical and Computational Neuroscience, University of Plymouth, 2005.

[3] M. Coath and S. L. Denham. Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience. *Biol Cybern*, 93(1):22–30, July 2005.

[4] M. Coath, J. M. Brader, S. Fusi, and S. L. Denham. Multiple views of the response of an ensemble of spectro-temporal features support concurrent classification of utterance, prosody, sex and speaker identity. *Network*, 16(2-3):285–300, 2005.

[5] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103:588, 1998.

[6] B. R. Glasberg and B. C. Moore. Derivation of auditory filter shapes from notched noise data. *Hear Res*, 47(1):103–138, 1990.

[7] M. Elhilali, J. B. Fritz, D. J. Klein, J. Z. Simon, and S. A. Shamma. Dynamics of precise spike timing in primary auditory cortex. *J Neurosci*, 24(5):1159–72, Feb 2004.

[8] P. Brossier, J. P. Bello, and M. D. Plumbley. Real-time temporal segmentation of note objects in music signals. *Proceedings of the ICMC.*

[9] P. Masri. *Computer modeling of Sound for Transformation and Synthesis of Musical Signal*. PhD thesis, University of Bristol, 1996.

[10] N. P. M. Todd. The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research*, 23(1):25–70, 1994.

[11] L. M. Smith. Modelling rhythm perception by continuous time-frequency analysis. In *Proceedings of the International Computer Music Conference*, pages 392–5. International Computer Music Association, 1996.

[12] L. M. Smith and P. Kovesi. A continuous time-frequency approach to representing rhythmic strata. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 197–202, Montreal, Quebec, August 1996. Faculty of Music, McGill University.

[13] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998. 577p.

[14] A. Grossmann, R. Kronland-Martinet, and J. Morlet. Reading and understanding continuous wavelet transforms. In J. M. Combes, A. Grossmann, and P. Tchamitchian, editors, *Wavelets*, pages 2–20. Springer-Verlag, Berlin, 1989.

[15] P. Fraisse. Rhythm and tempo. In Diana Deutsch, editor, *The Psychology of Music*, pages 149–80. Academic Press, New York, 1st edition, 1982.

[16] L. M. Smith. *A Multiresolution Time-Frequency Analysis and Interpretation of Musical Rhythm*. PhD thesis, Department of Computer Science, University of Western Australia, July 1999.

[17] J. Janer and E. Maestre. Phonetic-based mappings in voice-driven sound synthesis. In *Proceedings of International Conference on Signal Processing and Multimedia Applications*, 2007.

[18] P. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, Centre for Digital Music, Queen Mary University of London, 2000.

[19] M. Coath and S. L. Denham. The role of transients in auditory processing. *Biosystems*, Nov 2006.