Connection Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713411269

Model cortical responses for the detection of perceptual onsets and beat tracking in singing

Martin Coath [a]; Susan L. Denham [a]; Leigh M. Smith [b]; Henkjan Honing [b]; Amaury Hazan [c]; Piotr Holonowicz [c]; Hendrik Purwins [c]
[a] Centre for Theoretical and Computational Neuroscience, University of Plymouth, UK [b] Music Cognition Group, ILLC, Universiteit van Amsterdam, The Netherlands [c] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

Online Publication Date: 01 June 2009

To cite this Article Coath, Martin, Denham, Susan L., Smith, Leigh M., Honing, Henkjan, Hazan, Amaury, Holonowicz, Piotr and Purwins, Hendrik(2009)'Model cortical responses for the detection of perceptual onsets and beat tracking in singing',Connection Science,21:2,193 — 205

To link to this Article: DOI: 10.1080/09540090902733905

URL: http://dx.doi.org/10.1080/09540090902733905

Connection Science
Vol. 21, Nos. 2–3, June–September 2009, 193–205

Taylor & Francis
Taylor & Francis Group

Model cortical responses for the detection of perceptual onsets and beat tracking in singing

Martin Coath[a]*, Susan L. Denham[a], Leigh M. Smith[b], Henkjan Honing[b], Amaury Hazan[c], Piotr Holonowicz[c] and Hendrik Purwins[c]

[a]*Centre for Theoretical and Computational Neuroscience, University of Plymouth, UK;* [b]*Music Cognition Group, ILLC, Universiteit van Amsterdam, The Netherlands;* [c]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*

We describe a biophysically motivated model of auditory salience based on a model of cortical responses and present results that show that the derived measure of salience can be used to identify the position of perceptual onsets in a musical stimulus successfully. The salience measure is also shown to be useful to track beats and predict rhythmic structure in the stimulus on the basis of its periodicity patterns. We evaluate the method using a corpus of unaccompanied freely sung stimuli and show that the method performs well, in some cases better than state-of-the-art algorithms. These results deserve attention because they are derived from a general model of auditory processing and not an arbitrary model achieving best performance in onset detection or beat-tracking tasks.

Keywords: auditory modelling; cortical modelling; salience; transients; onsets; rhythm; singing

1. Introduction

When listening to a sequence of sounds, particularly music, we tend to find certain events perceptually salient. Even sounds that change continuously tend to be perceived in terms of sequences of events. In previous work (Coath 2005; Coath and Denham 2005; Coath, Brader, Fusi and Denham 2005), we have explored the way in which discrete percepts might be extracted from incoming sounds and developed a model of auditory processing, which generated discrete, temporally sparse bursts of activity in response to a continuous sound input. The activity in the model essentially corresponds to responses across an ensemble of filters that model cortical responses. We also showed that the pattern of responses within these bursts contained information about each event which supported perceptual distinctions such as those people perceive. We have argued that the reason for the success of this approach may be found partly in experimental results that suggest that perception and classification require integration over relatively short time scales, as short as 10 ms, around those parts of the signal that exhibit maximum change (Furui 1986). The model is consistent with the results from animal studies showing that firing rate of primary auditory cortical neurons averaged over short periods is well correlated with the discrimination ability for speech sounds (Engineer et al. 2008).

*Corresponding author. Email: mcoath@plymouth.ac.uk

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540090902733905
http://www.informaworld.com

Analysing the signal locally in time with respect to an ensemble of filters and integrating these responses is equivalent to a saliency map in the temporal domain (Koch and Ullman 1985), which is a time-varying measure that indicates the presence of 'interesting locations in complex scenes' (Koch and Ullman 1985). A similar approach in the spectro-temporal domain has subsequently been proposed and supported by perceptual experiments (Kayser, Petkov, Lippert and Logothetis 2005). But this approach is also consistent with the hypothesis that 'the most promising developments for onset detection schemes lie in the combination of cues from different detection functions, which is most likely the way human perception works' (Bello et al. 2005). It seems then that it is at least possible that the identification of cues that engender rhythm is a task that might be performed as well, or better, by a model of auditory processing as by what one reviewer of this paper described as 'an arbitrary model achieving best performance'.

In the current work, we take the first steps in investigating whether our novel, biophysically motivated approach for the segmentation and identification of salient events in a stimulus is suitable for the field of automatic music processing. We employ the algorithms developed in Coath et al. (2005) and elsewhere with some additional processing to show that this same approach successfully identifies positions in music samples, which agree well with positions of annotated perceptual onsets. The model is evaluated using corpora that consist of unaccompanied singing, regarded as one of the harder problems in onset detection (Kapanci and Pfeffer 2004). We also show that the salience measure generated by the model is useful in beat-tracking tasks.

In general, the extraction of signal salience using models of auditory processing can be seen as one of a family of approaches that utilise a mid-level representation of sounds as the basis for further processing (Davies and Plumbley 2005). In some ways, our approach is similar to that documented by Scheirer (1998). Although we do not explicitly extract the signal envelope within each frequency channel, the transient responses in the model do serve to mark changes in the signal envelope. Another alternative is to use the temporal structure in the output from a bank of pitch trackers to derive information about the stimulus and to use a measure of pitch stability to detect onsets (Scheirer 1997; Collins 2005). The model we propose does not explicitly extract pitch, and hence it might be supposed that it will be insensitive to onsets defined only by pitch changes. However, changes in pitch of more than a very small degree also produce changes in the spectro-temporal representation within the model and therefore these onsets can be detected.

One of the most important percepts to emerge in human auditory perception is the tactus which can be thought of as a regular isochronous pattern (the beat) that is activated while listening to music. This beat, which can easily be tapped or clapped along to, is a central issue in time keeping in music performance. But also for non-experts, the process seems to be fundamental to the processing, coding and appreciation of temporal patterns (Jones 1976). The induced beat carries the perception of tempo and is the basis of temporal coding of temporal patterns. Furthermore, it determines the relative importance of notes in, for example, the melodic and harmonic structure (Desain and Honing 1999). In order to quantify objectively the performance of any model that identifies and tracks the tactus, the 'clapping positions' or the *beat markers* need to be manually annotated.

Here we show that the continuous measure of salience that emerges from our model can be used to mark the timing of events, and we report the results that compare our approach with a reference method. However, the usefulness of discrete event markers notwithstanding, we also show, using a wavelet decomposition model, that the continuous salience variable can be used to derive the tactus without explicitly extracting perceptual onset times and that models of emerging rhythmic perception can be built using model cortical responses without any recourse to complex discussions about what constitutes an event or perceptual onset (Gouyon and Dixon 2005).

2. Methods

The results presented in this paper are derived from two independent sets of ideas: the model of auditory salience (Section 2.1) and the multi-resolution representation of rhythmic structure based on a wavelet transform (Section 2.2). In the following sections, we briefly outline the key aspects of these methods.

2.1. *Model of auditory salience*

Our model of auditory salience was developed in order to investigate the online representation and classification of complex sounds (Coath 2005; Coath and Denham 2005; Coath et al. 2005). An outline representation of this process is shown in Figure 1 and representations of the response at each stage can be seen in Figure 2.

Starting with the waveform of the stimulus, an example of which is shown in Figure 2(a), what follows are three stages describing the cochlear frequency analysis, subcortical transient enhancement leading to onset and offset responses and a 'cortical' processing stage consisting of an ensemble of spectro-temporal filters.

2.1.1. *Cochlear model*

In the first stage, the stimulus is processed by a linear gammatone filter bank which models the process of spectral decomposition found in the cochlea (Slaney 1993). This is followed by half-wave rectification and low-pass filtering, with a cut-off frequency of 1000 Hz, to simulate the phase-locking characteristics of auditory nerve firing. We use 30 cochlear filters with centre frequencies (CFs) ranging from 50 to 8000 Hz equally spaced on the ERB scale (Glasberg and Moore 1990). The resulting cochleographic representation is illustrated in Figure 2(b). The output of the cochlear model is down-sampled to 1000 Hz, which is not only a computational convenience but also reflects the decrease in fine detail (phase locking) in the cochlear response above this frequency (Palmer and Russell 1986) modelled by the low-pass filter.

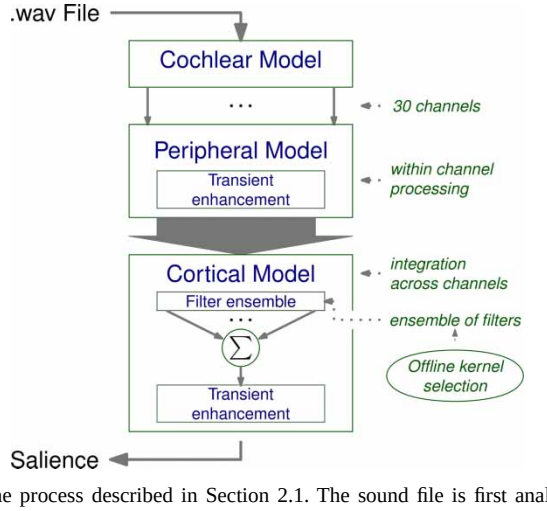

Figure 1. An outline of the process described in Section 2.1. The sound file is first analysed by a cochlear model (Section 2.1.1) and the output from each of the channels passed to a stage, which models the transient sensitivity of the auditory periphery (Section 2.1.2). The third stage consists of convolution with an ensemble of filters that have properties similar to those measured in cortex. The ensemble response is summed and the salience (Section 2.1.3) derived by a further stage of transient enhancement.

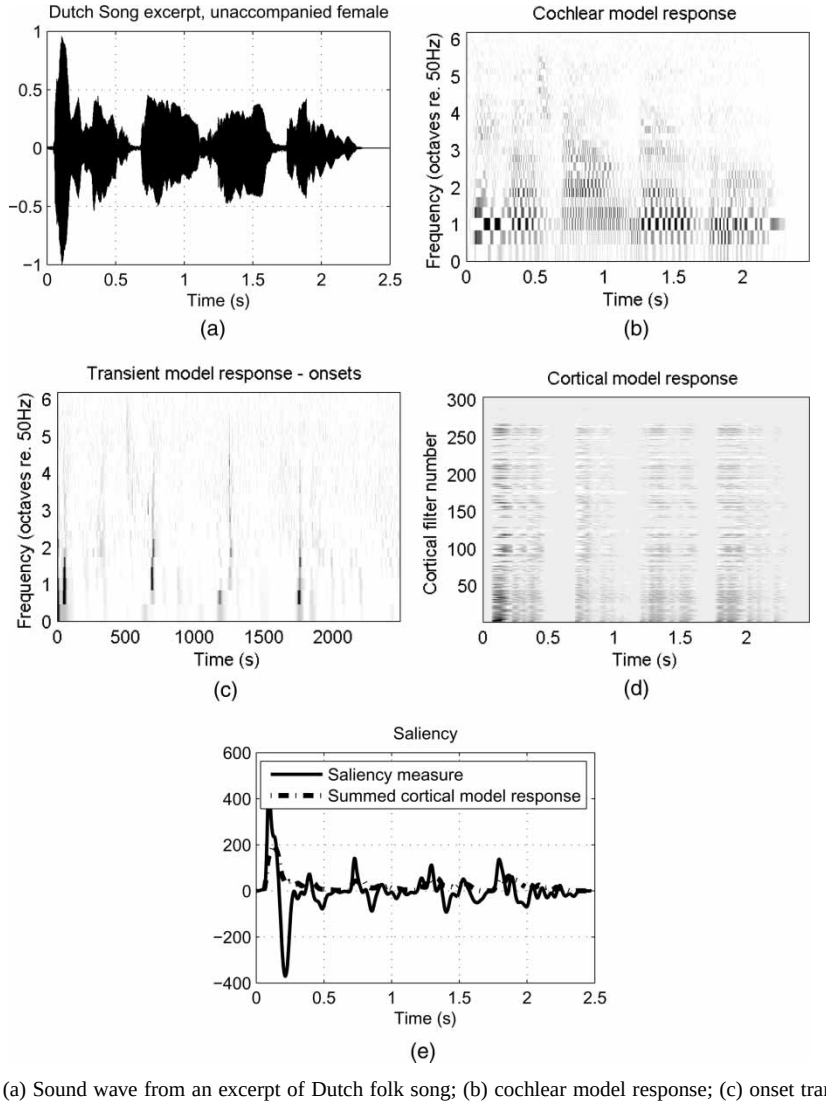

Figure 2. (a) Sound wave from an excerpt of Dutch folk song; (b) cochlear model response; (c) onset transients; (d) results of convolution with cortical filters; (e) the summed response from (d) (broken line), and the salience measure derived from the summed cortical response (solid line). For explanation, see Section 2.1.3.

2.1.2. *Transient enhancement*

The next stage of processing enhances envelope transients within each frequency channel, as is found in the subcortical auditory system. In this model (details in Coath and Denham 2005; Coath et al. 2005), we do not consider the extraction of any other acoustic features. The mean level of activity within each channel is calculated in overlapping temporal windows of duration twice the period of the CF but with a minimum window size of 2.5 ms at high frequency (Wiegrebe 2001). The skewness, or third central moment z, of the distribution of energy across four successive windows is then calculated:

$$z = \frac{1}{N}\sum_{j=1}^{N}\left(\frac{y_j - \overline{y}}{\sigma^3}\right), \tag{1}$$

where N is the number of windows, y_j the energy in the jth window, $\overline{y}$ the mean and σ a variance. Short-term skewness is a sensitive indicator of rising and falling energy and has a value near zero when the energy is approximately unchanging. One advantage of calculating energy distributions within short-time windows is that it essentially gives rise to a rapidly adapting threshold, and hence a roughly level independent representation (Phillips, Hall and Boehnke 2002). Furthermore, the maximum skewness is locked to the onset of the transient and its timing depends upon the dynamics of the onset envelope, as found experimentally (Heil 1997). In effect, this processing amounts to edge detection in the temporal domain and the result is a spectro-temporal map of envelope transients in response to the processed sound. Furthermore, the growth of short-term skewness depends on the dynamics of the transients and varies with the maximum rate of change and acceleration in a way which is consistent with neural behaviour (Heil and Irvine 1997). Specifically, the latency at which the integrated short-term skewness exceeds some threshold can be related to the maximum acceleration (for $\cos^2$ ramps) or to the maximum rate of change (for linear ramps) in a way which is very similar to the first spike latencies measured in auditory cortex (Heil and Irvine 1997).

As mentioned above, the duration of the window for transient enhancement varies with the CF of the channel and is either 0.01 s or eight times the period of the CF whichever is shorter. These values were chosen partly empirically but reflect the evidence that parallel frequency channels are processed on different, frequency-dependent time scales and that, at least for pitch perception, these times are between four and eight times the period (Wiegrebe 2001; Krumbholz, Patterson, Seither-Preisler, Lammertmann and Ltkenhner 2003). Therefore, in order to calculate the transient response, a variable duration, short-term memory of up to 160 ms (i.e. $8^*1/50$ Hz) is required. Both onset and offset transients are found in this way, and both types of responses have been observed in subcortical areas of the auditory pathway, particularly in the mid-brain (Kadner and Berrebi 2008). However, it is not clear what role offset responses might play in the emergence of rhythmic perception and as a result only the onset transients are used in further processing (Figure 2(c)). The output of the transient module is downsampled to 200 Hz; this sampling rate is slow enough to offer significantly lower computational overhead while still capturing the fine detail of speech stimuli, such as the differences in voice onset time necessary for categorical perception of consonants (Sinnott and Adams 1987).

2.1.3. *Cortical model*

The third stage consists of convolving the onset transient activity with an ensemble of kernels representing cortical filters (Coath and Denham 2005). These filters are a set of fragments of stimuli chosen to maximise information with respect to a set of formative sounds, in this case speech. The method employed to select these kernels, the fast correlation-based filter (FCBF) (Yu and Liu 2004), addresses the twin problems of (1) removing both irrelevant and redundant features and (2) reducing the computational overhead of the search in a high-dimensional space. According to Yu and Liu, there are two types of feature-selection algorithm, which they refer to as *feature-weighting* and *subset-search*. The first evaluates the usefulness of individual features. This approach is fast but does not remove redundant features, i.e. it retains features that provide essentially the same information as others already chosen. The second type of search, based on evaluating the usefulness of subsets of features (which is equivalent to an *ensemble*) has a high computational cost. The FCBF algorithm is an attempt to produce a feature-selection algorithm that takes into account the redundancy of features as well as the usefulness of individual features at moderate computational cost.

The FCBF algorithm starts with the kernel whose response is most correlated to the class vector representing a range of representative stimuli, in this case speech sounds. Then it removes

all kernels whose responses are closely correlated to the chosen one, the *redundant peers*. The chosen kernel is designated a *predominant feature*. This is then repeated with the kernel with the next most highly correlated response and so on. With each new choice of predominant feature, a great many redundant peers are eliminated and the algorithm halts when there are no more kernels to be considered. In this way, an ensemble of dominant feature is selected.

The FCBF selection method is not highly parametric but is necessary to fix the minimum correlation value below which a kernel might be regarded as irrelevant. Setting this value too high means that useful features are not considered for inclusion. Setting it too low means that a great many useless features are considered for inclusion, which simply increases the time taken by the process to halt. It is also necessary to set a value for the correlation between two responses above which they might be regarded as redundant.

The size of the resulting ensemble is particularly sensitive to the second parameter. In previous work, we have generated and used an ensemble of 303 kernels, which was the result of just one such value, although we have found that previous classification results were not overly sensitive to ensemble size (Coath 2005).

The properties of the kernels chosen by this process are comparable to those of spectro-temporal receptive fields estimated for neurons in mammalian primary auditory cortex (Elhilali, Fritz, Klein, Simon and Shamma 2004). An important characteristic of the responses of these filters is that they generate a set of well-spaced, brief activity bursts that mark salient events in the ongoing sound (Figure 2(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.

In the final stage of processing, the rise and fall in energy of the summed cortical response are detected using the third-order moment of the response distribution within a sliding window in a way similar to that described above; this constitutes the salience measure and is illustrated in Figure 2(e)(solid line).

2.2. *The wavelet transform*

Multi-resolution representations of rhythm have been demonstrated to reveal periodicities in the temporal structure of onsets (Todd 1994; Smith 1996; Smith and Kovesi 1996; Smith and Honing 2008). The continuous wavelet transform (CWT) (Mallat 1998) 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 localised by the window over short time scales and low frequencies are localised 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 the discrimination of expressive timing. Morlet and Grossmann's mother-wavelet (Grossmann, Kronland-Martinet, and Morlet 1989) is used in this application, being a scaled complex Gabor function,

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

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 localisation (Grossmann et al. 1989). This enables short-term periodicities contained in the rhythm to be represented in the analysis.

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 a musical rhythm, a time–frequency ridge is an oscillation at a rhythmic frequency. Over a period of time, this frequency varies to track the *rubato* in the performance (the expressive variations in tempo). 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, i.e. the tactus.

A CWT analysis is used here to identify time-varying periodicities in the salience signal described above (Section 2.1). While the thresholded discrete perceptual onsets can also be analysed with the CWT, the continuous salience measure 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 s (Fraisse 1982) 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. This ridge, together with the phase component of the analysis, can be used to compute a rhythmic oscillator that can be used to clap to accompany the original rhythm at the tactus rate (Smith and Honing 2008). The clapping is locked to the phase of the first large peak in the salience.

We tackle at length issues of biological plausibility surrounding such approaches in Smith and Honing (2008). In short, biological plausibility can be addressed by delaying the window over time, or by using a causal wavelet, with some cost in time–frequency resolution. Using a non-causal wavelet, kernel with the most theoretically compact time–frequency representation (Grossmann et al. 1989) demonstrates the extent of information that can be revealed by such multi-resolution analyses. This then sets the baseline for comparing other causal wavelets. Although there is evidence of self-organization of Gabor kernels (Kohonen 1995), in contrast to Todd (1994), we do not propose the existence of *in vivo* rhythmotopic neural maps. We propose instead that there are cognitive behaviours that have a functional output that is reflected by the time–frequency representation of our multi-resolution approach.

2.3. *Comparison method for onset detection*

For comparison purposes, we have used the algorithm described in Klapuri (1999). The particular implementation is that of Ricard (2005). For convenience, we will refer to this as the Klapuri algorithm in Section 3. The algorithm uses a cochlear model (from Slaney (1993) – 10 filters from 100 to 10000 Hz) where the amplitude envelope is extracted in each band by full-wave rectification and low-pass filtering, which is then downsampled to 245 Hz to reduce the computation time. The output is then smoothed by a 50 ms half-Hanning window, which preserves sudden changes, but masks rapid modulation (Klapuri 1999). The derivative of the smoothed log-amplitude envelope is used in the detection function. In addition, in order to compute a one-dimensional novelty function, we have computed the sum of the derivatives of the smoothed log-amplitude envelope over all frequency channels. This produces a one-dimensional novelty function, also sampled at 245 Hz.

2.4. *Evaluations corpora*

Unaccompanied and freely sung stimuli are not those that would typically be chosen for evaluating systems that identify perceptual onsets or beats. This is one of the hard problems in the automatic processing of music (Gouyon and Dixon 2005). Such stimuli do not contain percussive elements or other events characterised by large or abrupt changes in amplitude or spectral contour; but they do have a rhythmic structure that is clear to listeners.

The first corpus used for evaluation consisted of a set of melodies, $N_s = 94$ mean duration 4.5 s, sung without words, to imitate the sound of a saxophone (Janer and Maestre 2007). 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 annotated and then checked and adjusted manually. This relatively large corpus of metrically uncomplicated examples serves as a useful preliminary comparison between the proposed model and the reference system. Since each melody is rather brief, we report only perceptual onset detection results for this corpus and limit the tactus evaluations to the more challenging SUNG-FOLK corpus (see below).

The second corpus, for which we report perceptual onset and tactus results, consists of six freely sung unaccompanied folk songs. We will refer to this as the SUNG-FOLK corpus. Three are Austrian folk songs from the Essen collection recorded at the Music Technology Group in Barcelona and the remaining three are Dutch folk songs from the collection of the Meertens Institute in Amsterdam. The latter represent a yet greater challenge as they are sung very freely and include beat omissions and unpredictably variable tempi. These stimuli have been annotated, for both perceptual onsets and tactus positions, by two or more of the authors who are experienced musicians. 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 measure (stars).

To assess the performance of both models, the times of the perceptual onsets or tactus beats are compared with the annotated values. If a detected onset or beat falls within some tolerance window of an annotated event, then it is considered to be correct, otherwise it is considered to be an error. We have examined the results using a range of tolerance window sizes from 10 to 90 ms (Figure 4). 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 as a proportion of all annotated onsets (Van Rijsbergen 1979). Both measures need to be taken into account in the evaluations. A combination of P and R 'punish' an algorithm that detects too many onsets; for example, a continuous series of detected onsets separated by less than the time window within which an onset was judged correct would result in an excellent recall, but its precision would be very low. On the other hand, very few detected onsets, if they

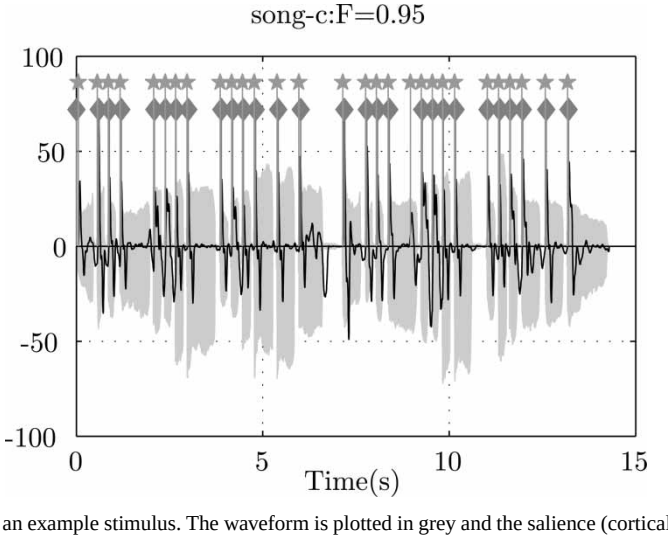

Figure 3. Events in an example stimulus. The waveform is plotted in grey and the salience (cortical response) in black. Peaks in this trace are used to identify onsets (stars). Annotated onsets are marked with diamonds. The height of the onset stem markers is unrelated to saliency and simply chosen to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

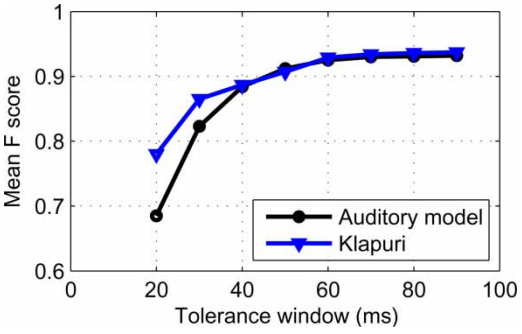

Figure 4. Graph showing the change in the F-scores with a range of tolerance window sizes for the SUNG-SAX corpus ($N_s = 94$ and $N_e = 4936$) for both methods. On this measure, the performance of the two methods is very similar for tolerance windows of $\geq$40 ms

are all correct, would receive a high precision, but the recall value will be low. A good result will exhibit both a high precision and a high recall, and so a combined measure known as the *F-score* is used (Brossier 2006).

$$F = \frac{2 \times P \times R}{P + R}. \tag{3}$$

To aid interpretation of these results, we also show the total number of annotated events N_e, either perceptual onsets or tactus beats as appropriate, in each stimulus or corpus.

3. Results

3.1. *Perceptual onsets using singing samples*

3.1.1. *The SUNG-SAX corpus*

The results for the identification of perceptual onsets obtained from the SUNG-SAX corpus is summarised in Figure 4. This shows the comparison between mean F-scores at a range of tolerance window sizes. Results from the two methods are similar for windows of $\geq$40 ms and the Klapuri detector outperforms the auditory salience model with windows less than 30 ms. The value of 50 ms was chosen for all subsequent evaluations as this is in accordance with evaluations of the MIREX database.

3.1.2. *The SUNG-FOLK corpus*

The results derived from the second small corpus of sung folk songs are shown in Table 1. For the Austrian folk songs (a,b,c), the F-scores are comparable to those obtained using the SUNG-SAX corpus (Figure 4). For the much more difficult Dutch examples (d,e,f: particularly song -d), the results for both methods are much lower, with the Klapuri model scoring better in two out of three cases. This is discussed further in Section 4.

3.2. *Beat marking in singing samples*

In this section, we present the results obtained from the continuous salience output (as detailed in Section 2.1) and from the one-dimensional novelty function derived from the Klapuri method (Section 2.3), both used as inputs to the CWT algorithm (Section 2.2) that extracts periodicity and

Table 1. The F-scores for perceptual onset position detection in the six folk songs using the auditory model and Klapuri algorithm.

F-scores ($N_e = 4936$)	Song a	Song b	Song c	Song d	Song e	Song f	Mean
Auditory model	0.921	1.000	0.950	0.343	0.500	0.609	0.721
Klapuri	0.909	0.915	0.778	0.417	0.772	0.598	0.732
Mean F-scores		Songs a–c			Songs d–f		
Auditory model		0.957			0.484		
Klapuri		0.867			0.596		

Note: Songs a–c are the Austrian examples sung in a clean, rhythmically precise style, and songs d–f are the more difficult Dutch examples.

Table 2. The results of the F-scores, precision and recall for the tactus events in the SUNG-FOLK corpus.

	Song a	Song b	Song c	Song d	Song e	Song f	Mean
F scores ($N_e = 1944$)							
Auditory model	0.500	0.639	0.394	0.750	0.457	0.141	0.480
Klapuri	0.563	0.438	0.269	0.723	0.468	0.042	0.417
Precision							
Auditory model	0.375	0.479	0.292	0.750	0.444	0.104	0.407
Klapuri	0.563	0.467	0.205	0.739	0.458	0.042	0.412
Recall							
Auditory model	0.750	0.958	0.609	0.750	0.471	0.217	0.626
Klapuri	0.563	0.412	0.391	0.708	0.478	0.041	0.423

Note: These are derived from comparison of annotated times with the times from the CWT and Klapuri models. Songs a–c are the Austrian examples sung in a clean, rhythmically precise style, and songs d–f are the more difficult Dutch examples. Note that although overall scores are similar the auditory model has higher recall scores in all but one case.

makes predictions of the tactus. The predicted tactus positions are compared with annotations of the stimulus. The results are presented in Table 2.[1] As previously mentioned, the stimuli (d–f) are sung in a very loose style that would be challenging even for a human listener to clap to on first hearing.

Stimuli a–c show recall scores well above precision scores indicating that the model predicts the tactus beat positions well; there is, however, a tendency, reflected in the low precision scores for both methods, to 'clap too often'; for example, in song f which is in six to eight times the auditory 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. This results in a particularly low precision and recall scores for song f. However, as is audibly apparent, the beat selected for song f results in clapping in tight polyrhythm with matching rubato to the original rhythm. The tuning and application of tempo weighting is a current research task. In general, however, these results show that the auditory model detects the annotated tactus positions in these difficult examples more reliably than the comparison model as is reflected in the higher recall scores.

4. Conclusions and discussion

The method we describe in this paper is novel in that it is inspired by models of the auditory system. 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 punctate bursts in the model cortical response mark events that are salient; i.e. the pattern of responses during these periods can be used to classify stimuli in a number of behaviourally important ways (Coath and Denham 2005, Coath et al. 2005). Importantly,

this approach represents a departure from methods based on the stimulus amplitude envelope, concentrating instead on a wider concept of *change*, which leads to parts of the signal becoming perceptually salient.

In common with speech perception, judgement of speaker sex, prosody of utterance and so on, examined in Coath et al. 2005, the emergence of rhythmic structure as a high level percept certainly is perceptually and behaviourally important, at least in humans. Here we show that it is plausible that the responses that lead to the perception of rhythm are derived from the same underlying mechanisms for identifying the salient, or 'interesting' parts of the stimulus.

Given that there are concerns that methods might be over-fitted to particular data sets, we suggest that it is also possible that methods could be being over-fitted to particular *tasks* and that the wider principles of all behaviour that involves auditory feature extraction might be obscured. In this light, it is a key feature of the method herein described that it was not developed, or optimised, for the present purpose. Nor indeed was it optimised for its original task in any sense other than that the cortical filters were derived from speech and selected on the basis of the entropy of their responses to speech. Likewise, the CWT model of rhythm has been developed and tested on sparse impulse representations (Smith and Honing 2008).

In additional work, not reported here, we have the results that suggest that these filters perform less well in the detection of perceptual onsets in non-sung music stimuli. The implication is, clearly, that the population characteristics of the ensemble of filters used can be optimised 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 peripheral model is itself robust to interference by some types of noise (Coath 2005, Coath and Denham 2005), so there is a 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 given the existence of appropriate short-term memory (160 ms). However, the wavelet decomposition algorithm applies each dilation of the wavelet to a short-term window and if the system is delayed until the lowest frequency (longest time) dilation, then it would react far too late compared with human behaviour. The solution to this problem is likely to be a combination of forward expectation interacting with a retrospection process. Approaches to this are currently being explored. A fully causal system or a system based on a combination of causal and predictive processes 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) and EPSRC EP/C010841/1 (COLAMN). 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.

Note

1. But are best appreciated by listening to the sound files of the original stimuli overlaid with synthesised percussion at the times identified by the model; http://emcap.iua.upf.es/.

References

Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. (2005), 'A Tutorial on Onset Detection in Music Signals', *Speech and Audio Processing, IEEE Transactions on*, 13(5 Part 2), 1035–1047.

Brossier, P. (2006), 'Automatic Annotation of Musical Audio for Interactive Applications', Centre for Digital Music, Queen Mary University of London.
Coath, M. (2005), 'A Computational Model of Auditory Feature Extraction and Sound Classification', Centre for Theoretical and Computational Neuroscience, University of Plymouth.
Coath, M., and Denham, S.L. (2005), 'Robust Sound Classification Through the Representation of Similarity Using Response Fields Derived From Stimuli During Early Experience', *Biological Cybernetics*, 93(1), 22–30.
Coath, M., Brader, J.M., Fusi, S., and Denham, S.L. (2005), '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.
Collins, N. (2005), 'Using a Pitch Detector for Onset Detection', *Proceedings of the ISMIR 2005*.
Davies, M., and Plumbley, M. (2005), 'Comparing Midlevel Representations for Audio Based Beat Tracking', *Proceedings of the DMRN Summer Conference*, pp. 36–41.
Desain, P., and Honing, H. (1999), 'Computational Models of Beat Induction: The Rule-Based Approach', *Journal of New Music Research*, 28(1), 29–42.
Elhilali, M., Fritz, J.B., Klein, D.J., Simon, J.Z., and Shamma, S.A. (2004), 'Dynamics of Precise Spike Timing in Primary Auditory Cortex', *Journal of Neuroscience*, 24(5), 1159–1172.
Engineer, C.T., Perez, C.A., Chen, Y.H., Carraway, R.S., Reed, A.C., Shetake, J.A., Jakkamsetti, V., Chang, K.Q., and Kilgard, M.P. (2008), 'Cortical Activity Patterns Predict Speech Discrimination Ability', *Nature Neuroscience*, 11(5), 603–608.
Fraisse, P. (1982), 'Rhythm and Tempo,' in *The Psychology of Music*, ed. D. Deutsch, New York: Academic Press, pp. 149–180.
Furui, S. (1986), 'On the Role of Spectral Transition for Speech Perception', *The Journal of the Acoustical Society of America*, 80(4), 1016–1025.
Glasberg, B.R., and Moore, B.C. (1990), 'Derivation of Auditory Filter Shapes From Notched Noise Data', *Hearing Research*, 47(1), 103–138.
Gouyon, F., and Dixon, S. (2005), 'A Review of Automatic Rhythm Description Systems', *Computer Music Journal*, 29(1), 34–54.
Grossmann, A., Kronland-Martinet, R., and Morlet, J. (1989), 'Reading and Understanding Continuous Wavelet Transforms', in *Wavelets*, eds. J. Combes, A. Grossmann and P. Tchamitchian, Berlin: Springer-Verlag, pp. 2–20.
Heil, P. (1997), 'Auditory Cortical Onset Responses Revisited. I. First-spike Timing', *Journal of Neurophysiology*, 77(5), 2616–2641.
Heil, P., and Irvine, D. (1997), 'First-Spike Timing of Auditory-nerve Fibers and Comparison With Auditory Cortex', *Journal of Neurophysiology*, 78(5), 2438–2454.
Janer, J., and Maestre, E. (2007), 'Phonetic-based Mappings in Voice-driven Sound Synthesis', in *Proceedings of International Conference on Signal Processing and Multimedia Applications*.
Jones, M. (1976), 'Time, Our Lost Dimension: Toward a New Theory of Perception, Attention, and Memory', *Psychological Review*, 83(5), 323–355.
Kadner, A., and Berrebi, A.S. (2008), 'Encoding of Temporal Features of Auditory Stimuli in the Medial Nucleus of the Trapezoid Body and Superior Paraolivary Nucleus of the Rat', *Neuroscience*, 151(3), 868–887.
Kapanci, E., and Pfeffer, A. (2004), 'A Hierarchical Approach to Onset Detection', in *Proceedings of the International Computer Music Conference*. Available at http://www.eecs.harvard.edu/∼avi/Papers/onset13.pdf
Kayser, C., Petkov, C.I., Lippert, M., and Logothetis, N.K. (2005), 'Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map', *Current Biology*, 15(21), 1943–1947.
Klapuri, A. (1999), 'Sound Onset Detection by Applying Psychoacoustic Knowledge', *Proceedings of the Acoustics, Speech, and Signal Processing, 1999 on 1999 IEEE International Conference*, Washington DC: IEEE Computer Society, p. 6.
Koch, C., and Ullman, S. (1985), 'Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry', *Human Neurobiology*, 4(4), 219–227.
Kohonen, T. (1995), '2', *Emergence of Invariant Feature Detectors in Self-Organization*, New York: IEEE Press, pp. 17–31.
Krumbholz, K., Patterson, R.D., Seither-Preisler, A., Lammertmann, C., and Ltkenhner, B. (2003), 'Neuromagnetic Evidence for a Pitch Processing Center in Heschl's Gyrus', *Cerebral Cortex*, 13(7), 765–772.
Mallat, S. (1998), *A Wavelet Tour of Signal Processing*, San Diego: Academic Press, p. 577.
Palmer, A.R., and Russell, I.J. (1986), 'Phase-locking in the Cochlear Nerve of the Guinea-pig and Its Relation to the Receptor Potential of Inner Hair-cells', *Hearing Research*, 24(1), 1–15.
Phillips, D., Hall, S., and Boehnke, S. (2002), 'Central Auditory Onset Responses, and Temporal Asymmetries in Auditory Perception', *Hearing Research*, 167(1–2), 192–205.
Ricard, J. (2005), 'An Implementation of Multiband Onset Detection', in *MIREX, Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*. Available at http//www.music-ir.org/evaluation/mirex-results/articles/onset/ricard.pdf
Scheirer, E. (1997), 'Pulse Tracking with a Pitch Tracker', in *Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY: IEEE, p. 4.
Scheirer, E.D. (1998), 'Tempo and Beat Analysis of Acoustic Musical Signals', *The Journal of the Acoustical Society of America*, 103, 588.
Sinnott, J.M., and Adams, F.S. (1987), 'Differences in Human and Monkey Sensitivity to Acoustic Cues Underlying Voicing Contrasts', *The Journal of the Acoustical Society America*, 82(5), 1539–1547.

Slaney, M. (1993), 'An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank', Apple Technical Report 35, Technical Report, Apple Computer Inc.

Smith, L.M. (1996), 'Modelling Rhythm Perception by Continuous Time–Frequency Analysis', in *Proceedings of the International Computer Music Conference*, International Computer Music Association, pp. 392–395.

Smith, L.M., and Honing, H. (2008), 'Time–Frequency Representation of Musical Rhythm by Continuous Wavelets', *Journal of Mathematics and Music*, 2(2), 81–97.

Smith, L.M., and Kovesi, P. (1996), 'A Continuous Time–Frequency Approach To Representing Rhythmic Strata', in *Proceedings of the Fourth International Conference on Music Perception and Cognition*, August, Montreal, Quebec: Faculty of Music, McGill University, pp. 197-4-202.

Todd, N.P.M. (1994), 'The Auditory "Primal Sketch": A Multiscale Model of Rhythmic Grouping', *Journal of New Music Research*, 23(1), 25–70.

Van Rijsbergen, C.J. (1979), *Information Retrieval* (2nd edn), Department of Computer Science, University of Glasgow, London: Butterworths.

Wiegrebe, L. (2001), 'Searching for the Time Constant of Neural Pitch Extraction', *The Journal of the Acoustical Society of America*, 109(3), 1082–1091.

Yu, L., and Liu, H. (2004), 'Redundancy Based Feature Selection for Microarray Data', in *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDDD-04), August* 2004, Seattle, pp. 737–742.