

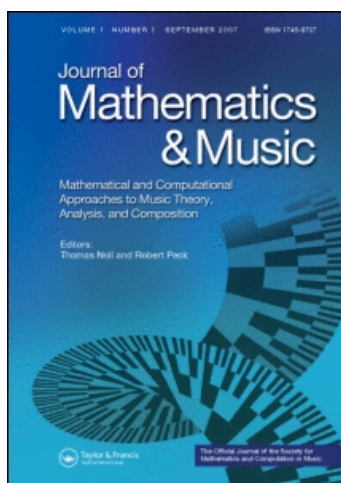
This article was downloaded by: [Ircam]

On: 17 October 2008

Access details: *Access Details: [subscription number 794557483]*

Publisher *Taylor & Francis*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Mathematics and Music

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t741809807>

Time-frequency representation of musical rhythm by continuous wavelets

Leigh M. Smith^a; Henkjan Honing^a

^a Music Cognition Group/ILLC, Universiteit van Amsterdam, The Netherlands

Online Publication Date: 01 July 2008

To cite this Article Smith, Leigh M. and Honing, Henkjan(2008)'Time-frequency representation of musical rhythm by continuous wavelets',*Journal of Mathematics and Music*,2:2,81 — 97

To link to this Article: DOI: 10.1080/17459730802305336

URL: <http://dx.doi.org/10.1080/17459730802305336>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

RESEARCH ARTICLE

Time–frequency representation of musical rhythm by continuous wavelets

Leigh M. Smith* and Henkjan Honing

*Music Cognition Group/ILLC, Universiteit van Amsterdam, Plantage Muidergracht 24,
1018TV, The Netherlands*

A method is described that exhaustively represents the periodicities created by a musical rhythm. The continuous wavelet transform is used to decompose an interval representation of a musical rhythm into a hierarchy of short-term frequencies. This reveals the temporal relationships between events over multiple time-scales, including metrical structure and expressive timing. The analytical method is demonstrated on a number of typical rhythmic examples. It is shown to make explicit periodicities in musical rhythm that correspond to cognitively salient 'rhythmic strata' such as the *tactus*. *Rubato*, including *accelerandos* and *ritardandos*, are represented as temporal modulations of single rhythmic figures, instead of timing noise. These time varying frequency components are termed ridges in the time–frequency plane. The continuous wavelet transform is a general invertible transform and does not exclusively represent rhythmic signals alone. This clarifies the distinction between what perceptual mechanisms a pulse tracker must model, compared to what information *any* pulse induction process is capable of revealing directly from the signal representation of the rhythm. A pulse tracker is consequently modelled as a selection process, choosing the most salient time–frequency ridges to use as the *tactus*. This set of selected ridges is then used to compute an accompaniment rhythm by inverting the wavelet transform of a modified magnitude and original phase back to the time domain.

Keywords: rhythm; rhythmic strata; expressive timing; continuous wavelet transform; time–frequency analysis; beat tracking

MCS/CCS/AMS Classification/CR Category numbers: J.5, H.1.2

1. Introduction

Despite a long history of computational modelling of musical rhythm [1,2], the performance of these models has yet to match human performance. Humans can quickly and accurately interpret rhythmic structure, and can do so very flexibly, for example, they can easily distinguish between rhythmic, tempo and timing changes [3]. What are the representations and relevant features that humans so successfully use to interpret rhythm? We investigate these questions using a computational representation of musical rhythm. This demonstrates how a pattern of time intervals can reveal a structure that informs the understanding of human cognition. We aim to mimic the overall behaviour of human rhythm cognition, as a precursor to future attempts to apportion and separately model each perceptual process involved.

*Corresponding author. Email: lsmith@science.uva.nl

A musical rhythm can be distinguished, memorized and reproduced independently of the music's original pitch and timbre. Even using very short impulse-like clicks, a listener can recognize a familiar rhythm, or comprehend and tap along with an unfamiliar rhythm. The rhythm is thus described from the *inter-onset intervals* (IOI's) between events alone, that is, the temporal structure.

The rhythmic interpretation of those temporal patterns has received considerable research, notable summaries include Fraisse [4], Clarke [5] and London [6]. Yeston [7] characterized musical time as consisting of a hierarchy of time periods spanning individual notes, bars and phrases, terming these periods 'rhythmic strata'. Lerdahl and Jackendoff [8] made the important distinction between the process of grouping of events across time, and the induction of the musical meter that arises from the regular re-occurrence of accented events. These researchers both noted the role of *tactus*, constituting the most prominent musical pulse rate and phase that is induced by the listener. The *tactus* typically appears as the rate and time points that listeners will tap to (typically around 600 ms [4]) when hearing a musical rhythm.

Such attending to a musical rhythm formed from IOIs is proposed by Jones [9] to use two strategies, future-oriented and 'analytic' processes. These processes project expectancies forward in time, and retrospectively assess experienced (i.e. heard) events, respectively. Both strategies are argued to occur simultaneously, and to be judged with respect to the hierarchical time levels established by the rhythm. The perception of a rhythmic pulse and its ongoing attending can be characterized as composed of two resonant processes. These consist of a 'bottom-up' *beat induction* process, and a schematic expectation process providing 'top-down' mediation of the perception of new events [10,11]. Gouyon and Dixon [12] illustrate and characterize computational models according to a similar machine learning-oriented architecture, describing the top-down process as *pulse tracking*.

This tracking task has remained an unsolved research problem, owing in part to the effect of *expressive timing*. Musicians purposefully use expressive timing to emphasize the structural aspects of the piece, such as the metrical and phrase structure [13]. A representation of musical rhythm must therefore account for the establishment of rhythmic hierarchy, the induction of pulse, and the role and effect of expressive timing. In this paper, we describe a representation of musical rhythm using an established transformation, the continuous wavelet transform, applied to the inter-onset timing of events. This representation is demonstrated to make explicit the rhythmic hierarchy and expressive timing. Furthermore, it allows a number of beat tracking methods to be applied to determine the *tactus* implicit in the rhythmic signal.

1.1. *Rhythm signals*

Considerable research has been directed at designing models of both pulse induction and tracking processes towards the final goal of producing useful and robust models of musical time. Existing approaches are reviewed in detail in online Supplement 1. Common problems confronted and addressed in a diverse manner by these approaches are the representation of temporal context, order and hierarchy, and the role of expressive timing and tempo within the existing rhythmic structure.

Since musical rhythm can be induced from mere clicks alone, a rhythmic function for analysis is created by representing the time of each onset as a unit impulse function. The rhythm function for a piece of music is therefore a train of impulses with intervals matching the IOI between onsets. A pulse-train function can be seen to be a minimal, or *critical* sampling of, the auditory amplitude envelope at the lowest sampling frequency that still accurately represents the rhythm function. This yields one sample at the point in time at which each musical event becomes audible to the listener. This is an onset-based representation of rhythm, and is typically recovered by a

detection algorithm operating on the signal energy of the acoustic signal, $|y|^2$. This is effectively a rectification of the audio signal to deconvolve the amplitude envelope from the auditory carrier signal [14,15]. Alternatively, the onset times are obtained by transducing a musician's actions on a sensing instrument (e.g. MIDI). This is distinguished by Gouyon and Dixon [12] from a time-domain frame based system, which aims to determine the rhythmic signal directly from the auditory signal. However in practice, systems must rectify in order to deconvolve the rhythm from the auditory signal.

The use of the continuous wavelet transform as a means of analysing rhythms consisting of an impulse train of onsets was originally reported in Smith [16] and Smith and Kovesi [17]. The output of the transform is similar to Todd's rhythmogram [14], but more detailed. The representation reveals a hierarchy of rhythmic strata and the time of events by using a wavelet that has the best combined frequency and time resolution. Such bottom-up data-oriented approaches, including the multiresolution method described in this paper, do not fully account for human rhythm cognition. Rhythm perception is additionally influenced in a top-down manner by the listener's memory developed by a combination of explicit training and learning through exposure. A goal of this paper is to clarify the information that is inherent (i.e. retrievable) in the temporal structure of a musical rhythm. This aims to establish a base-line measure, to evaluate the contribution of different models of musical time before considering the effect of top-down processing. In short, the intention of this paper is to demonstrate how much structure can be retrieved from a sparse impulse representation of a rhythmic signal.

1.2. Proposed method

The paper is organized as follows: the analytical technique of a continuous wavelet transform is reviewed in Section 2. The application of this transform to musical rhythm to produce a rhythmic hierarchy is described in Section 3. Musical rhythm is described in terms of signal processing theory, and distinguished from the auditory spectrum in Sections 3.1 and 3.2. A simplified schema model is used to determine the tactus (foot tapping rate) in Section 3.3. This extracted tactus is then used to compute a beat tracking rhythm to accompany the original rhythm in Section 3.4. The analysis and representation of rhythms with expressive timing is demonstrated in Section 4.

2. The continuous wavelet transform

Expressive timing, agogic, dynamic and other objective accents produce complex, frequency and amplitude varying rhythmic signals that require a non-stationary signal analysis technique. Analytical wavelets are well suited to this task. The following section is a review of the continuous wavelet transform, further detail is provided in Smith [18].

Wavelet theory has historical roots in the analysis of time varying signals, the principle being to decompose a one-dimensional signal $s(t)$ at time t , into a non-unique two-dimensional time-frequency distribution $W_s(t, f)$, representing frequencies changing over time [19,20].

Earlier signal analysis approaches have used the short-term Fourier transform (STFT), a time windowed version that decomposes a signal into harmonically related *basis functions* composed of weighted complex exponentials. The STFT is

$$F_s(t, f) = \int_{-\infty}^{\infty} s(\tau) \times \bar{h}(\tau - t) \times e^{-i2\pi f\tau} d\tau,$$

where $\bar{h}(t)$ is the complex conjugate of the window function. Significantly, the window function's time scale is independent of f . Any partial of the signal that changes in frequency over the time

extent of the analysis window will have its energy distributed across the spectral domain. This makes short-term changes within the window not resolvable by the STFT.

In contrast, the continuous wavelet transform (CWT) [20–22], decomposes the signal onto scaled and translated versions of a *mother wavelet* or *reproducing kernel* basis $g(t)$,

$$W_s(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(\tau) \times \bar{g}\left(\frac{\tau - b}{a}\right) d\tau, \quad a > 0, \quad (1)$$

where a is the scale parameter and b is the translation parameter. The scale parameter controls the dilation of the window function, effectively stretching the window geometrically over time. The translation parameter centres the window in the time domain. Each of the $W_s(b, a)$ coefficients weight the contribution of each basis function to compose $s(t)$. The geometric scale gives the wavelet transform a ‘zooming’ capability over a logarithmic frequency range, such that high frequencies (small values of a) are localized by the window over short time scales, and low frequencies (large values of a) are dilated over longer time scales [23]. This forms an *influence cone* [24] that has a time interval, for each scale and translation, between $[at_l + b; at_r + b]$ for a mother wavelet with *time support* over the interval $[t_l, t_r]$.

Resynthesis from the transform domain back to the time domain signal is obtained by

$$s(t) = \frac{1}{c_g} \times \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_s(b, a) \times g\left(\frac{t - b}{a}\right) \frac{da db}{a^2}, \quad (2)$$

where the constant c_g is set according to the mother wavelet chosen:

$$c_g = \int_{-\infty}^{\infty} \frac{|\hat{g}(\omega)|^2}{|\omega|} d\omega < \infty, \quad (3)$$

where \hat{g} is the Fourier transform of the mother wavelet.

The CWT indicated in Equation (1) is a scaled filter from a constant relative bandwidth (constant-Q, Equation 3) filter bank. A discrete version of the wavelet transform is used for implementation, so the scale parameter a must be discretized with a sufficient density of filters or v ‘voices’ per octave. The computation of each voice can be performed in the Fourier domain, which can be efficiently computed with the fast Fourier transform, requiring $O(N \log_2 N)$ operations. The number of scales over which the analysis is performed is at most $J = v \log_2(N/K)$ where K is the time support of the wavelet. The complexity of computing the wavelet analysis of the signal over the entire dilation range is therefore $O(JN \log_2 N) = O(vN(\log_2 N)^2)$ [22]. A higher value for v captures finer variations in frequency, but incurs greater computational cost.

2.1. Morlet wavelets

There are many choices for mother wavelets; orthogonal basis functions [20] produce a non-redundant transform for coding and compression applications. However these are unusable for signal analysis, because they do not preserve the phase of the signal, being translation dependent [25]. Grossmann and Morlet [21], have applied a complex-valued Gabor mother wavelet for signal analysis,

$$g(t) = e^{-t^2/2} \times e^{i\omega_0 t}, \quad (4)$$

where ω_0 is the frequency of the mother wavelet (before it is scaled). The frequency parameter $\omega_0 = 6.2$ was determined for this application by calibrating, using an isochronous rhythm of

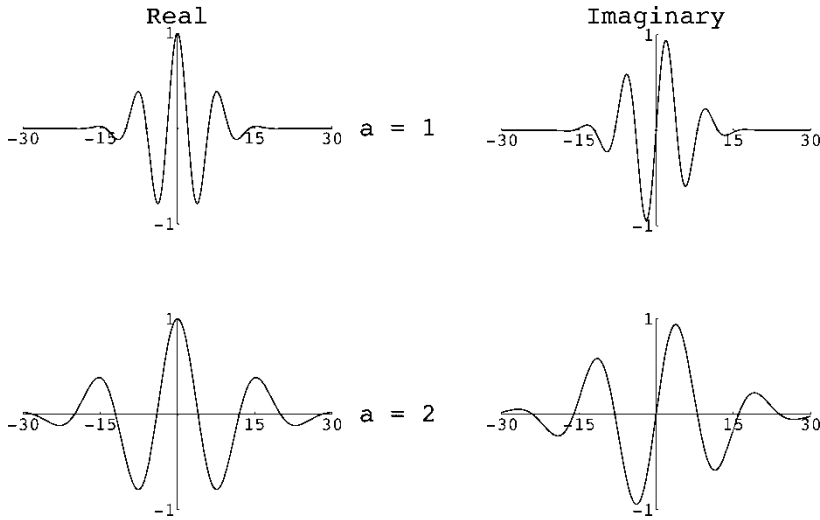


Figure 1. Time domain plots of Morlet wavelet kernels, showing real and imaginary components for the mother wavelet and a version dilated by $a = 2$.

known frequency against the maximum responding scale a for $v = 16$ voices per octave (see [18]). In essence, this is a Gaussian window over cosine and sine curves that are in the real and imaginary planes, respectively (see Figure 1). A Gaussian window function has the property that it is invariant between time and frequency domains, therefore producing the best simultaneous localization in both domains with respect to Heisenberg's uncertainty relation [26, p. 440], [22, p. 33]: $\delta t \times \delta \omega \geq 1/4\pi$.

This led Gabor [27] to propose its use for basis functions that incorporate both time and frequency. Subsequently, Kronland-Martinet, Morlet and Grossmann [23,28,29] applied such a wavelet to sound analysis; however, the research reported here (and in [16,17]) differs from their approach in that it is the rhythm signal (the function that modulates the auditory carrier) that is analysed using so-called Morlet wavelets—not the entire sound signal. Here the rhythm is analysed independently (effectively deconvolved from the sound signal) of the auditory carrier component.

However, the kernel of Equation (4) does not meet the admissibility condition of a zero mean for exact reconstruction [18,22]. The asymptotic tails of the Gaussian distribution envelope must be limited in time such that the residual oscillations will produce a non-zero mean. Given that much analysis can be performed without requiring exact reconstruction, this is not a problem in practice, particularly to the application of musical rhythm analysis. Likewise, the Gaussian envelope renders Equation (4) close to a 'progressive support' or 'analytic' wavelet, nearly satisfying the condition that $\hat{g}(\omega) = 0, \forall \omega < 0$.

Equations (1) and (4) produce complex valued results and, owing to their analytic nature, the real and imaginary components are the Hilbert transform of each other. The conservation of energy of progressive analytical wavelets allows the modulus of a wavelet transform to be interpreted as an energy density localized in the time/scale half-plane. An analytic (progressive) signal $Z_s(t)$ of $s(t)$ can be defined in polar coordinate terms of modulus $A_s(t)$ and phase $\phi_s(t)$ as

$$Z_s(t) = A_s(t)e^{i\phi_s(t)}. \quad (5)$$

The magnitude and phase of the wavelet coefficients $W_s(b, a)$ can then be plotted on a linear time axis and logarithmic scale axis as 'scaleogram' and 'phaseogram' plots (see for example Figure 5), first proposed by Grossmann et al. [30]. The discretized version of the phase of the wavelet

transform, $\Psi(b, a) = \arg[W_s(b, a)]$ (the phasogram), can be recovered owing to the nature of the near-analytic mother wavelet (Equation 4). Phase values are mapped onto a colour wheel or grey-scale to visualize the regularity of the progression of phase. To improve clarity, phase values are clamped to 0, where they correspond to low magnitude values; otherwise, $|W_s(b, a)| > \epsilon_m$, where the magnitude threshold, $\epsilon_m = 0.005$, registers the phase measure as valid to display.

2.2. Ridges

A group of researchers (well summarized by [22, Chapter 4]) have used points of stationary phase derived from wavelet analysis to determine ‘ridges’ that indicate the frequency modulation function of an acoustic signal. These ridges determine the frequency variations over time of the fundamental and a finite number of partials. The chief motivation of this research was to reduce the computation of the transform to only the ridges, collectively termed a ‘skeleton’ [22,31–33]. In that application, the signal analysed was the sampled sound pressure profile.

The motivation here is to extract the frequency modulation function for the purpose of determining a rhythmic partial that corresponds to the tactus. The skeleton is computed from the maximum magnitude $|W_s|$, normalized over scales for each time point b . The peak points $\rho(b, a) = |W_s(b, a)|$, with respect to the dilation scale axis a , are found at

$$\frac{\partial |W_s(b, a)|}{\partial a} = 0, \quad \text{when} \quad \frac{\partial^2 |W_s(b, a)|}{\partial a^2} < 0. \quad (6)$$

An implementation to reorder $\rho(b, a)$ into data structures of individual ridges is described in Section 3.3.

3. A multiresolution time–frequency representation of musical rhythm

3.1. Non-causality

As the time domain plots indicate (Figure 1), the Morlet wavelet is *non-causal*, running forward and backward in time. A causal system is one that depends on past and current inputs only, not future ones [34]. Non-causality implies that the wavelet transformation must be performed on a recorded copy of the entire signal, and is physically unrealizable in real-time. The wavelet is therefore considered in terms of an ideal theoretical analysis kernel, summarizing a number of cognitive processes, rather than one existing in vivo as a listener’s peripheral perceptual mechanism. However, it should be noted that Kohonen [35] has presented evidence for the self-organization of Gabor wavelet transforms; so, such representations are not impossible to realize with a biological system.

However, there are reasons to entertain the idea that the mechanisms used in the process of rhythm induction are not solely dependent on past information alone. Mere exposure to rhythms from previous listening has been shown to construct a schema used to aid perception [36]. The use of temporal context for attentional energy has been argued for rhythm by Jones et al. [9,37], and in terms of pulse sensations by Parncutt [38]. New rhythms are perceived with respect to previously heard rhythms and are organized and anticipated within the harness of a particular schematization. In that sense, the perception of a current beat has an expectancy weighting, projecting from the present into the future, and a retrospection, projecting from the present back into the past.

A purely causal model will be limited in its success because it does not take into account the prediction and retrospection possible during a musical performance. Gouyon and Dixon [12] illustrate the ambiguity of local vs. global tempo changes and timing changes, which can only be resolved by retrospection. Such timing changes are disambiguated over a span of time that may

be considered a moving window. For computational approaches, this requires a representation of a rhythmic schema, which may be considered as abstract structural regularities derived from the music to which the listener is exposed.

The non-causal projection of the Morlet wavelet can be viewed as an idealistic aggregation of such predictive memories. Backwards projection of the filter is a model of completion of an implied rhythm. It functions as retrospective assessment of the rhythm, as argued by Jones and Boltz [9] and Huron [39]. Its use does not seek to apportion rhythm perception behaviour between biological and cultural processes. Clearly, the Morlet wavelet is an oversimplification of the rhythm perception process. Despite the Morlet wavelet being a theoretic formalism, and being a basis for smooth functions, it has several positive attributes as a representation for rhythm analysis.

3.2. Input representation

The onset impulses are weighted by a measure of the phenomenal importance of each event. This summarizes the influence from all forms of phenomenal accent that impinges upon the perception of the event, not only dynamics, and including melodic and harmonic implication. This is denoted by $\iota(t) = c(v) \times \delta(t)$, where $\iota(t)$ is the rhythm function composed of sparse impulse values (0.0–1.0), and $c(v)$ is the normalized phenomenal accent. Minimally, $c(v)$ is the intensity of the amplitude of the onset. This is a simplifying assumption that there is a linear relationship between the perceptual salience of an individual dynamic accent and the intensity of a beat. This ignores the effect of masking of beats by temporal proximity and other nonlinearities between intensity and its final perceptual salience. Masking [40], auditory streaming [41], and expectation (for example, from tonal structure [42] and subjective rhythmization, [4]) can be modelled by a hypothetical nonlinear transfer function $c(v)$. This would summarize the effect of context on the perceptual impact of the note. Alternatively, if a frequency representation is used that preserves energy (by Parseval's relation [34]), such perceptual effects can be modelled in the frequency domain.

The sampling rate can be low (200 Hz) as the audible frequencies are not present. The multiple resolution analysis is therefore performed over the frequencies comprising expressively timed rhythm, spanning from 0.1 to 100 Hz. For analysing human performances, the very shortest scales (less than four samples, 20 ms) do not need to be computed. Hence, the CWT considers several time scales, including those commonly referred to as rhythm and expressive timing [5].

The perception of a polyphonic rhythm (comprising different instruments or sound sources) is assumed to involve segregation into separate simultaneous rhythmic patterns by using common sound features or changes. Where the listener can interpret a rhythm as comprising multiple rhythmic lines, rather than variations in accentuation of a single rhythm, this is assumed to introduce two or more independent rhythms running parallel in time. Furthermore, each is assumed to be analysed separately by parallel processes.

A clearer model of musical time can be constructed in terms of the time–frequency representation of rhythm, rather than strictly in the time domain. The invariance of the Gaussian envelope, between the time and frequency domains of the Morlet wavelet, provides the best simultaneous localization of change in time and frequency. Other kernels will achieve better resolution in one domain at the expense of the other. Arguably, the Morlet wavelet therefore displays the time–frequency components *inherent* in a rhythmic signal, prior to the perceptual processes of the listener. Using such wavelets allows for the quantifying of the representative abilities of other multiresolution approaches to rhythm models.

The explicit representation of multiple periodicities implied by the temporal structure of events can be considered as a pulse induction process that forms the pulse percept across an integrating period. The pulse induction process produces simultaneously ‘attendable’ pulses [9], matching the concept of multiple hypotheses used in beat tracking systems [43–45], but in the case of the

CWT, arising as a direct result of the basis representations (filter impulse responses). This is demonstrated in Section 4. The top down persistent mental framework process (schema) is then responsible for selection from the time–frequency plane of one or more ridges that constitute the most appropriate pulse to which to attend. This is described in Section 3.3.

3.3. Tactus determination

As a minimum demonstration of interpretation of a rhythm, the multiresolution model is used to determine the tactus. This tactus is verified by using it to compute a beat track to accompany the original rhythm. The tactus can be considered to function as the carrier in classical frequency modulation (FM) theory. An isochronous beat is a rhythmic periodicity of a single frequency, and the performer’s rubato constitutes a frequency modulation of this idealized frequency. In performance, the tactus of a rhythm is modulated but still elastically retains semi-periodic behaviour. A means of extracting the rubato frequency modulation (the ridge) of the tactus is required. The schematic diagram of Figure 2 describes this process.

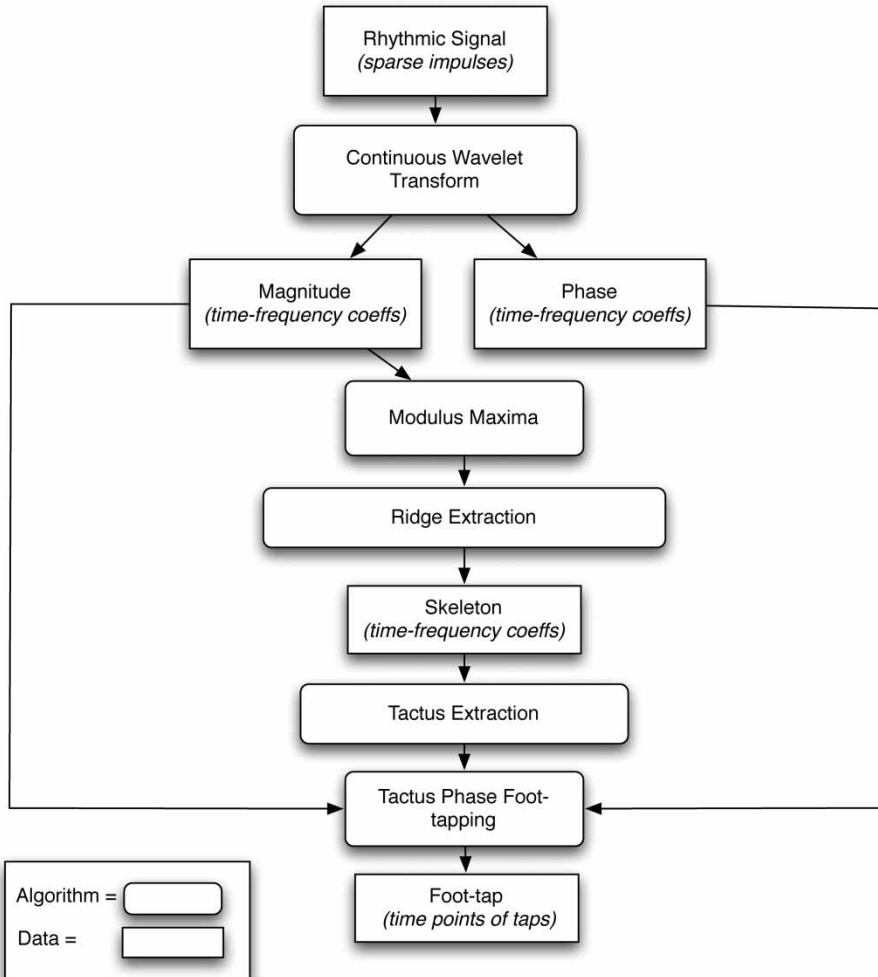


Figure 2. Schematic diagram of the multiresolution rhythm interpretation system.

Downloaded By: [Ircam] At: 11:37 17 October 2008

A process, which can be considered a pulse tracker, extracts the ridge determined to be the tactus. This tactus ridge is then transformed from the time–frequency plane back to the time domain, and its phase measure is used to compute the time points of a beat track.

3.3.1. Ridge extraction

Section 2.2 described the process of combining magnitude and phase measures, and of identifying the peak points $\rho(b, a)$ in the time–frequency plane. These peak points are then grouped by time–frequency proximity into a skeleton of ridges. The skeleton of ridges is implemented as a list of Common Lisp objects, each ridge defining its starting location in time and the scale at each subsequent time point. Representing such ridges as objects (equivalently Minskyian frames) aims to establish a knowledge framework that bridges the divide between symbolic and sub-symbolic representations of rhythm. An algorithm to reorder $\rho(b, a)$ into ridges $\varrho(b) = a$ is shown in Figure 3 and illustrated in Figure 4.

With a tolerance $t = 1$, this extracts ridges with a rate of change of at most one scale per time sample. Since the matching is done by scale number a , $t = 1$ is the minimum. For $t > 1$, the algorithm will accept discontinuities between scales of successive time samples for ridges tracking extremely rapid acceleration or deceleration. In practice however, setting $t > 1$ can also lead to the extraction of a single ridge, rather than extracting two closely parallel ridges. With sufficiently high time and frequency resolution (a sample rate of 200 Hz and voices per octave $v = 16$), $t = 1$ will track acceleration and deceleration for musical rhythms correctly.

Such a skeleton represents alternative accompaniment or attending strategies, and is query-able as such. This representation serves as a foundation to develop, in a modular fashion, alternative pulse tracking or schematic processes. This is to allow schemas to be developed that decouple the model of ridge selection (e.g. pulse tracking) from pulse induction. This decoupling seems

Algorithm: Extract-Ridges

```

1: for  $b \leftarrow 0 \dots B$  ; time extent of analysis window
2:    $S_b = a$  if  $\rho(b, a) > 0$  ; those peaks at each time point  $b$ 
3:    $(S'_b, S'_{b-1}) \leftarrow S_b \cap S_{b-1}$  ; match current peaks and those at previous time point
4:   if  $\varrho_r(b) \in S'_{b-1}$  ; update the existing ridges with scales that match
5:      $\varrho_r(b) \leftarrow S'_b$  ; those at the previous time point
6:    $\varrho_{r+1}(b) \leftarrow S_b - S_{b-1}$  ; create new ridges for scales that don't match

```

The sets $(J', K') \leftarrow J \cap K$ containing those elements of J and K matching within a tolerance t are defined as:

Algorithm: Match-Sets

```

0: ensure  $J$  and  $K$  are both sorted in numerical order
1:  $n \leftarrow 0$  ; the next unmatched index
2: for  $j \in J$ 
3:   for  $k \leftarrow n \dots \text{len}(K)$  ; find where  $j$  is located in  $K$ 
4:     if  $K_k > j + t$ 
5:       exit-for
6:     if  $|j - K_k| \leq t$ 
7:        $J' \leftarrow j \cup J'$  ; collect those elements from  $J$  that match
8:        $K' \leftarrow K_k \cup K'$  ; also collect those elements for  $K$  that match
9:        $n \leftarrow k + 1$ 
10:    exit-for
11:   $n \leftarrow k$  ; remove all earlier elements from future tests

```

Figure 3. Algorithm for the extraction and reordering of peak points ρ in the combined magnitude and phase measures into ridge structures ϱ ordered by time.

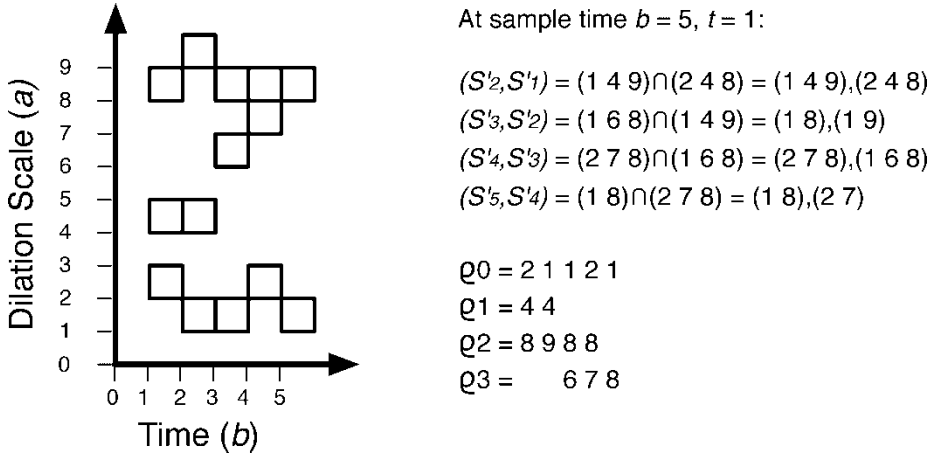


Figure 4. Example operation of the ridge extraction algorithm of Figure 3.

necessary in order adequately to model schematic knowledge, which may contradict the veridical expectancy derived from the surface material. For example, the tactus of a reggae rhythm being at half time to the quaver (eighth note) pulse.

3.3.2. Ridge selection

The simplest schema process for selecting a ridge as tactus is a heuristic that selects the longest, lowest frequency ridge that spans the analysis window. In effect, this can be considered the rhythmic ‘fundamental’ frequency. While such an axiom or principle may seem problematically reductionist, the heuristic proposed describes a broad behaviour that applies to the entire rhythm. This indicates that there is an inherent coherence in the temporal structure of the entire rhythmic sequence analysed. Indeed, the low-frequency components are derived from the contribution of all events analysed. The selection of such a ridge is shown on the phaseograms of Figures 5 to 8.

Other simple schemas are to establish tempo constraints in choosing a ridge as Parncutt [38] adopts with 1.38 Hz, and Todd [46] with 2 Hz band-pass filters. More complex schemas can then be introduced in the future, and compared to this obviously overly simplistic model. The query-able nature of the skeleton also enables manual selection and testing of ridges for evaluation of their role as modulated rhythmic strata by using reconstruction.

3.4. Reconstruction of the tactus amplitude modulation

Once the tactus has been extracted from the rhythm, it can be used to compute tap times. When sampling a tactus that undergoes rubato, it is not sufficient simply to sample the instantaneous frequency of the tactus ridge owing to the accumulation of error. Therefore, the tactus ridge is transformed from the time–frequency domain into an FM sinusoidal beat track signal in the time domain. Only the tactus ridge itself will contribute to the beat track signal. The sinusoidal nature of the resulting signal causes it to act as an amplitude envelope, i.e. as a rhythm frequency that modulates over time. This signal is reconstructed from both the scaleogram and phaseogram coefficients.

All scaleogram coefficients, other than those of the tactus ridge, are clamped to zero, while the original phase is retained. This altered magnitude and the original phase components are

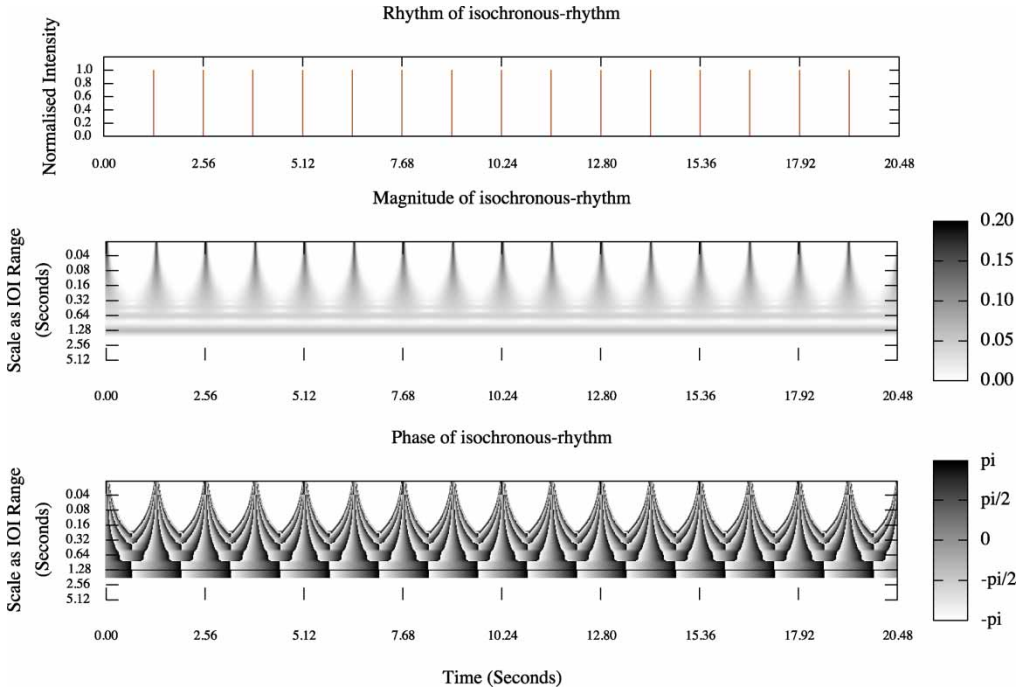


Figure 5. An isochronous rhythm (fixed IOI of 1.28 seconds) shown in the top graph as a time-amplitude plot of impulses, then represented as a scaleogram (magnitude), phaseogram and skeleton of the continuous wavelet transform coefficients. The most activated ridge (candidate for tactus) is centered on the scale corresponding to an interval of 1.28 seconds and is indicated by the horizontal black line on the phaseogram and highlighted line on the skeleton. A lower energy ridge is centered on the interval of 0.64 seconds. This secondary ridge occurs from interactions between secondary lobes of the wavelet.

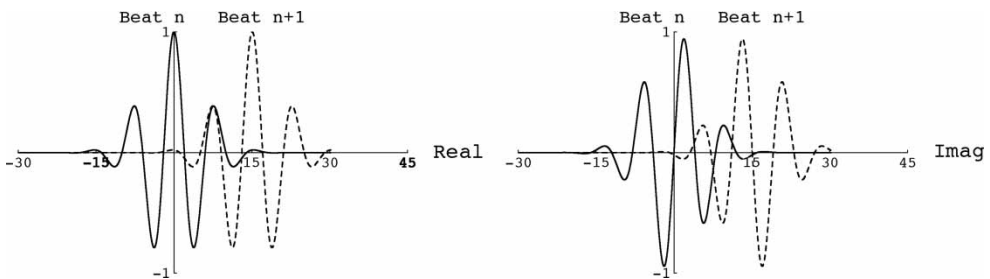


Figure 6. Time domain plots of the overlap of the real and imaginary components of Morlet wavelet kernels. These demonstrate the cause of the reduced energy second harmonic in the scaleogram.

converted back to $W_s(b, a)$ coefficients, and reconstructed back to a time domain signal using Equation (2). The constant $c_g = 1.7$ was determined by calibrating the original time signal with its reconstruction $s(t)$ to achieve energy conservation. Owing to the asymptotic tails of the Gaussian, the reconstruction cannot be perfect, but the reconstruction was determined still accurately to resynthesize the frequency and phase of signals. The real component of the reconstruction,

$$A_s(t) = \Re[s(t)], \tag{7}$$

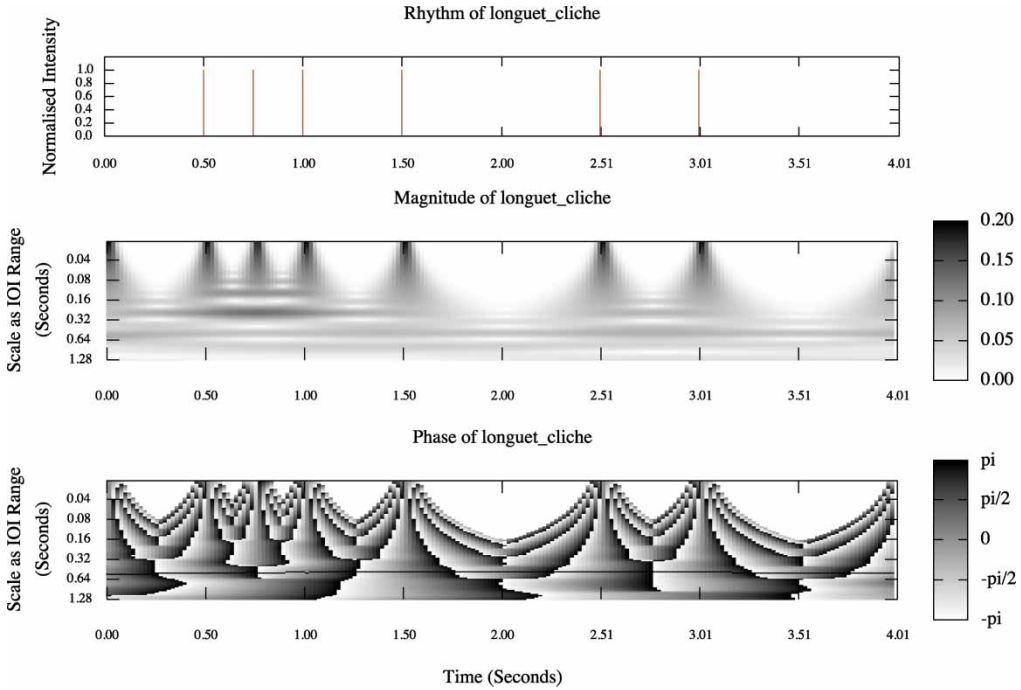


Figure 7. Scaleogram and phaseogram plots of an analysis of the score time of the musical cliché [51]. The multiple periodicities implied by the event times appear as time–frequency ridges. The lowest frequency ridge that extends across the entire analysis window is indicated on the phaseogram by the black line and is visible as energy on the scaleogram.

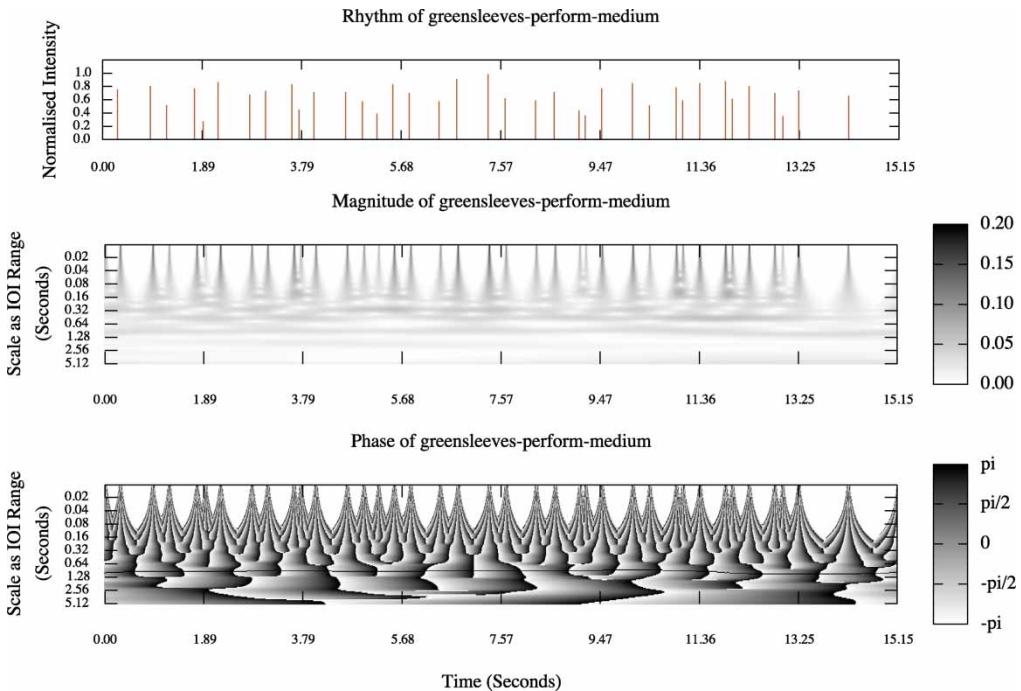


Figure 8. Rhythmic impulse function, scaleogram and phaseogram of a performed version of ‘Greensleeves’ possessing expressive timing modulations from the notated rhythm. The varying amplitudes of each beat are indicated by the density of events at the highest scales on the scaleogram. The tactus of the rhythm (derived using the algorithm described in Section 3.3) is shown as a black line on the phaseogram.

reproduces the sinusoid, while $\Im[s(t)]$ reproduces its analytic counterpart, i.e. phase shifted by $\pi/2$ radians. In addition, the phase of the reconstructed sinusoid can be easily obtained:

$$\phi_s(t) = \arg(\Im[s(t)], \Re[s(t)]). \quad (8)$$

While the peaks of the amplitude modulation (AM) (Equation 7) were verified to produce the correct beat track points for an isochronous tactus from the pulse (Figure 5), which aligned with the original rhythm, problems would arise with rhythms that were phase shifted from the occurrence of an anacrusis. Therefore the $\phi_s(t)$ value was noted for t at the first onset time at which to begin tapping, and the remaining beat times were selected for each $\phi_s(t)$ that matched that initial phase value. These beat times were then used to generate note event times with which to synthesize a beat track that could be mixed with the original rhythm.

This beat track is only synchronized to the original rhythm on one onset, all beats are computed relative to that first chosen onset. Unless human listeners have an existing memory of the rhythm, they will not begin clapping from the first beat, so a future research task is to identify the appropriate first tap beat for each rhythm. Currently, the onset time at which to begin tapping (selecting the phase of the beat track) is chosen to be the second beat for all rhythms.

4. Rhythm examples

4.1. Isochronous rhythm

The wavelet transform produces short-term, high-frequency basis functions for small values of the scaling parameter a and long-term, low-frequency versions for large values of a . Short wavelet basis functions isolate discontinuities in the time domain, while long basis functions analyse with high discrimination in the frequency domain.

An impulse is localized in time, but infinite in frequency content. A CWT of an impulse localizes the impulse's effect in the time domain at the higher-frequency scales (small values of a) and spreads the effect across longer finite time periods at lower scales. Owing to the non-causality of the Morlet wavelet, at each scale and translation of Equation (1) the impulse will be projected simultaneously forward and backward in time in the time–frequency plane, matching the support of the wavelet.

As detailed in [30,47,48] and [28, p. 279], a *singularity* such as an impulse will be marked by a localized increase in the modulus at high frequency scales, and by a constant phase across frequency scales, independent of the mother wavelet used. An analysis of an isochronous train of impulses with a bank of dilated Morlet wavelets is shown in Figure 5. The x -axis represents time in seconds. The y -axis is logarithmic, represented here by the time extent of each wavelet voice, again in seconds, with a scale resolution of $v = 16$ voices per octave. This resolution is sufficient to capture changes in frequency matching expressive timing. For a scale matching 1.28 seconds, the difference in time extent of one voice is $+56/-54$ ms.

The scale with the highest modulus represents energy density, and corresponds to the frequency of the beat. This frequency is the reciprocal of the IOI, as indicated by the horizontal band at 1.28 seconds across the magnitude plot of Figure 5. The timing of the onset intervals between beats will be reflected by the energized scales. The relative energy levels of each scale are indicated in Figure 5. In addition to the most highly activated scale corresponding to an IOI of 1.28 seconds, there is a secondary lobe of half amplitude energy at the first harmonic of the beat rate (0.64 seconds). This is caused by coincidence of the half-amplitude second oscillations of the kernels in the time domain (see Figure 6, which plots Equation 4 for $\omega_0 = 6.2$). The forward time projection of the n th beat will positively add with backward time projection of the $(n + 1)$ th beat at the first

and second oscillations of the kernel, producing energy at the first and second harmonic of the beat rate. These artifacts arise from the Morlet kernel and are dependent on the ω_0 value, more oscillations producing further low-energy harmonics. Therefore a very slight energy at the third harmonic corresponding to 0.426 seconds can be discerned in the magnitude plot and more so in the skeleton plot of Figure 5.

While artifactual in nature, these harmonics can be considered as representing a listener's lower propensity to perceive an isochronous rhythm as actually being at double the rate of the events. From another perspective, second and third harmonics from respective rhythms, at half and one third rates, will contribute to the total signal energy measured at a given rhythmic frequency. This effect occurs in the Morlet wavelet as a consequence of the nature of the Gaussian envelope of the Gabor kernel, which is modelling the Heisenberg inequality of time and frequency representation.

This implies that secondary preferences for doubling or, to a lesser extent, tripling a rhythm is inherent in this model of rhythm, rather than learned. Categorization of rhythmic ratios towards 2:1 in reproduction tasks [4,49,50] does indeed show that these ratios are privileged. Musical performance practice and other activities involving doubling a motor behaviour are easily accomplished by humans. That practice is reflected in the ubiquity of duple and triple representations in musical notation of rhythm. It is quite possible that this motor production optimization is matched by an inherent perceptual process favouring simple subdivisions of time.

4.2. *Non-isochronous rhythm*

The representation of a typical musical rhythm is shown in Figure 7. The resolution of the CWT makes the short-term frequencies of the rhythm apparent. Scales are reinforced from time intervals that overlap, and fade where onsets no longer contribute impulsive energy. This produces a set of simultaneous ridges in the scaleogram as the rhythm progresses across time, and creates different intersecting intervals. The most highly activated ridges are the ones that receive most reinforcing contributions from re-occurring intervals. The phase diagram illustrates the presence of periodicities at individual scales by the steady progression of phase (i.e. the spectrum) across time.

The impulse signal is padded with repetitions of the signal up to the next dyadic length to account for edge conditions so that the last interval is part of the analysed signal. Therefore the entire signal is analysed in the context of its being repeated. This does not compromise the time–frequency resolution of the CWT, since signals are not assumed to be periodic within the analysis window, as is the case for a Fourier representation. An alternative representation is to pad the analysed signal with silence on both ends. This has minimal impact on representations of onsets within the signal, but reduces the contribution of the first and last beats.

Further examples of analyses of rhythms are demonstrated in online Supplement 2.

4.3. *A performed rhythm: Greensleeves*

Figure 8 demonstrates the CWT applied to a performed version of a well-known rhythm example with multiple IOIs grouped in musically typical proportions. The rhythm timing is obtained from a performance played by tapping the rhythm on a single MIDI drum-pad. The scaleogram and phaseogram indicate the hierarchy of frequencies implied at each time point owing to the intervals between notes falling within each scaled kernel's support. Repeated rhythmic figures, such as the shortened semi-quaver (sixteenth note) followed by a crochet (quarter note), produces energy ridges starting at 1.74, 3.6, 9.07 and 10.91 seconds in the scaleogram. These and other timing characteristics produce temporal structure that is revealed in the lower-frequency scales. The phaseogram in Figure 8 indicates higher-frequency periodicities, and the enduring

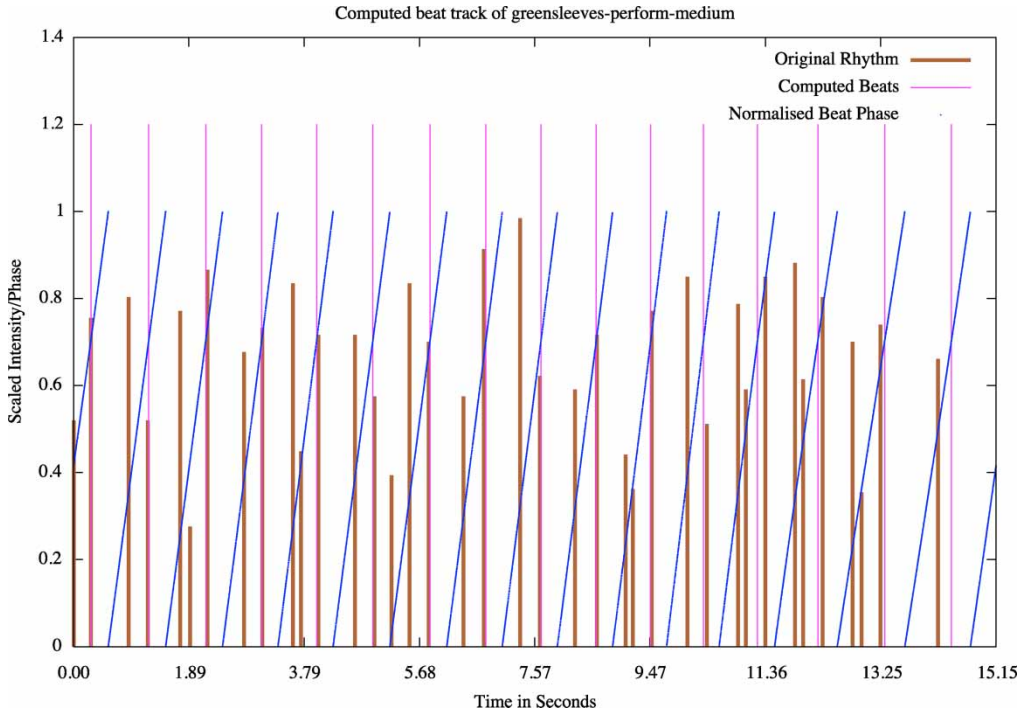


Figure 9. This diagram demonstrates beat tracking to the tactus of the performance of ‘Greensleeves’ analysed in Figure 8. The original time-points of the notes of the rhythm are shown, together with the reconstructed phase $\phi_s(t)$ and the computed beat track determined from phase-locking to the second note, the nominated event to begin tapping to. In this diagram the computed beat track is artificially scaled to a value of 1.2, and the phase (0.0–1.0) to improve readability. Available in colour online.

lower-frequency periodicity with an IOI mostly around 0.91 seconds nominated as the tactus (see Section 3.3). The original rhythm, the computed beat track, and the reconstructed phase $\phi_s(t)$ of Equation (8) are displayed in Figure 9.

5. Conclusion

This paper proposes and demonstrates phase-preserving Morlet wavelets as a means of analysing musical rhythm by revealing the rhythmic strata implicit in a rhythmic signal. The transform represents the rhythmic effects generated by dynamic and temporal accents in establishing hierarchies of rhythmic frequencies. Such a hierarchical representation bears similarities with the metrical and grouping structure theory of Lerdahl and Jackendoff [8]. It does not require explicit generative rules to produce such an interpretation. Hence the proposed method examines the information contained within a musical rhythm before any perceptual or cognitive processing is performed. This method attempts to make explicit the structure inherent in a rhythm signal. It can be viewed as a formalization of rhythmic decomposition.

In addition, the preservation of energy, and therefore invertability of the CWT (Section 2.1), enables cognitive models to be built in the time–frequency domain, as an alternative to purely time domain models. This allows representations to be built that directly address the change over time that musical rhythm undergoes. The degree to which the time–frequency ridges produced by this model matches human cognition suggests the degree to which musical rhythm is an example of emergent cognition, arising from mechanisms that attune to features of environmental stimuli.

Nevertheless, the system still has a number of limitations. The heuristic of choosing a tactus on the basis of the lowest frequency and longest time extent ridge is obviously simplistic. However, it does demonstrate that the CWT can identify the tactus as a time–frequency ridge. Work is currently underway to improve the tactus selection method to include sensitivity to phase and global tempo. Further work is also required to extend the tactus selection to one that is determined by exposure to a musical environment, and to verify this approach with a large test set that uses a wide range of musical rhythms. A systematic evaluation is currently underway to determine the limits of rhythmic signals that can be represented. Finally, the non-causal implementation of the method currently prevents a direct online application, and the frequency resolution used produces a relatively high computational burden.

Despite these current issues, we think the multiresolution analysis model contributes to an understanding of how much information can be obtained from the rhythmic signal itself, including both categorical (temporal structure) and continuous (expressive timing) information. Since it does not use additional ‘top-down’ modelling, it may serve as a baseline for cognitive models of rhythm perception.

Acknowledgements

This research was realized in the context of the EmCAP (Emergent Cognition through Active Perception) project funded by the European Commission (FP6-IST, contract 013123). Thanks go to Peter Kovesi, Robyn Owens, Olivia Ladinig, Bas de Haas and anonymous reviewers for comments on drafts of this paper.

References

- [1] P. Desain and H. Honing, *Final report NWO-PIONIER project “Music, Mind, Machine”*, X-2004-02, Institute for Logic, Language and Computation, Universiteit van Amsterdam, 2004.
- [2] E.W. Large and M.R. Jones, *The dynamics of attending: How people track time-varying events*, *Psychol. Rev.* 106 (1999), pp. 119–59.
- [3] H. Honing, *Structure and interpretation of rhythm and timing*, *Tijdschr. V. Muziekth.* (Dutch J. Music Theory) 7 (2002), pp. 227–32.
- [4] P. Fraisse, *Rhythm and tempo*, in *The Psychology of Music*, D. Deutsch, ed., 1st ed., Academic Press, New York, 1982, pp. 149–80.
- [5] E.F. Clarke, *Rhythm and timing in music*, in *The Psychology of Music*, D. Deutsch, ed., 2nd ed., Academic Press, San Diego, CA, 1999, chap. 13, pp. 473–500.
- [6] J. London, *Rhythm*, in *The New Grove Dictionary of Music and Musicians*, S. Sadie and J. Tyrrell, eds., 2nd ed., Macmillan, London, UK, 2001, pp. 277–308.
- [7] M. Yeston, *The Stratification of Musical Rhythm*, Yale University Press, New Haven, CT, 1976, 155pp.
- [8] F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music*, MIT Press, Cambridge, MA, 1983, 368pp.
- [9] M.R. Jones and M. Boltz, *Dynamic attending and responses to time*, *Psychol. Rev.* 96 (1989), pp. 459–491.
- [10] P. Desain and H. Honing, *Computational Models of Beat Induction: The Rule-Based Approach*, *J. New Music Res.* 28 (1999), pp. 29–42.
- [11] P. Desain and H. Honing, *Modeling the effect of meter in rhythmic categorization: Preliminary results*, *Japan. J. Music Percept. & Cogn.* 7 (2001), pp. 145–156.
- [12] F. Gouyon and S. Dixon, *A review of automatic rhythm description systems*, *Comput. Music J.* 29 (2005), pp. 34–54.
- [13] E.F. Clarke, *Levels of structure in the organization of musical time*, *Contemp. Music Rev.* 2 (1987), pp. 211–238.
- [14] N.P. Todd, *The auditory “Primal Sketch”: A multiscale model of rhythmic grouping*, *J. New Music Res.* 23 (1994), pp. 25–70.
- [15] E.D. Scheirer, *Tempo and beat analysis of acoustic musical signals*, *J. Acoust. Soc. America* 103 (1998), pp. 588–601.
- [16] L.M. Smith, *Modelling rhythm perception by continuous time–frequency analysis*, in *Proceedings of the International Computer Music Conference*, 1996, pp. 392–395.
- [17] 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*, August, Montreal, Quebec, 1996, pp. 197–202.
- [18] L.M. Smith, *A multiresolution time–frequency analysis and interpretation of musical rhythm*, Department of Computer Science, University of Western Australia, 1999.
- [19] O. Rioul and M. Vetterli, *Wavelets and signal processing*, *IEEE Signal Process. Mag.* 8 (1991), pp. 14–38.
- [20] I. Daubechies, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992, 357pp.

- [21] A. Grossmann and J. Morlet, *Decomposition of Hardy functions into square integrable wavelets of constant shape*, SIAM J. Math. Anal. 15 (1984), pp. 723–736.
- [22] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, New York, 1998, 577pp.
- [23] R. Kronland-Martinet and A. Grossmann, *Application of time-frequency and time-scale methods (wavelet transforms) to the analysis, synthesis, and transformation of natural sounds*, in *Representations of Musical Signals*, G.D. Poli, A. Piccialli and C. Roads, eds., MIT Press, Cambridge, MA, 1991, pp. 45–85.
- [24] M. Holschneider, *Wavelets: An Analysis Tool*, Clarendon Press, Oxford, 1995, 423pp.
- [25] M. Vetterli and C. Herley, *Wavelets and filter banks: Theory and design*, IEEE Trans. Signal Process. 40 (1992), pp. 2207–2232.
- [26] B. Boashash, *Time-frequency signal analysis*, in *Advances in Spectrum Analysis and Array Processing*, S. Haykin, ed., Prentice-Hall, Englewood Cliffs, NJ, 1990, pp. 418–517.
- [27] D. Gabor, *Theory of communication*, IEE Proc. 93 (1946), pp. 429–457.
- [28] R. Kronland-Martinet, J. Morlet, and A. Grossmann, *Analysis of sound patterns through wavelet transforms*, Int. J. Pattern Recog. & AI 1 (1987), pp. 273–302.
- [29] R. Kronland-Martinet, *The wavelet transform for analysis, synthesis and processing of speech and music sounds*, Comput. Music J. 12 (1988), pp. 11–20 (Sound examples on soundsheet with Vol. 13, No. 1, 1989).
- [30] A. Grossmann, R. Kronland-Martinet, and J. Morlet, *Reading and understanding continuous wavelet transforms*, in *Wavelets*, J. Combes, A. Grossmann and P. Tchamitchian, eds., Springer-Verlag, Berlin, 1989, pp. 2–20.
- [31] P. Tchamitchian and B. Torrèsani, *Ridge and skeleton extraction from the wavelet transform*, in *Wavelets and Their Applications*, M.B. Ruskai, ed., Jones and Bartlett, Boston, MA, 1992, pp. 123–151.
- [32] N. Delprat, B. Escudé, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrèsani, *Asymptotic wavelet and Gabor analysis: Extraction of instantaneous frequencies*, IEEE Trans. Inf. Theory 38 (1992), pp. 644–664.
- [33] P. Guillemain and R. Kronland-Martinet, *Characterization of acoustic signals through continuous linear time-frequency representations*, Proc. IEEE 84 (1996), pp. 561–585.
- [34] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*, 2nd ed., Macmillan, New York, 1992, 969pp.
- [35] T. Kohonen, *Emergence of invariant feature detectors in self-organization*, in *Computational Intelligence: A Dynamic System Perspective*, M. Palaniswami, Y. Attikiouzel, R.J. Marks II, D. Fogel and T. Fukuda, eds., IEEE Press, New York, 1995, pp. 17–31.
- [36] H. Honing and O. Ladinig, *The Effect of Exposure and Expertise on Timing Judgments: Preliminary Results*, M. Baroni, A.R. Adessi, R. Caterina and M. Costa, eds., Bologna, Italy, 2006, pp. 80–85.
- [37] M.R. Jones, *Attentional rhythmicity in human perception*, in *Rhythm in Psychological, Linguistic, and Musical Processes*, J.R. Evans and M. Clynes, eds., Charles Thomas, Springfield, IL, 1986, chap. 2, pp. 13–40.
- [38] R. Parncutt, *A perceptual model of pulse salience and metrical accent in musical rhythms*, Music Percept. 11 (1994), pp. 409–464.
- [39] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*, MIT Press, Cambridge, MA, 2006.
- [40] D.W. Massaro, *Preperceptual images, processing time, and perceptual units in auditory perception*, Psychol. Rev. 79 (1972), pp. 124–145.
- [41] A.S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, MA, 1990, 773pp.
- [42] J.J. Bharucha, *Tonality and expectation*, in *Musical Perceptions*, R. Aiello and J. Sloboda, eds., Oxford University Press, Oxford, 1994, pp. 213–239.
- [43] P.E. Allen and R.B. Dannenberg, *Tracking musical beats in real time*, in *Proceedings of the International Computer Music Conference*, 1990, pp. 140–143.
- [44] D.F. Rosenthal, *Machine rhythm: Computer emulation of human rhythm perception*, MIT Media Lab, Cambridge, MA, 1992.
- [45] M. Goto and Y. Muraoka, *Music understanding at the beat level—real-time beat tracking for audio signals*, in *IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 1995, pp. 68–75.
- [46] N.P.M. Todd, D.J. O’Boyle, and C.S. Lee, *A sensory-motor theory of rhythm, time perception and beat induction*, J. New Music Res. 28 (1999), pp. 5–28.
- [47] A. Grossmann, M. Holschneider, R. Kronland-Martinet, and J. Morlet, *Detection of abrupt changes in sound signals with the help of wavelet transforms*, in *Inverse Problems: An Interdisciplinary Study; Advances in Electronics and Electron Physics*, P.C. Sabatier, ed., Academic Press, New York, 1987, pp. 289–306.
- [48] L. Solbach, R. Wöhrmann, and J. Kliewer, *The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis*, in *Working Notes of the Workshop on Computational Auditory Scene Analysis at the International Joint Conference on Artificial Intelligence*, August, Montreal, Quebec, 1995.
- [49] S. Handel, *Listening: An Introduction to the Perception of Auditory Events*, MIT Press, Cambridge, MA, 1989, 597pp.
- [50] D.J. Povel, *Internal representation of simple temporal patterns*, J. Exp. Psychol. - Human Percept. & Perform. 7 (1981), pp. 3–18.
- [51] H. Christopher Longuet-Higgins, *The perception of music*, Proceedings of the Royal Society of London 205 (1979), pp. 307–322.

Time–frequency representation of musical rhythm by continuous wavelets

Leigh M. Smith* and Henkjan Honing

*Music Cognition Group/ILLC, Universiteit van Amsterdam, Plantage Muidergracht 24,
1018TV, The Netherlands*

1. Previous work in rhythm representation

Considerable research has been directed at designing models of both pulse induction and at tracking processes, towards the final goal of producing useful and robust models of musical time. Existing models have used various approaches including grammars [1,2], expectancy [3], statistics [4,5], Minskyian agents [6], oscillator entrainment [7,8] and other self-organising connectionist systems [9–11]. A recent review of rhythm description systems is provided by Gouyon and Dixon [12]. Common problems confronted and addressed in a diverse manner by these approaches are the representation of temporal context, order and hierarchy, and the role of expressive timing and tempo within the existing rhythmic structure.

Since musical rhythm can be induced from clicks alone, a rhythmic function for analysis is created by representing the time of each onset as a unit impulse function. The rhythm function for a piece of music is therefore a train of impulses with intervals matching the IOI between onsets. A pulse-train function can be seen to be a minimal or *critical* sampling of the auditory amplitude envelope at the lowest sampling frequency which still accurately represents the rhythm function. This yields one sample at the point in time at which each musical event becomes audible to the listener. This is an onset-based representation of rhythm, and is typically recovered by a detection algorithm operating on the energy of the acoustic signal, $|y|^2$. This is effectively a rectification of the audio signal in order to deconvolve the amplitude envelope from the auditory carrier signal [13,14]. Alternatively the onset times are obtained by transducing a musicians actions on a sensing instrument (e.g., MIDI). This is distinguished by Gouyon and Dixon [12] from a time domain frame-based system which aims to determine the rhythmic signal directly from the auditory signal. However, in practice, systems directly processing the audio signal must rectify it in order to deconvolve the rhythm before further processing.

The explicit treatment of a rhythm as a signal, applicable to digital signal processing methods, has only recently become widely adopted as a computational approach to rhythm perception. Notable early examples have been the use of autocorrelation methods by Desain and de Vos [15] and Brown [16] to detect periodicities in MIDI data and audio recordings, respectively. Goto

*Corresponding author. Email: lsmith@science.uva.nl

and Muraoka [17] developed a beat tracking system capable of processing musical audio in real time. Their system manages multiple hypotheses created by beat prediction agents. These agents track “onset time finders” which operate on spectral energy representations. The frequency and sensitivity parameters of the onset time finders are tuned by the agents. Domain knowledge is used to pattern match against characteristic spectral features (snare and bass drums) and typical popular music rhythms (strong-weak-strong-weak repetition). Beat intervals are determined by interval histograms weighted by reliability of tracking estimates. As the authors admit, the system makes strong assumptions on the music to be tracked, such as popular songs in $\frac{4}{4}$ and constrained tempo range and variability.

Formalising the dynamic attending model of Jones and Boltz [18], Jones and Large [8] have developed a model based on coupled self-oscillating “attending rhythms” that entrain to external (performed) rhythms. Phase and period (frequency) coupling of the attending rhythm to the external rhythm enable variations in the timing of the external rhythm to be tracked. Internal coupling between oscillators is designed to pull pairs into a 2:1 phase-locked relationship. The model relies on an attentional focus period during each attentional rhythm’s oscillation that creates a period of enhanced responsiveness (“attentional targetting” or a “temporal responsive field” [19, p. 80]) to strengthen the coupling. The period of focus narrows with repetition, responding to onsets which are expected to fall within this attended period. Conversely, onsets which fall at half periods (twice the rate) lie outside the attentional focus. Hence, the model is insensitive to change at these points in time [19]. In the model described in this article, to begin to track an external rhythm which then changes to double time, a new oscillator must begin tracking, starting from a wide attentional focus, or the currently attending rhythm must adapt to this new double rate rhythm. The dynamics of each oscillator form a time and frequency varying filter which can adapt to timing variations from the musical rhythm.

Scheirer [14] demonstrated that the sum of the amplitude envelopes of a bank of subbands is sufficient to induce a rhythmic percept that matches the original musical signal. His model uses a six band filter bank to attempt to separate polyphonic input by frequency bin. This assumes that any polyphony occurring within that frequency band will not be rhythmically contradictory (i.e., play in polyrhythm). The band-pass “envelope channel” output is rectified, smoothed, differentiated and then half-wave rectified in order to sharpen the amplitude envelope of the sub-band to an approximation of an onset signal. This produces impulsive energy to a causal resonant comb filter bank which resonates at the periodicity of the sub-band. The comb filters only resonate with onsets at periods matching their delay times; a sufficient number of resonators is required to track tempo deviations in real time. Therefore, Scheirer’s tests must be applied to music that exhibits a “strong beat”.

In a similar fashion, Sethares and Staley [14] used a signal decomposition method that extracts strictly periodic components from the RMS energy of audio sub-bands. They select basis elements according to several different criteria, including the best correlation of the elements to the sub-band signal. As the authors note, “When the tempo of the performance is unsteady, the periodicity methods fail, highlighting the methods’ reliance on a steady underlying pulse” [20, p. 152]. Such “unsteady” tempo includes musically essential gestures such as *ritardando* and *accelerando*; so, this approach and Scheirer’s seems to be limited in their applicability for music that exhibits such expressive timing.

Klapuri et al. [21] generalises the preprocessing of the audio signal of Scheirer [14] and Goto and Muraoka [17] to produce many sub-bands which are then summed to a subset of four channels, termed “accent signals”. These accent signals are then subject to periodicity detection. This is performed by a bank of comb filter resonators (matching Scheirer’s) whose output are summed to a single measure of current periodicities. Three hidden Markov models form a probabilistic model to estimate the periods and phases of the tactus, bar, and shortest rhythmic interval (“tatum”) from the summed and individual resonator energies, respectively. While described probabilistically,

the periods and phase probabilities are not determined by training, they are derived from features chosen by hand, estimated to have a proportional influence. Phase of the bar is chosen by pattern matching against two expected energy time profiles derived by inspection. Using such an approach outperforms two reference systems, Scheirer [14] and Dixon [22], although pieces which exhibit substantial expressive timing were not tested.

Modelling tempo tracking as a stochastic dynamical system, Cemgil et al. [23] represent tempo as a hidden state variable estimated by Kalman filtering. They model tempo as a strict periodicity with an additional noise term that describes expressive timing as a Gaussian distribution. The Kalman filter uses a multiscale representation of a real performance, the “tempogram”. This is used to provide a Bayesian estimation of the likelihood of a local tempo given a small set of onsets. The tempogram is calculated by convolving onset impulses with a Gaussian function, which is then decomposed onto multiscale basis functions. These bases are similar to Scheirer’s comb filters, forming a local constant tempo that is used to match against the incoming rhythm. By considering rhythm as locally constant, and timing deviations as noise, the system does not take advantage of the underlying structure of a performer’s expressive timing.

Todd [13] applied banks of Mexican hat filters, analogous to the primal sketch theory of human vision [24], towards an auditory primal sketch of rhythm. This produces a “rhythmogram” representation of an audio signal. A recent publication by Todd [25] postulates a rhythm perception model based in the auditory periphery and controlling directly the musculoskeletal system. While such a neurobiologically inspired model may be plausible, it is difficult to measure the contribution of each component of the model against a base-line. For example, the performance of a model may be due to its accurate representation of neurobiology, or alternatively mostly due to the behaviour of the signal processing systems incorporated therein. A model that attempts to relate directly to components of the human auditory periphery may not provide the simplest explanation for the output, making verification difficult. Given the massive connectionism present in neurobiology, a model may not reflect neurological architecture in a sufficient manner to produce accurate results, or be simply too computationally expensive to test thoroughly.

The use of the continuous wavelet transform as a means of analysing rhythms consisting of an impulse train of onsets was originally reported in Smith [26] and Smith and Kovesi [27]. The output of the transform is similar to Todd’s rhythmogram, but more detailed. The representation reveals a hierarchy of rhythmic strata and the time of events by using a wavelet that has the best combined frequency and time resolution. Such bottom-up data-oriented approaches, including the multiresolution method described in this paper, do not fully account for human rhythm cognition. Rhythm perception is additionally influenced in a top-down manner by the listener’s memory, developed by a combination of explicit training and learning through exposure. A goal of this paper is to clarify the information which is inherent (i.e., retrievable) in the temporal structure of a musical rhythm. This aims to establish a base-line measure to evaluate the contribution of different models of musical time before considering the effect of top-down processing. In short, the intention of this paper is to demonstrate how much structure can be retrieved from a sparse impulse representation of a rhythmic signal.

Acknowledgements

This research was realized in the context of the EmCAP (Emergent Cognition through Active Perception) project funded by the European Commission (FP6-IST, contract 013123). Thanks go to Peter Kovesi, Robyn Owens, Olivia Ladinig, Bas de Haas and anonymous reviewers for comments on drafts of this paper.

References

- [1] H.C. Longuet-Higgins and C.S. Lee, *The perception of musical rhythms*, Perception 11 (1982), pp. 115–28.

- [2] D.F. Rosenthal, *Machine Rhythm: Computer Emulation of Human Rhythm Perception*. PhD thesis, MIT Media Lab, Cambridge, Mass, 1992.
- [3] P. Desain, A *(De)Composable theory of rhythm perception*, Music Percept. 9(4) (1992), pp. 439–54.
- [4] C. Palmer, and C.L. Krumhansl, *Mental representations for musical meter*, J. Exp. Psychol. – Human Percept. Perform. 16(4) (1990), pp. 728–41.
- [5] R. Parncutt, *A Model of Beat Induction Accounting for Perceptual Ambiguity by Continuously Variable Parameters*, in *Proceedings of the Proceedings of the International Computer Music Conference*, 1994, pp. 83–4.
- [6] J. Chung, An Agency for the Perception of Musical Beats, or, If I Only Had a Foot. . . MIT media laboratory report, Massachusetts Institute of Technology, 1989.
- [7] E.W. Large and J.F. Kolen, *Resonance and the perception of musical meter*, Conn. Sci. 6(2+3) (1994), pp. 177–208.
- [8] E.W. Large and M.R. Jones, *The dynamics of attending: How people track time-varying events*, Psycho. Rev. 106(1) (1999), pp. 119–59.
- [9] P. Desain and H. Honing, *The quantization of musical time: A connectionist approach*, Comput. Music J. 13(3) (1989), pp. 56–66.
- [10] S.C. Roberts and M. Greenhough, *Rhythmic Pattern Processing Using A Self-Organising Neural Network*. In: *Proceedings of the Proceedings of the International Computer Music Conference*, 1995, pp. 412–9.
- [11] M.P. Page, *Modelling the perception of musical sequences with self-organizing neural networks*, Conn. Sci. 6(2+3) (1994), pp. 223–46.
- [12] F. Gouyon and S. Dixon, *A review of automatic rhythm description systems*, Comput. Music J. 29(1) (2005), pp. 34–54.
- [13] N.P. Todd, *The auditory “Primal Sketch”: A multiscale model of rhythmic grouping*, J. New Music Res. 23(1) (1994), pp. 25–70.
- [14] E.D. Scheirer, *Tempo and beat analysis of acoustic musical signals*, J. Acoustic. Soc. Am. 103(1) (1998), pp. 588–601.
- [15] P. Desain and S. de Vos, *Autocorrelation and the Study of Musical Expression*, in *Proceedings of the Proceedings of the International Computer Music Conference*, 1990, pp. 357–360.
- [16] J.C. Brown, *Determination of the meter of musical scores by autocorrelation*, J. Acoustic. Soc. Am. 94(4) (1993), pp. 1953–7.
- [17] M. Goto, and Y. Muraoka, *Music Understanding At The Beat Level – Real-time Beat Tracking For Audio Signals*, in *Proceedings of the IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 1995, pp. 68–75.
- [18] M.R. Jones and M. Boltz, *Dynamic attending and responses to time*, Psychol. Rev. 96(3) (1989), pp. 459–91.
- [19] H. Honing, *Is there a perception-based alternative to kinematic models of tempo rubato?* Music Percep. 23(1) (2005), pp. 79–85.
- [20] W.A. Sethares and T.W. Staley, *Meter and periodicity in musical performance*, J. New Music Res. 30(2) (2001), pp. 149–58.
- [21] A.P. Klapuri, A.J. Eronen, and J.T. Astola, *Analysis of the meter of acoustic musical signals*, IEEE Trans. Audio, Speech Lang. Process. 14(1) (2006), pp. 342–55.
- [22] S. Dixon, *Automatic extraction of tempo and beat from expressive performances*, J. New Music Res. 30(1) (2001), pp. 39–58.
- [23] A.T. Cemgil, B. Kappen, P. Desain, and H. Honing, *On tempo tracking: Tempogram representation and kalman filtering*, J. New Music Res. 29(4) (2000), pp. 259–73.
- [24] D. Marr, and E. Hildreth, *Theory of edge detection*, Proc. R. Soc. Lond. B 207 (1980), pp. 187–217.
- [25] N.P.M. Todd, D.J. O’Boyle, and C.S. Lee, *A sensory-motor theory of rhythm, time perception and beat induction*, J. New Music Res. 28(1) (1999), pp. 5–28.
- [26] L.M. Smith, *Modelling Rhythm Perception by Continuous Time-Frequency Analysis*, in *Proceedings of the Proceedings of the International Computer Music Conference*, 1996, pp. 392–5.
- [27] Smith, L.M. and Kovesi, P., 1996, *A Continuous Time-Frequency Approach To Representing Rhythmic Strata*, in *Proceedings of the Proceedings of the Fourth International Conference on Music Perception and Cognition*, August, Montreal, Quebec, pp. 197–202.

Time-frequency representation of musical rhythm by continuous wavelets

Leigh M. Smith* and Henkjan Honing

*Music Cognition Group/ILLC, Universiteit van Amsterdam, Plantage Muidergracht 24,
1018TV, The Netherlands*

1. Dynamic accents

Figure 1 illustrates an analysis of a rhythm composed of a meter changing from an intensified beat every three beats to an accent every four beats and then returning to every three beats. The IOI's remain equal across the pulse train; only the beat which is intensified is changed. As can be seen, a band of frequency scales corresponding to the interval between accented beats is established during the $\frac{3}{4}$ meter period, dips downwards for the $\frac{4}{4}$, and returns to the previous scale. This demonstrates the zooming of the multiresolution model, and its ability to track a short term change in frequency. The phaseogram in Figure 1 indicates congruence over ranges of scales corresponding to the rhythmic band. Additionally, the phase highlights the points of change in the signal, where a frequency (meter) change occurs. The non-causal nature of the convolution operator used in the continuous wavelet transform (CWT) pin-points the rhythmic alternation.

2. Timing

A key feature of the CWT representation is that this pulse induction process can identify a time modulated (retarding/accelerating) rhythmic pulse. It does not require pulses to be isochronous and can therefore avoid representing local deviations in rhythmic frequency as “noise”. As a demonstration, two scaleograms of rhythms are displayed in Figures 2 and 3. Both analyse the same number of events, all impulses are equal in amplitude and the size of deviations in expressive timing are also equal. The variations in the grouping, in four (Figure 2) and in three (Figure 3), are made visible by the multiresolution representation. The multiple timescales reflect the different arrangement of the timing groups, and the periodicity of these groups appears as a low energy (but still visible and detectable) ridge. The ability to discriminate quasi-periodicity means that grouping from expressive timing that is not strictly periodic can also be represented. However, this timing shift is subtle, when supported only by timing accentuation alone. Co-occurring dynamic

*Corresponding author. Email: lsmith@science.uva.nl

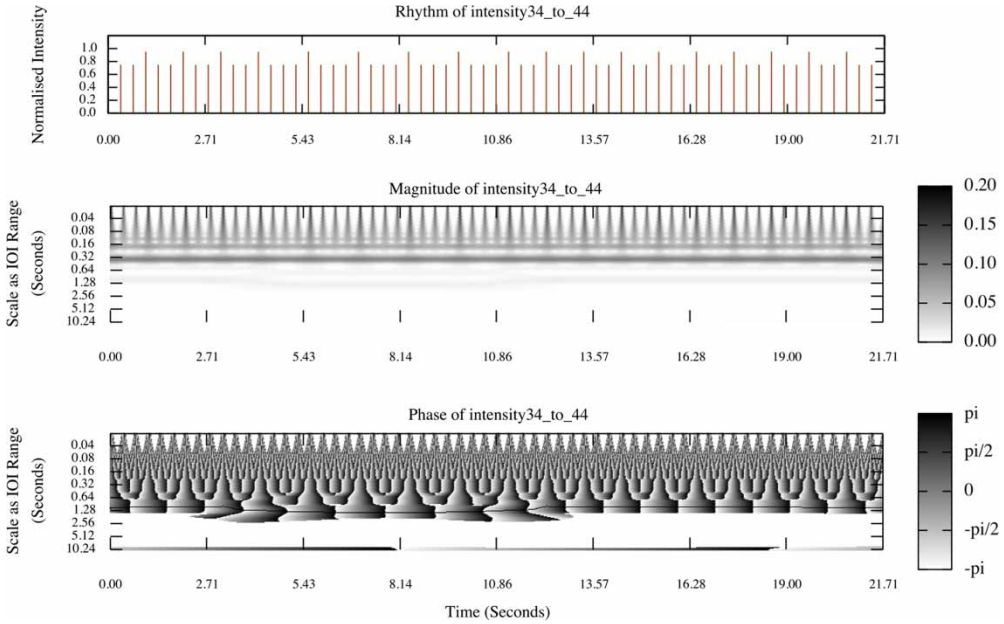


Figure 1. An isochronous rhythm changing in meter by variations in amplitude. The upper plot shows the impulse amplitudes, with the meter changing from $\frac{3}{4}$ to $\frac{4}{4}$ over the period of 4.2 to 11.2 seconds. The scaleogram (middle) and phaseogram (lower) plots display a continuous wavelet transform of the rhythmic impulse function. The intensity variations of the impulses are discernible in the scaleogram at short IOI scales, and the time-frequency ridge with the most energy is at 0.35 seconds matching the IOI. A lower energy ridge is visible on the scaleogram, and more clearly on the phaseogram, changing in its period from 1.05 seconds to 1.4 seconds matching the duration of the bar. It is marked on the phaseogram as a black line.

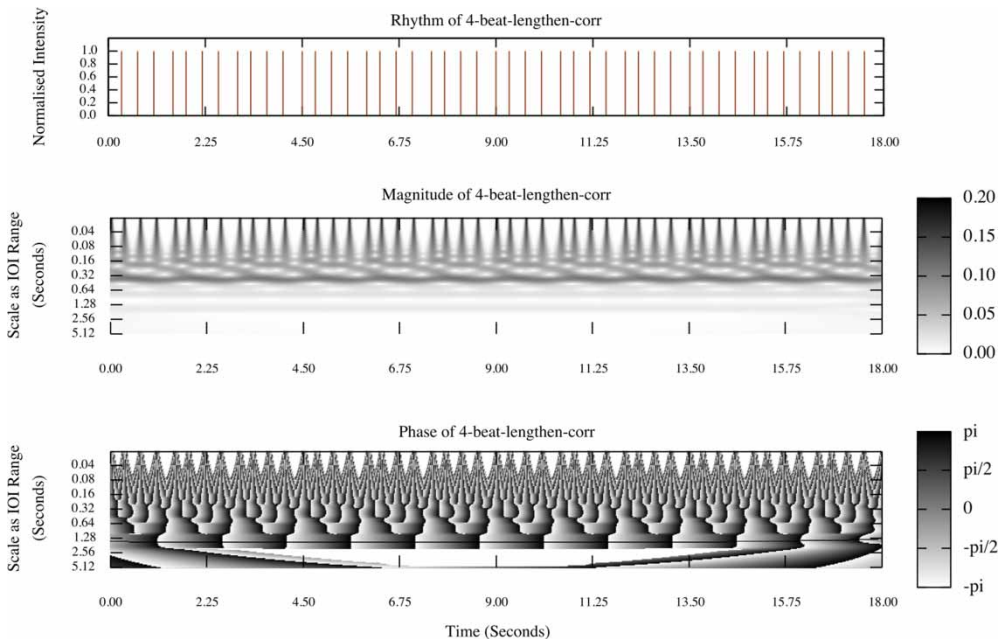


Figure 2. A rhythm that has a repeating IOI pattern of 0.305, 0.375, 0.375, 0.445 seconds. The period of the pattern (1.5 seconds) is shown on the scaleogram and is marked on the phaseogram as a black line.

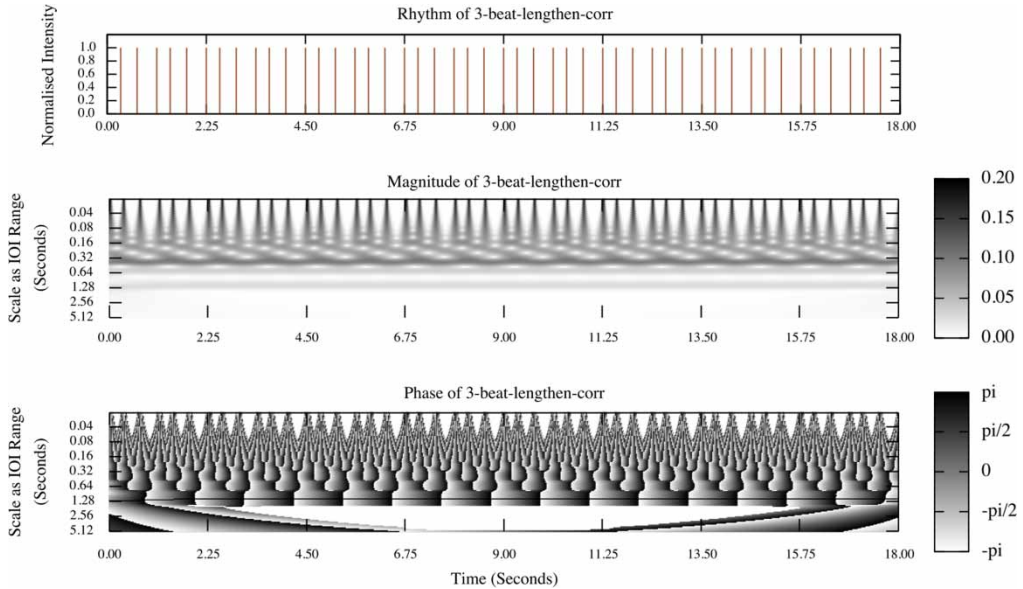


Figure 3. A rhythm that has a repeating IOI pattern of 0.305, 0.375, 0.445 seconds. The slight timing variation differences between this rhythm and that shown in Figure 2 are indicated in the different patterns of the high frequency scales, while a low energy ridge at the lower frequency scales matching the different periods (1.125 seconds and 1.5, respectively) of the repeated patterns is visible on the magnitude diagram, and is plotted on the phase diagram.

accentuation would make the grouping more apparent to the listener and in the scaleogram. Such accentuation is reflected by the CWT, as demonstrated in Figure 1.

Acknowledgements

This research was realized in the context of the EmCAP (Emergent Cognition through Active Perception) project funded by the European Commission (FP6-IST, contract 013123). Thanks go to Peter Kovcsi, Robyn Owens, Olivia Ladinig, Bas de Haas and anonymous reviewers for comments on drafts of this paper.