A Continuous Time-Frequency Approach To Representing Rhythmic Strata

Leigh M. Smith and Peter Kovesi
Department of Computer Science
University of Western Australia

Motivation

Modelling the cognition of musical rhythm offers insights into the theory of time perception, quantification of musical theories of performance and expression [6], and non-verbal artificial intelligence knowledge representation. Existing models of rhythm have used various approaches including grammars, expectancy [1], statistics, Minskyian agents, oscillator entrainment [5] and other self-organising connectionist systems.

A common problem confronted and addressed in a diverse manner by these approaches is the representation of temporal context, order and hierarchy, and the role of expressive timing within the existing rhythmic structure. This paper describes an approach to exhaustively represent the periodicities which are created by temporal relationships between beats over multiple timescales, including both metrical and agogic times. This multiple resolution analysis is performed using Morlet wavelets [3] over frequencies from 0.1 to 100Hz. The result is a decomposition of a musical rhythm into short term low frequency components to reveal transient details from which may be constructed an executable theory [6] of rhythmic structure.

Rhythmic strata, tactus and expectancy

Yeston has argued for the conception and representation of rhythm as a hierarchy of strata [11], with a meter arising from accents created by the interaction between hierarchial levels. Lerdahl and Jackendoff [7] have argued for two hierarchies, a metrical structure and a grouping structure. Their metrical structure is a decomposition of the listeners sense of repetitive beat, again by interactions between levels of their metrical grid. The *tactus* (roughly the foot tapping rate), is considered as the most salient hierarchial level. The grouping structure is responsible for establishing boundaries of hierarchial grouping over phrases, motives and other units of musical time, interacting with the metrical structure.

Desain's decomposable rhythm perspective has similarities to a wavelet approach [1]. He forward projects expectancy curves in time which are composed from Gaussian sections with parameters determined from the ratio of previous time intervals. The curves are also weighted by an absolute time component, creating tempo dependency. The expectancy curve is calculated by summing the expectancies determined from all of the possible intervals between all onsets. Each time point of highest expectancy positions a window where beats are identified.

Frequency approaches

An isochronous rhythm may be considered as a single periodicity of a measurable frequency — the reciprocal of the inter-onset-interval (IOI) between beats. Objective (for example dynamic) accenting of regularly spaced beats creates a meter and thereby two frequencies, that implied by the period of the beat, and that by the period of the measure, a lower frequency. Expressive timing deviations from a strict metrical grid may be reviewed as relatively short term frequency deviations and modulations from a metrical "carrier frequency". A rhythm composed of syncopations and varied IOIs may be conceived of as a linear combination of short term frequency components, in the sense of Fourier theory. A sufficiently fine grained *time-frequency* analysis will then reveal these components over their time period.

Recovering the rhythm from an acoustic signal can be achieved by rectification [8], alternatively, capturing the monophonic timing from a MIDI controller avoids this processing. The rhythm is represented as impulses at the time of note onset, weighted by it's MIDI velocity value. This makes an assumption of a linear relationship between intensity and cognitive salience and ignores timbre, pitch and duration effects. It does however succintly represent the cognitive structure of percussive rhythms, which can be memorised and recalled independent of timbre, pitch and duration.

Neural oscillator entrainment models use a hierarchy of oscillators to effectively respond to periodicities in the rhythm, within frequency bands defined by the dynamics of the phase locking behaviour. Such models are not actually interlocked between hierarchies [5, pp. 198], suggesting independent stratified layers of rhythmic times produced by a time-frequency analysis will reveal the signal on which the oscillators are self-organising.

Wavelets

The discrete Fourier transform decomposes an arbitary signal onto infinite time, complex exponential bases in harmonic relationship to the analysis period. This localises frequency while losing representing frequency change over time. The short term Fourier transform (STFT) addresses this by a time window over the signal, *fixed across the frequency plane*. By contrast, the continuous wavelet transform (CWT) [3] decomposes a time t varying signal $s(t)$ onto scaled and translated versions of a *mother-wavelet* $g(t)$,

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

where $\bar{g}(t)$ is the complex conjugate and a is the scale parameter, controlling the dilation of the window function, effectively stretching the window geometrically over time. The translation parameter b centres the window in the time domain. The geometric scale 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. The CWT indicated in Equation 1 is a scaled and translated instance from a bank of an infinite number of constant relative bandwidth (Q) filters. For a discrete implementation, a sufficient density of scales (a) or "voices" per octave is required.

Morlet and Grossmann's mother-wavelet [2] for $g(t)$ is a scaled complex Gabor function,

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

where ω_0 is the frequency of the mother-wavelet before it is scaled. Choices of $\omega_0 \geq \pi\sqrt{2/\ln 2}$ will be close to a zero mean [3]. The Gaussian envelope over the complex exponential provides the best possible simultaneous time/frequency localisation [2], respecting the Heisenberg uncertainty relation. This ensures that all short term periodicities contained in the rhythm will be captured in the analysis. The time domain of $s(t)$ which can influence the wavelet output $W_s(b_0, a_0)$ at the point b_0, a_0 is an inverted logarithmic cone with its vertex at b_0, a_0, equally extending bidirectionally in time. Where impulses fall within the time extent of a point, $W_s(b_0, a_0)$ will return a high energy value.

The phase of Equation 2 is constant. Real valued wavelets such as Marr's [8] do not provide an anti-symmetric component to produce a phase which oscillates independent of the signal. This produces magnitudes which will not be congruent across scales, deviating forward and backward in time. By the "progressive" nature of Equation 2 [2, 3], the real and imaginary components of $W_s(b, a)$ are the Hilbert transform of each other. These can be computed as magnitude and phase components and then plotted in grey scales on a "scalogram" and "phasogram" (Figure 2) respectively. Phase values are mapped from the domain $0 - 2\pi$ to black through to white. The transition from white to black indicates a return to 0. Vertical lines of constant shade indicates a congruence of phase over a range of frequencies.

In order to preserve phase, Morlet wavelets are non-causal, convolving a signal with the wavelet family in both directions in time. Phase has been shown to be important in discrimination at audio frequencies [10]. Given the effects of backwards masking and auditory persistence [8], there may be an integrating period, such that the auditory system may not be totally causal. However this does not account for the non-causality of the convolution operator over longer time scales in a rhythm. Convolution places the events within a global time construct and is therefore in some sense a theory about the cognition of a complete rhythmic structure, at the time of performance, rather than directly the listeners perception.

Local Energy

Phase indicates the progression of a periodic wave though its cycle. Phase values of a voice regularly oscillating linearly between 0 to 2π, indicates the frequency represented by that voice is present in the signal. Morrone and Owens have reported compelling evidence for the *local energy* model for feature detection in image processing, proposing that features of an image are perceived at points where the Fourier components are most in phase [9]. Peaks in the local energy function can be used to indicate points of maximum phase congruency. The local energy function $E(t)$ of the signal $s(t)$ can be defined as,

$$E(t) = \sqrt{\left[\sum_n^N \mathcal{R}[W_s(t,n)]\right]^2 + \left[\sum_n^N \mathcal{I}[W_s(t,n)]\right]^2},$$

where N is the number of voices, and $\mathcal{R}[x]$, $\mathcal{I}[x]$ are the real and imaginary outputs from the CWT of Equation 1 for each voice respectively. With a progressive mother-wavelet, a *singularity* such as an impulse will be marked by a constant phase [2].

The local energy function will typically produce high values at the impulse times. Applied in the 1-D case, phase congruency produces a dimensionless measure of the alignment of beat frequencies within a global temporal context. The similarity between phase congruent temporal feature detection and Yeston's theory of inter-hierarchial accents is striking.

Example Analysis

The snare drum rhythm of Ravel's "Bolero" was chosen to demonstrate the behaviour of the CWT on a musical example (See Figure 1). The rhythm's metrical durations were used with a tempo of 60 bpm to generate an unaccented impulse train at 200Hz sampling rate. Morlet wavelets were discretised over 16 "equally tempered" voices per octave, for 10 octaves, with the maximum analysing period of the wavelets ranging from 2 to 2048 samples. The scalogram and phasogram of two repeats of the rhythm are shown in Figure 2, with scales plotted in rhythmic notation according to the tempo. At the highest scales, the impulses are discernable due to the localisation over a short time window. At lower scales, the frequencies created by the IOIs of the impulses are indicated by regular phase oscillations and grey scaled magnitude values. Figure 3 is a 3-D "rhythmic contour plot" combining scale and phase to illustrate the rhythm frequency energy variation over time for a region centered at the 750th sample (3.75 seconds). Figure 4 is a cross-sectional plot illustrating the frequencies present at that time point. The repeated rhythm induces a high (dark) ridge at the dotted semibreve voice and a clear hump for the triplet semiquaver IOI voice, a train of which surround 750th sample. The crochet voice – the tactus for this rhythm, undulates in energy value, notably being the highest scale which remains distinct across the window. While impulses will create high phase congruency (Figure 5), the quaver impulse points which contrast to the triplet semiquaver sequences over longer timescales actually create lower points of local energy than their surrounding intervals. The repeat of the phrase produces higher local energy values than the first sequence due to the congruent dotted semibreve voice.

Future Work

This paper has demonstrated a phase congruent wavelet analysis for revealing rhythmic strata. This finds common ground with Lerdahl and Jackendoff's metrical and grouping structure process, however it does not represent a generative grammatical approach to interpreting those structures. Additionally, it is not suggested that listeners utilise a direct Gabor filter multiresolution process in their rhythm perception – while noting that Kohonen has recently presented evidence for the self organisation of Gabor wavelet transforms [4]. Rather, the aim of this paper has been to examine the information contained within a musical rhythm before any perceptual processing is performed. The intention is to make explicit that information which is inherent in the rhythm, viewed as a formalisation of the decomposition conceptual model.

In the current model, the energy density for each wavelet is proportional to it's frequency. Further work is to investigate rescaling the scalogram to better represent a cognitive model. While it is tempting to draw hypotheses for methods of derivation of the tactus by "ridge-tracing" or the well-formedness [7] of the global continuation of a voice, further research is required to build a model of tactus in respect of perceptual issues.

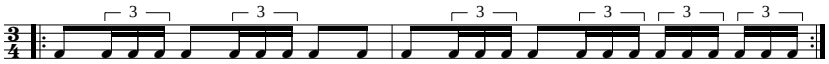

Figure 1: The snare drum rhythm of “Bolero”.

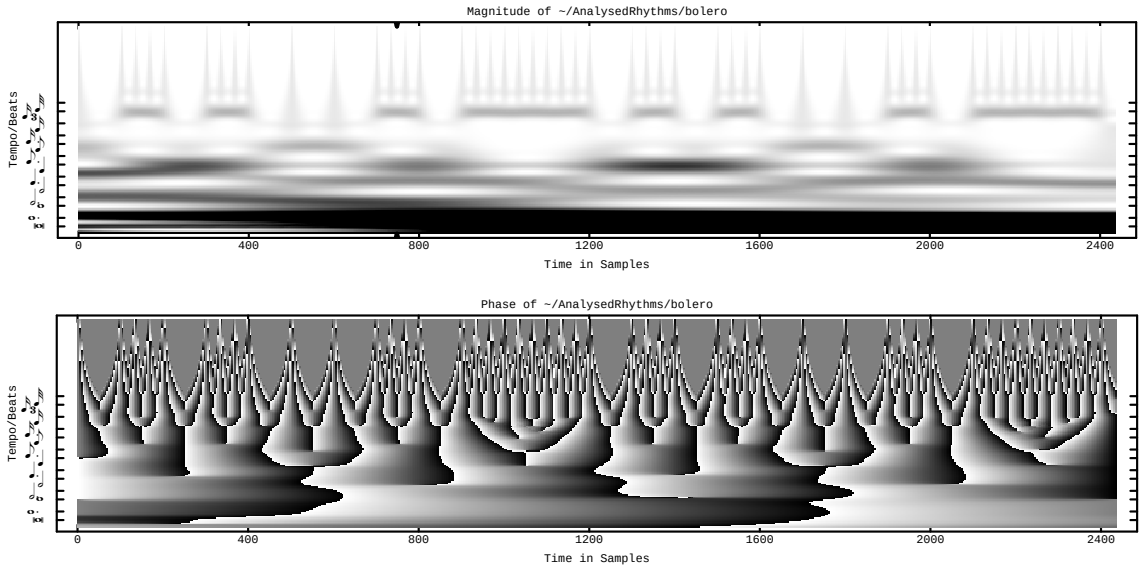

Figure 2: Time-Scale scalogram (top) and phasogram (bottom) displays of a CWT of the rhythmic impulse function of two repetitions of the rhythm in Figure 1.

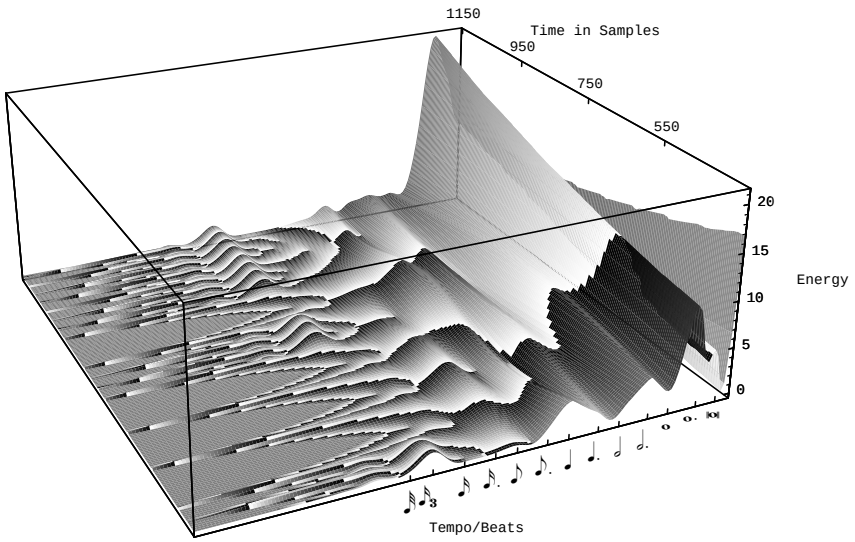

Figure 3: Combined magnitude and phase display of a portion of the CWT of Figure 2.

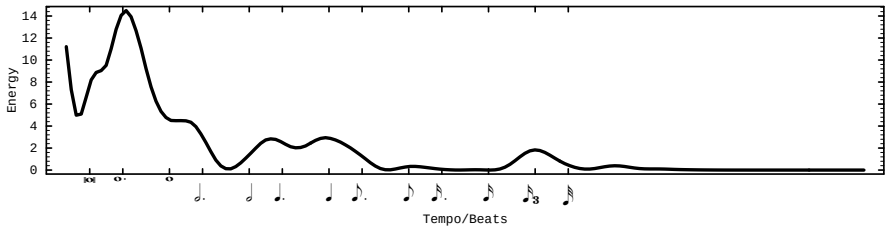

Figure 4: Cross-section at the 750th sample through the scalogram in Figure 2. The logarithmic scale produces dotted rhythmic units which do not evenly subdivide surrounding times.

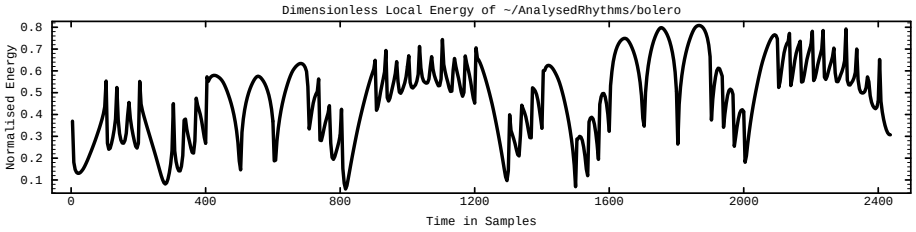

Figure 5: Local energy display of a CWT of the rhythm in Figure 1.

References

[1] P. Desain. A (de)composable theory of rhythm perception. *Music Perception*, 9(4):439–54, 1992.

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

[3] M. Holschneider. *Wavelets: An Analysis Tool*. Clarendon Press, 1995. 423 p.

[4] T. Kohonen. Emergence of invariant feature detectors in self-organization. In M. Palaniswami, Y. Attikiouzel, R. J. Marks II, D. Fogel, and T. Fukuda, editors, *Computational Intelligence: A Dynamic System Perspective*, chapter 2, pages 17–31. IEEE Press, New York, 1995.

[5] E. W. Large and J. F. Kolen. Resonance and the perception of musical meter. *Connection Science*, 6(2+3):177–208, 1994.

[6] O. E. Laske. Artificial intelligence and music: A cornerstone of cognitive musicology. In M. Balaban, K. Ebcioglu, and O. E. Laske, editors, *Understanding Music with AI*, pages 3–28. Massachusetts Institute of Technology, Cambridge, Mass, 1992.

[7] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. Massachusetts Institute of Technology, Cambridge, Mass, 1983. 368p.

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

[9] M. C. Morrone and R. A. Owens. Feature detection from local energy. *Pattern Recognition Letters*, 6:303–313, December 1987.

[10] A. V. Oppenheim and J. S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–41, 1981.

[11] M. Yeston. *The stratification of musical rhythm*. Yale University Press, New Haven, 1976. 155p.