Modelling Rhythm Perception by Continuous Time-Frequency Analysis[*]

Leigh M. Smith
Department of Computer Science
University of Western Australia
Crawley, W.A. 6907
leigh@cs.uwa.edu.au

Abstract

The use of linear phase Gabor transform wavelets is demonstrated as a robust analysis technique capable of making explicit many elements of human rhythm perception behaviour. Transforms over a continuous time-frequency plane (the scalogram) spanning rhythmic frequencies (0.1 to 100Hz) capture the multiple periodicities implied by beats at different temporal relationships. Wavelets represent well the transient nature of these rhythmic frequencies in performed music, in particular those implied by agogic accent, and at longer time-scales, by rubato.

The use of the scalogram phase information provides a new approach to the analysis of rhythm. Measures of phase congruence over a range of frequencies are shown to be useful in highlighting transient rhythms and temporal accents. The performance of the wavelet transform is demonstrated on examples of performed monophonic percussive rhythms possessing intensity accents and rubato. The transform results indicate the location of such accents and from these, the inducement of phrase structures.

1 Rhythm as a Signal

Despite a history of computational approaches to rhythm perception and production, the treatment of a rhythm as a signal applicable to digital signal processing methods has not been widespread. Notable recent exceptions have been Todd's approach of applying banks of Mexican hat filters [MH80] to an acoustic signal towards an auditory primal sketch of rhythm [MT94], and Desain's use of autocorrelation to determine periodicities in rhythm [DdV90].

Engineering approaches of representing an acoustical phenomenon as a signal have brought a conceptual rigour and considerable application to the field of acoustics and music. Careful conceptualisation of musical rhythm in signal processing terms furthers our understanding and utilisation of rhythm in musical contexts. Equally careful consideration of psychological and musicological issues is necessary to build accurate, useful signal representations.

The use of signal processing techniques uncovers the information inherent in the rhythm prior to our perceptual processing. The time perception effects such as masking [Cow84] and top-down expectancy from subjective rhythmisation [Fra82], are considered in this paper as processes that occur in parallel and with respect to a rhythmic frequency map.

1.1 Rhythm Frequencies

Theorists have argued for representations of musical rhythm with respect to the *tactus* [LJ83], and that a rhythm can be decomposed into a hierarchy of rhythmic strata of increasing time spans, subdividing the tactus. With a constant tempo, and no dynamic or timing accents, a repetitive beat (an *isorhythm*) can be considered as a single periodicity with an underlying frequency $f = \frac{1}{\lambda}$, where λ is the inter-onset-interval (IOI) of the beat and f is the rhythmic frequency — the rate of event presentation. An example in performance terms is performing a $\frac{4}{4}$ crochet rhythm; objectively accenting the downbeats of each measure by intensifying them indicates two frequencies, that implied by the period of the crochet, and that by the period of the semi-breve, at a quarter of the crochet frequency.

1.2 Capturing Musical Intention

In acoustic terms, the rhythmic signal is described by the amplitude behaviour over time. Essentially, this is an amplitude modulation of a carrier signal in the auditory frequency range by a modulating signal in the rhythmic frequency range. Frequency analysis of the *rectification* of the time-amplitude signal will separate the low frequency rhythmic signal from the acoustic signal [MT94].

[*] Many thanks go to Peter Kovesi for some example code of 1-D Morlet wavelets and Robyn Owens for comments on drafts of this paper. Further detail and examples are available from http://www.cs.uwa.edu.au/~leigh/rhythmfreq.html and during the conference.

An alternative pragmatic approach is to capture the musician's intention rather than capturing the acoustic result. This is done by sampling the rhythmic signal before it is made audible, that is, before it is "multiplied" with the auditory carrier signal, and subsequently producing an audible rhythm. For the present purposes of analysis, electronic MIDI drum pads were used to transduce the time of a drum strike and a measure of intensity of the strike (MIDI velocity). While timbre from an acoustic drum is used to produce accenting, given that a rhythm can be induced by a listener from timing alone, a lack of timbral information is not significant.

1.3 Frequency Analysis of Rhythm

Wavelet transforms [RV91] were originally conceived to simultaneously localise analysis in both the time and frequency domains. Wavelets represent time-varying frequency information with better accuracy than alternative transformation techniques such as the short time Fourier transform (STFT). Expressive timing, agogic, dynamic and other objective accents will produce multiple, short term frequency and amplitude varying rhythmic signals which can be revealed with such a non-stationary signal analysis technique.

Using very short impulse-like taps, a familiar rhythm can be recognised, or a new rhythm comprehended and tapped along with. Therefore one promising representation of a rhythmic function to an analysis transform is to take the short duration tap in the limit and represent the time of the *onset* of each beat as a unit impulse function weighted by a normalised measure of the intensity of the beat:

$$\iota(t) = \begin{cases} v/127 & \text{if } t = 0 \\ 0 & \text{otherwise,} \end{cases}$$

where t is the sample index at the onset of the note, $\iota(t)$ is the impulse value (0.0–1.0), and v is the MIDI velocity value (1–127). The rhythm function for a piece of music is therefore an uneven train of pulses. The sparse, uneven pulse train can also be viewed as an oversampling of a much lower frequency signal. The sampling rate can be low (200Hz) as the audible frequencies are not present.

Weighting the impulse incorporates the effect of dynamic accent, assuming 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 [Cow84] of beats by temporal proximity. Further work is therefore needed to match the impulse weighting to the actual psychoacoustic effect, but a simple linear mapping seems an appropriate first approximation.

2 The Wavelet Transform

In contrast to the STFT, the continuous wavelet transform (CWT) [RV91] decomposes a time t varying signal $s(t)$ onto scaled and translated versions of a *mother-wavelet* $g(t)$,

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

where $\bar{g}(t)$ is the complex conjugate of $g(t)$, 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.

There are many choices for mother wavelets; Kronland-Martinet and Grossmann [KMG91] have applied a complex Gabor mother-wavelet to sound analysis,

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

where ω_0 is the frequency of the mother-wavelet before it is scaled. This kernel does not meet "admissibility conditions" to reconstruct the signal from the $W_s(b,a)$ coefficients, in contrast to orthogonal basis functions [RV91]. While losing reconstruction, Equation 2 preserves the phase during analysis. It is close to a "progressive" [KMG91] wavelet, nearly satisfying $\forall\omega < 0 : \hat{g}(\omega) = 0$, where $\hat{g}(\omega)$ is the Fourier transform of $g(t)$.

The CWT indicated in Equation 1 is a scaled and translated filter from a constant relative bandwidth (Q) filter bank, comprised of an infinite number of filters or "voices". For implementation, a sufficient density of voices per octave is required for a discrete approximation.

Due to the progressive nature of Equation 2, 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 and then plotted in grey scales on a "scalogram" and "phasogram" (Figure 1) 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.

2.1 Phase Congruency and Local Energy

Phase indicates the progression of a periodic wave though its cycle. Therefore an oscillating phase at a scale, characterised by regularly spaced dark to white transitions, indicates that frequency is present in the signal. Image processing research in feature detection has found compelling evidence for the *local energy* model, proposing that features of an image are perceived at points where the Fourier components are most in phase [MO87]. 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 $s(t)$ at time 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 scales in the discretisation, and $\mathcal{R}[x]$, $\mathcal{I}[x]$ produce the real and imaginary outputs from the CWT of Equation 1 at each scale respectively. It is feasible that some phase congruent temporal feature detection may occur in the 1-D case. A similar approach has been taken by Todd with respect to Marr's primal sketch theory of human vision [MT94].

An impulse is localised in time, but infinite in frequency content. A CWT of an impulse localises the impulse's effect in the time domain at the higher frequency scales and spreads the effect across longer finite time periods at lower scales [KMG91]. Using a progressive mother-wavelet, a *singularity* such as an impulse will be marked by a constant phase [GHKMM87]. The local energy function will therefore indicate points where the impulses fall and where points of lesser congruence lie.

3 Examples

To understand the decomposition behaviour of the CWT, these analyses have used monophonic percussive rhythms, exhibiting intensity accenting. It is hypothesised that listeners use timbral, spatial localisation, pitch and other objective differences between sound sources to distinguish between independent rhythmic patterns. Thus, a polyphonic rhythm would be represented by a number of parallel wavelet analyses.

The input data to the wavelet transform were recorded from MIDI then converted into an impulse train at 200Hz sample rate. The wavelet transformation extended over 10 octaves to a maximum analysis window of 2048 samples, with 16 voices per octave.

The x-axis of the scalogram plots time in samples, and the y-axis plots the frequency scale of the dilation of the wavelet in samples of its time period. At the highest scales (the highest y-axis values), the time window is very short, two samples, and the original impulse is apparent. At lower scales, the frequency localisation is more apparent and the rhythms are seen as parallel frequency bands corresponding to the frequencies implied by impulses at different intervals.

The CWT is demonstrated on an expressive performance of the rhythm of "Greensleeves" with multiple IOIs grouped in typical proportions. Figure 1 indicates the hierarchy of frequencies implied at each time point due to the intervals between beats falling within each scaled kernel's support. An exaggerated shortening of the semi-quavers produces characteristic sweeps at 480, 1000, 2480 and 3000 samples in the scalogram. The dark patterns beginning around the 1600th sample at the lowest scale are caused by the edge of the window. Extending the analysis over longer times (at the cost of processing) would remove these effects. The phasogram in Figure 1 indicates higher frequency periodicities but does not as clearly indicate the enduring low scale periodicity with an IOI of approximately 860 samples. Inspection of the local energy (Figure 2), reveals points of high phase congruency fall on impulses (beats) at phrase endings. Local energy produces a measure of structural importance of a beat, weighted by its intensity and its temporal context within the rhythmic frequency hierarchy.

4 Assessment and Further Work

Phase preserving Gabor wavelets have been proposed here as a means of analysing musical rhythm. The transform represents the rhythmic effects generated by dynamic and temporal accents in establishing hierarchies of rhythmic frequencies. This hierarchical representation conforms closely with music theories of the inducement of temporal structure, meter and expressive timing.

A powerful model of rhythm perception has a number of computer music applications — transcription, scorefile editing and computer accompaniment such as score following or interactive performance systems. To achieve the latter application with the continuous wavelet model requires a real-time unambiguous determination of the tactus or beat. This in turn requires modelling the constraints and behaviour of human perception and performance. Todd has proposed two peak receptive bandwidths in the rhythmic frequency spectrum (corresponding to body sway and foot-tapping tempos) which strongly influence the listener's perception of tactus [MT95]. Investigation into the applicability of such a theory to the described model is necessary.

References

[Cow84] N. Cowen. On short and long auditory stores. *Psychological Bulletin*, 96(2):341–70, 1984.

[DdV90] P. Desain and S. de Vos. Autocorrelation and the study of musical expression. In *Proceedings of the International Computer Music Conference*, pages 357–360. International Computer Music Association, 1990.

[Fra82] P. Fraisse. Rhythm and tempo. In D. Deutsch, editor, *The Psychology of Music*, pages 149–80. Academic Press, New York, 1982.

[GHKMM87] 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*, Supplement 19, pages 289–306. Academic Press, New York, 1987.

[KMG91] 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 G. D. Poli, A. Piccialli, and C. Roads, editors, *Representations of Musical Signals*, pages 45–85. Massachusetts Institute of Technology, Cambridge, Mass, 1991.

[LJ83] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music.* Massachusetts Institute of Technology, Cambridge, Mass, 1983. 368p.

[MH80] D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London B*, 207:187–217, 1980.

[MO87] M. C. Morrone and R. A. Owens. Feature detection from local energy. *Pattern Recognition Letters*, 6:303–313, December 1987.

[MT94] N. P. McAngus Todd. The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research*, 23(1):25–70, 1994.

[MT95] N. P. McAngus Todd. The kinematics of musical expression. *Journal of the Acoustical Society of America*, 97(3):1940–9, 1995.

[RV91] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, October 1991.

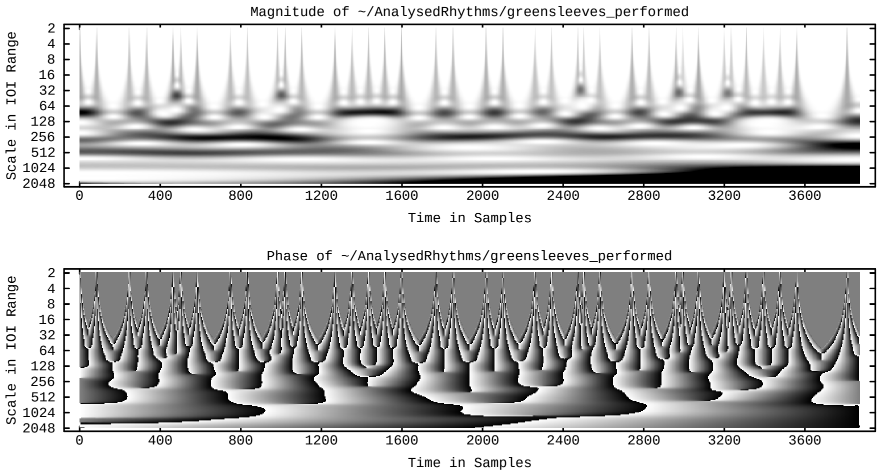

Figure 1: Time-Scale scalogram (top) and phasogram (bottom) displays of a CWT of the rhythmic impulse function of a performed version of "Greensleeves" possessing expressive timing modulations from the notated rhythm.

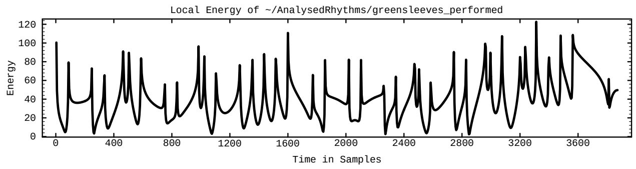

Figure 2: Local energy display of a CWT of the same performed "Greensleeves" rhythmic impulse function as Figure 1.