Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC 10). Sapporo, Japan.
Ken'ichi Miyazaki, Yuzuru Hiraga, Mayumi Adachi, Yoshitaka Nakajima, and Minoru Tsuzaki (Editors)

A Multiresolution Model of Rhythmic Expectancy

Leigh M. Smith and Henkjan Honing
Music Cognition Group, ILLC / Universiteit van Amsterdam
lsmith@science.uva.nl, www.musiccognition.nl

ABSTRACT

We describe a computational model of rhythmic cognition that predicts expected onset times. A dynamic representation of musical rhythm, the multiresolution analysis using the continuous wavelet transform is used. This representation decomposes the temporal structure of a musical rhythm into time varying frequency components in the rhythmic frequency range (sample rate of 200Hz). Both expressive timing and temporal structure (score times) contribute in an integrated fashion to determine the temporal expectancies. Future expected times are computed using peaks in the accumulation of time-frequency ridges. This accumulation at the edge of the analysed time window forms a dynamic expectancy. We evaluate this model using data sets of expressively timed (or performed) and generated musical rhythms, by its ability to produce expectancy profiles which correspond to metrical profiles. The results show that rhythms of two different meters are able to be distinguished. Such a representation indicates that a bottom-up, data-oriented process (or a non-cognitive model) is able to reveal durations which match metrical structure from realistic musical examples. This then helps to clarify the role of schematic expectancy (top-down) and it's contribution to the formation of musical expectation.

I. MUSICAL EXPECTATION

Understanding the processes behind the generation of expectancy in music has become a key research question (Meyer, 1956; Jones and Boltz, 1989; Huron, 2006). Given only rhythmic stimuli (everything else being equal), how do temporal expectations of musical events arise?

Bharucha (1993, 1994) distinguished between *veridical* and *schematic* musical expectancies. The former describes expectations during the performance of a particular piece of music while the latter form of expectation arise from abstracting from particular pieces to unifying mental schemas. Huron (2006) most recently has distinguished the terminology further, reserving veridical expectancy for the expectation of a performance of a previously heard piece. He then termed *dynamic expectation* as the prediction of future events while listening to a piece of music that has been previously unheard.

While many dimensions of music invoke expectations, such percepts can arise in rhythm alone, purely from a temporal structure, with no distinguishing melodic, intensity or other accentuations. To study rhythmic expectation, we propose a multiresolution model of musical rhythm in Section II. and evaluate that in Section III. with data sets of generated and recorded rhythms.

II. A MULTIRESOLUTION MODEL OF EXPECTANCY

A number of models of musical rhythm have been proposed, including oscillator based approaches (Scarborough et al., 1990; Large and Kolen, 1994; Large and Jones, 1999). A less researched approach is the use of multiple resolution representations (Todd, 1994a; Smith and Kovesi, 1996; Todd et al., 1999). These represent a rhythmic signal as a pyramid of time-frequency components (wavelets), decomposing the rhythm into short-term periodicities. This representation brings out salient periodicities, similar to the behaviour of a large (more than 100) bank of highly damped oscillators.

Expectation is modelled as a set of predictions of future onsets generated from a combined time-frequency representation of a rhythm. This representation is generated by a *continuous wavelet transform* (CWT) operating on a temporal window containing past events. This represents musical time as a bank of simultaneous short term periodicities or oscillations. Such multiresolution representations of rhythm have been previously demonstrated to reveal periodicities in the temporal structure of onsets matching rhythmic structure of the music (Todd, 1994b; Smith, 1996; Smith and Kovesi, 1996; Smith and Honing, 2007, 2008).

A. Continous Wavelet Transform

The CWT (Holschneider, 1995; Mallat, 1998) decomposes a time t varying signal $s(t)$ onto scaled and translated versions of a *mother-wavelet* $g(t)$,

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

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 centers 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.

Grossmann et al. (1989)'s mother-wavelet for $g(t)$ is a scaled complex Gabor function (Gabor, 1946),

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

where ω_0 is the frequency of the mother-wavelet before it is scaled. The Gaussian envelope over the complex exponential provides the best possible simultaneous time/frequency localisation (Grossmann et al., 1989), 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_{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_{b_0,a_0} will return a high energy value. In this application $\omega_0 = 6.45$ by calibrating the maximum output W_{b_i,a_i} against an isochronous impulse train.

By the "progressive" nature of Equation 2 (Grossmann et al., 1989; Holschneider, 1995), the real and imaginary components of $W_{b,a}$ are the Hilbert transform of each other. These are computed as magnitude and phase components and can then be reduced to time-frequency *ridges* which minimally describe the time varying frequency components in the signal, known collectively as a *skeleton* (Tchamitchian and Torrésani, 1992; Smith and Honing, 2008).

Since a musical rhythm can be induced from mere clicks alone, the rhythm is typically represented for CWT analysis as a sparse set of impulses at the time of each onset, sampled at 200Hz, capturing the temporal structure. An alternative representation derived directly from an audio signal, the *onset saliency trace* has also been successfully used to analyse and accompany the rhythm of sung vocals (Coath et al., 2008). When applied to musical rhythm, a ridge is an oscillation at a rhythmic frequency, over a period of time, incorporating rubato. Ridges function as beat periods of a rhythm that are perceptually prominent. For each rhythm, its skeleton then represents the entire candidate set of beat periods available to a listener to attend to.

B. Dynamic Temporal Expectancy

Each wavelet coefficient $W_{b,a}$ represents a short-term periodicity at every time point b, so the frequency at an instant in time t can be determined from the scale parameter a and therefore also its wavelength. These may be interpreted as the forward projection (i.e. an estimate) in time for a future onset $t_k = t + 2^{a/v}$, where v is the number of voices per octave (16 in this application).

Within the analyzed time window, the magnitude of the wavelet coefficient $|W_{b,a}|$ is used as a measure of confidence (likelihood) of the expectancy prediction.

Ridges, which identify scales a of magnitude peaks, correspond to projection times with highest likelihoods of an onset occurring. The multiple ridges that may exist at a particular time point in the skeleton represents multiple simultaneous hypotheses of the next expected onset time. Dynamic temporal expectancy is then defined as a weighted set of all expectations from a given moment in time.

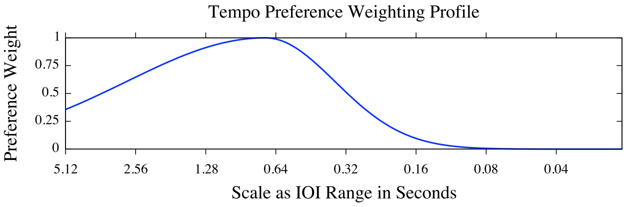

Figure 1. Tempo preference profile used for weighting expectation confidences.

Expectation into the future—beyond the rhythm signal currently recorded—is determined at the most recent edge of the analysis window. In terms of Bayesian probability, the likelihood of each estimated projection time is determined from the evidence observed in the time window. The evidence over the time window is the *ridge presence* P_a (Smith and Honing, 2007), amassed by summing the occurrence of ridge scales a over the time of the rhythm and normalising for its duration B:

$$P_a = \sum_{b=0}^{B-1} \frac{\mathrm{ridge}(W_{b,a})}{B}, \tag{3}$$

where $\mathrm{ridge}()$ is the normalised ridge peak function, derived from the magnitude local maxima of each wavelet coefficient $W_{b,a}$, described in detail in Smith and Honing (2008). Summing the ridges rather than simply integrating the scaleogram magnitude reduces the averaging effect of time-frequency uncertainty, resulting in more accurate predictions in time.

The ridge presence profile (over all scales $a \in A$) is then weighted for absolute tempo constraints. This consists of a concatenated Gaussian envelope with a mean at a period of 720 milliseconds (Parncutt, 1994), shown in Figure 1. Time periods shorter than the mean are weighted by a Gaussian of 1 octave per standard deviation, periods longer than the mean are weighted by a Gaussian of 2 octaves per standard deviation. This is designed to allow lower confidence long term projections to still be produced. Peaks in the ridge presence profile which are $w = 0.5$ standard deviations above the mean ridge presence peak values are then chosen as the projected expectations.

III. EVALUATION

To evaluate the model, two experiments were performed. The first using a Monte-Carlo simulation of the space of possible metrical rhythms to test the ability to produce expectation. The second test used a data set of performed musical rhythms (Temperley, 2007). Additionally, individual rhythms were verified for correct expectation times.

A. Sampling the Metrical Rhythm Space

The model was tested on sets of rhythms drawn from the total space of possible strictly metrical rhythms. These were generated randomly, whilest conforming to a given meter. A Monte-Carlo simulation was used to select from the large meter space (Desain and Honing, 1999). Profiles of the metrical position of onsets of the sample rhythms are shown in Figure 2 for two different meters. These are derived by weighting an empty interonset-interval (IOI) occurrence at each metrical level to 40% chance. These profiles consistently match the theoretical hierarchies reported by Palmer and Krumhansl (1990, Figure 1, pp. 731). Each rhythm used a fixed minimum semiquaver (16th note) of 150 milliseconds (30 samples). The binary $\frac{4}{4}$ meter rhythms were generated with 6 measures and the ternary $\frac{3}{4}$ meter rhythms with 8 measures, producing identical duration rhythms so only the temporal structure differed between the two meter groups.

These generated rhythms were then analysed with the multiresolution rhythm model described in Section II.. The expectation histograms for the two sets of metrical rhythms are shown in Figure 3. The accumulated confidence of an expectation time is the summation each rhythm's confidences of that time. Therefore the confidences are compared in relative terms. Since the duration of each generated rhythm measure (bar) is known, the expectation times are plotted on the abscissa axis as divisions of the measure. This is to compare the expectation times to the established rhythmic context.

Plotted behind the expectancy histograms are the corresponding metrical tree structures. These trees compare closely to the metrical profiles in Figure 2. The confidence accumulated over the set of rhythms shows noticeable peaks at divisions of the measure which corresponds to metrical subdivision boundaries. For example, for the $\frac{3}{4}$ metrical set in Figure 3, peaks occur at the 8.33, 8.66 and 9 measure positions, corresponding to the three crotchets (quarter notes) of that meter.

The expectation times and their relative confidence can also be compared to the occurrence of a given interval in the rhythms, as shown in Figure 4. For the $\frac{4}{4}$ meter example, there is a relatively strong peak at the seventh measure boundary compared to the IOI of 16 semiquavers (one measure), and a very strong peak at approximately half the measure (6.5, a minim duration), compared to the IOI of 8 semiquavers.

The expectations are well spread over several measures. This is due to there being a number of alternative expectation times generated at the end of each rhythm. While there are multiple alternatives, the relative confidence weights the likelihood of such intervals. This allows for possible subdivisions such as triplets in a binary rhythm. The relative confidence decays with further distance from the end of the rhythm, modelling a recency bias. This is an artifact of the energy conservation of the CWT, such that low frequency components have lower energy (i.e. confidence) spread over greater periods of time.

B. Performed Rhythms

The expectation model was also tested on a data set of performed musical rhythms, a set of MIDI keyboard performances of a subset of the Essen folk song collection (Temperley, 2007). Since some examples were significantly longer than others, a maximum of the starting 15 seconds of the rhythm was used. This was intended to test if the expectation can be formed quickly, matching human skills. Only the $\frac{3}{4}$ and $\frac{4}{4}$ pieces in the data set were tested in this experiment. Since the pieces were performed with a freely chosen tempo, the period of the measure is not fixed over the piece, or between pieces. In order to then evaluate the accuracy of the expectation times, they were divided by the minimum IOI, constituting the hypothesised *temporal atom* (Bilmes, 1993, (aka *tatum*)).

The accumulated confidences for the two metrical sets of expectancies are shown in Figure 5. On the rhythms in the $\frac{3}{4}$ meter, expectations appear at ternary multiples of the tatum, that is, around 3, 6 and 12 multiples of the minimum IOI (roughly corresponding to a semiquaver). This does not match the structure of the meter and would appear to be binary subdivisions of the meter period. With the strong peak at 6 tatums, the expectancies for this set would seem closer to $\frac{6}{8}$. On rhythms in the $\frac{4}{4}$ meter, expectations accumulate around binary multiples of the tatum, at 4, 8 and 16 multiples and more closely matches the intended meter. There does seem to be sufficient evidence to distinguish the two meters, however, since the profiles do significantly differ.

IV. IMPLICATIONS

This paper demonstrates that an expectation profile can be produced which corresponds to the transcribed meter of a rhythm. This indicates the degree that meter may emerge from a dynamic (bottom-up) expectation process. This then helps to clarify the role of schematic expectancy (top-down) and it's contribution to the formation of complete musical expectation. It can be hypothesised that schematic expectancy acts as a selection mechanism, rationing attentional resources (Jones and Boltz, 1989) to select from the candidate dynamic expectation. However this would also seem to be a task specific process, more attention would seem to be needed to accompany a rhythm, and adjust for contradicted expectations, than simply to listen, expecting and then confirming onsets falling over a short time span. The separation of the bottom-up and top-down processes enables these task specific processes to be explored.

Despite the current results, there is at least one shortcoming of the approach. Estimating time from the frequency (scale), is inherently inaccurate, and certainly accounts for part of the spread of expectations. Using the phase derived from the multiresolution analysis to address this is a current project. The CWT analysis functions across a time window in a non-causal fashion. This models, and therefore implies, that there is a leaky integration process constituting the short term memory. The exact behaviour of the update of this win-

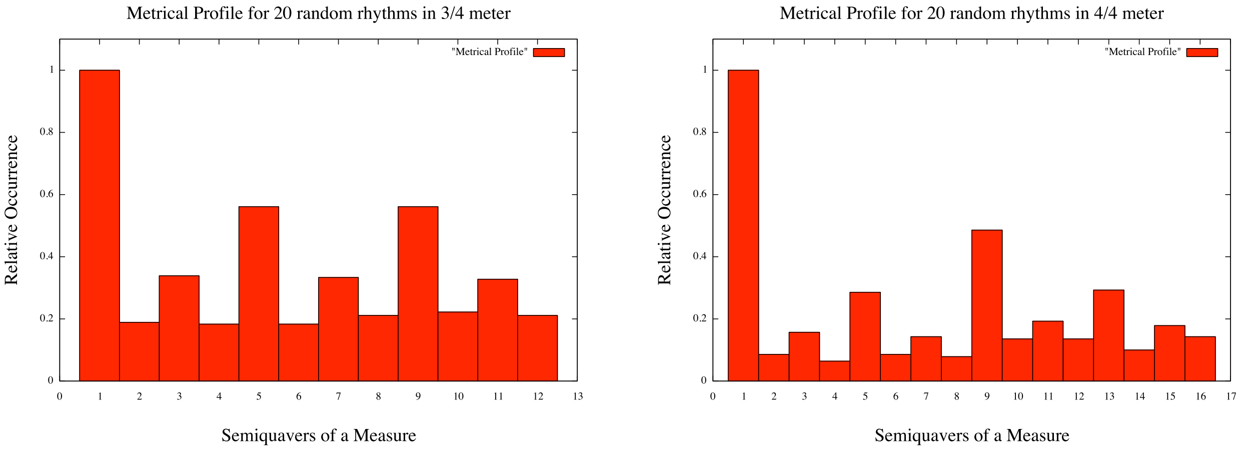

Figure 2. Metrical profiles for random samples of randomly generated metrical rhythms.

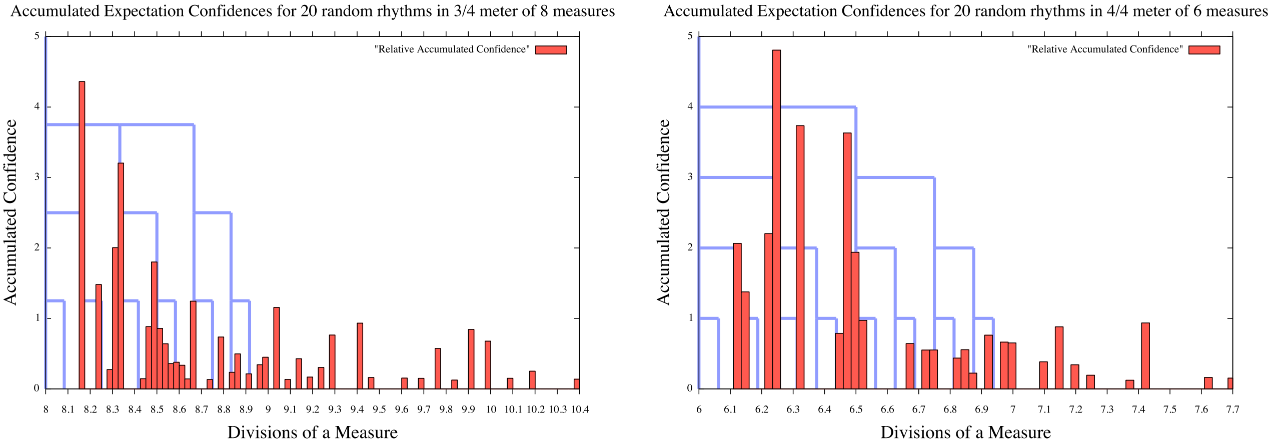

Figure 3. Accumulated expectancy profiles for random samples of randomly generated metrical rhythms. The canonical metrical trees are shown in blue behind the expectancy profiles. Peaks in the expectancy profiles correspond to major metrical subdivisions.

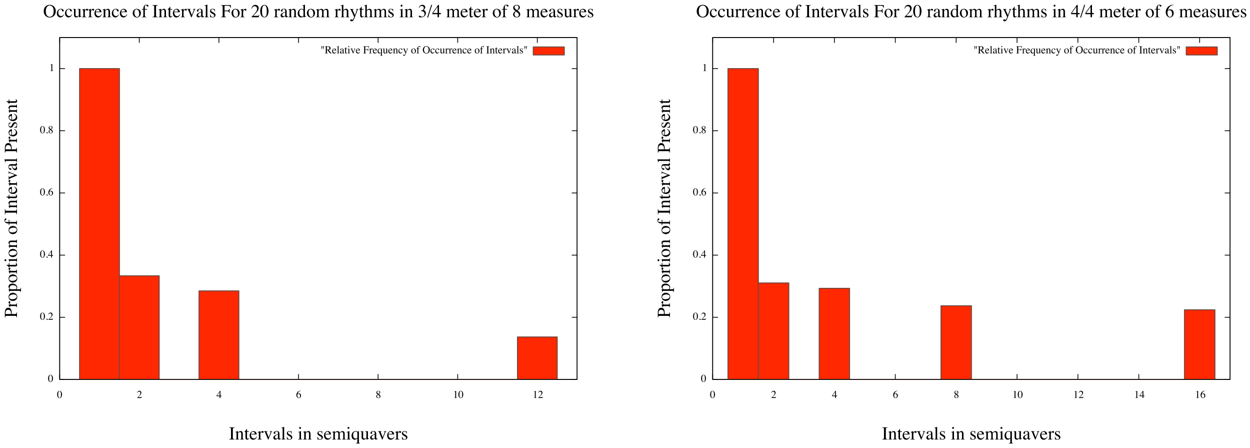

Figure 4. Histograms of the interonset intervals found in the set of rhythms analysed in Figure 3.

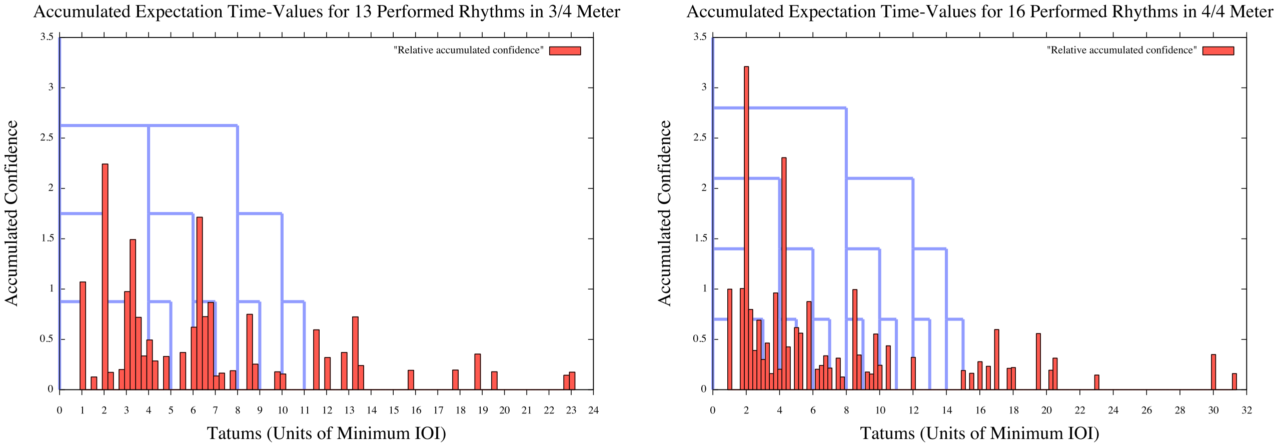

Figure 5. Accumulated expectancy profiles for rhythms taken from Temperleys performances of the Essen folk song collection (Temperley, 2007) for two meters. The abscissa axis is in *tatums*, representing the expectation time as a ratio of the minimum IOI in each rhythm. For tatum multiples approximating semiquavers, there are peaks in the accumulated expectancy approximating the metric multiples (4, 8 and 16) for $\frac{4}{4}$. There is only the measure period (12 tatums) as evidence for $\frac{3}{4}$, with peaks appearing for ternary subdivisions (3, 6 and 9).

dowed short term memory remains an open question.

V. 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 are due to Ricard Marxer Piñon for discussions and suggestions.

REFERENCES

Bharucha, J. J. (1993). MUSACT: A connectionist model of musical harmony. In S. M. Schwanauer and D. A. Levitt (Eds.), *Machine Models of Music*, pp. 497–510. Cambridge, Mass: MIT Press.

Bharucha, J. J. (1994). Tonality and expectation. In R. Aiello and J. Sloboda (Eds.), *Musical Perceptions*, pp. 213–239. Oxford University Press.

Bilmes, J. A. (1993, September). Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology.

Coath, M., S. Denham, L. M. Smith, H. Honing, A. Hazan, P. Holonowicz, and H. Purwins (2008). An auditory model for the detection of perceptual onsets and beat tracking in singing. *Connection Science*. (in press).

Desain, P. and H. Honing (1999). Computational models of beat induction: The rule-based approach. *Journal of New Music Research 28*(1), 29–42.

Gabor, D. (1946). Theory of communication. *IEE Proceedings 93*(3), 429–57.

Grossmann, A., R. Kronland-Martinet, and J. Morlet (1989). Reading and understanding continuous wavelet transforms. In J. Combes, A. Grossmann, and P. Tchamitchian (Eds.), *Wavelets*, pp. 2–20. Berlin: Springer-Verlag.

Holschneider, M. (1995). *Wavelets: An Analysis Tool*. Clarendon Press. 423 p.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, Mass: MIT Press.

Jones, M. R. and M. Boltz (1989). Dynamic attending and responses to time. *Psychological Review 96*(3), 459–91.

Large, E. W. and M. R. Jones (1999). The dynamics of attending: How people track time-varying events. *Psychological Review 106*(1), 119–59.

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

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press. 577p.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. University of Chicago Press. 307p.

Palmer, C. and C. L. Krumhansl (1990). Mental representations for musical meter. *Journal of Experimental Psychology - Human Perception and Performance 16*(4), 728–41.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception 11*(4), 409–64.

Scarborough, D. L., B. O. Miller, and J. A. Jones (1990). PDP models for meter perception. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ, pp. 892–9. Erlbaum Associates.

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

Smith, L. M. and H. Honing (2007). Evaluation of multiresolution representations of musical rhythm. In *Proceedings of the International Conference on Music Communication Science*, Sydney, Australia. Published online as `http://marcs.uws.edu.au/links/ICoMusic/Full_Paper_PDF/Smith_Honing.pdf`.

Smith, L. M. and H. Honing (2008). Time-frequency representation of musical rhythm by continuous wavelets. *Journal of Mathematics and Music 2*(2). (in press).

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

Tchamitchian, P. and B. Torrésani (1992). Ridge and skeleton extraction from the wavelet transform. In M. B. Ruskai (Ed.), *Wavelets and Their Applications*, pp. 123–51. Boston, Mass.: Jones and Bartlett Publishers.

Temperley, D. (2007). *Music and Probability*. Cambridge, Mass: MIT Press.

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

Todd, N. P. (1994b). Metre, grouping and the uncertainty principle: A unified theory of rhythm perception. In I. Deliége (Ed.), *Third International Conference on Music Perception and Cognition*, pp. 395–6. European Society for the Cognitive Sciences of Music.

Todd, N. P. M., D. J. O'Boyle, and C. S. Lee (1999). A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research 28*(1), 5–28.