A MULTIRESOLUTION TIME-FREQUENCY ANALYSIS AND INTERPRETATION OF MUSICAL RHYTHM

This thesis is
presented to the
Department of Computer Science
for the degree of
Doctor of Philosophy
of
The University of Western Australia

By
Leigh M. Smith
October 2000

Abstract

This thesis describes an approach to representing musical rhythm in computational terms. The purpose of such an approach is to provide better models of musical time for machine accompaniment of human musicians and in that attempt, to better understand the processes behind human perception and performance.

The intersections between musicology and artificial intelligence (AI) are reviewed, describing the rewards from the interdisciplinary study of music with AI techniques, and the converse benefits to AI research. The arguments for formalisation of musicological theories using AI and cognitive science concepts are presented. These bear upon the approach of research, considering ethnographic and process models of music versus traditionally descriptive methods of music study. This enquiry investigates the degree to which the human task of music can be studied and modelled computationally. It simultaneously performs the AI task of problem domain identification and constraint.

The psychology behind rhythm is then surveyed. This reviews findings in the literature of the characterisation of elements of rhythm. The effect of inter-onset timing, duration, tempo, accentuation, meter, expressive timing (rubato), the interrelationship between these elements, the degree of separability between the perception of pitch and rhythm, and the construction of timing hierarchy and grouping is reported. Existing computational approaches are reviewed and their degrees of success in modelling rhythm are reported.

These reviews demonstrate that the perception of rhythm exists across a wide range of timing rates, forming hierarchial levels within a wide-band spectrum of frequencies of perceptible events. Listeners assign hierarchy and structure to a rhythm by an arbitration of bottom-up phenomenal accents and top-down predictions. The predictions are constructed by an interplay between temporal levels. The construction of temporal levels by the listener arises from quasi-periodic accentuation.

Computational approaches to music have considerable problems in representing

musical time. In particular, in representing structure over time spans longer than short motives. The new approach investigated here is to represent rhythm in terms of frequencies of events, explicitly representing the multiple time scales as spectral components of a rhythmic signal.

Approaches to multiresolution analysis are then reviewed. In comparison to Fourier theory, the theory behind wavelet transform analysis is described. Wavelet analysis can be used to decompose a time dependent signal onto basis functions which represent time-frequency components. The use of Morlet and Grossmann's wavelets produces the best simultaneous localisation in both time and frequency domains. These have the property of making explicit all characteristic frequency changes over time inherent in the signal.

An approach of considering and representing a musical rhythm in signal processing terms is then presented. This casts a musician's performance in relation to an abstract rhythmic signal representing (in some manner) the rhythm intended to be performed. The actual rhythm performed is then a sampling of that complex "intention" rhythmic signal. Listeners can reconstruct the intention signal using temporal predictive strategies which are aided by familarity with the music or musical style by enculturation. The rhythmic signal is seen in terms of amplitude and frequency modulation, which can characterise forms of accents used by a musician.

Once the rhythm is reconsidered in terms of a signal, the application of wavelets in analysing examples of rhythm is then reported. Example rhythms exhibiting duration, agogic and intensity accents, accelerando and rallentando, rubato and grouping are analysed with Morlet wavelets. Wavelet analysis reveals short term periodic components within the rhythms that arise. The use of Morlet wavelets produces a "pure" theoretical decomposition. The degree to which this can be related to a human listener's perception of temporal levels is then considered.

The multiresolution analysis results are then applied to the well-known problem of foot-tapping to a performed rhythm. Using a correlation of frequency modulation *ridges* extracted using stationary phase, modulus maxima, dilation scale derivatives and local phase congruency, the tactus rate of the performed rhythm is identified, and from that, a new foot-tap rhythm is synthesised. This approach accounts for expressive timing and is demonstrated on rhythms exhibiting asymmetrical rubato and grouping. The accuracy of this approach is presented and assessed.

From these investigations, I argue the value of representing rhythm into time-frequency components. This is the explication of the notion of temporal levels

(strata) and the ability to use analytical tools such as wavelets to produce formal measures of performed rhythms which match concepts from musicology and music cognition. This approach then forms the basis for further research in cognitive models of rhythm based on interpretation of the time-frequency components.

Acknowledgements

Like a D.W. Griffith film, this thesis has one name as director, but has a cast of thousands ...

Philip Hingston acted as Masters by Research supervisor for the first two years until his departure to private industry. Nick Spadaccini then supervised for six months until his sabatical leave. Peter Kovesi then took the reins alone for six months until he was joined by Robyn Owens when I made the decision to convert to a PhD. Peter and Robyn both provided numerous suggestions of research directions, proof reading, cross-checking results and much needed advice and moral support. Both were instrumental in achieving an ARC grant which enabled research equipment to be purchased. When Robyn and Peter took simultaneous long service leave, C.P. Tsang filled as temporary supervisor for four months. Andy Milburn of tomandandy generously provided equipment and time for me to finish corrections.

Fellow Computer Science PhD students Matt Bellgard and Rameri Salama, and Jason Forte in the Psychology Department, provided me with many stimulating conversations and tricky questions to address. Bernard Cena's work with wavelets in vision research enabled me to discuss many concepts and approaches. Dave Cake and CompMuser[1] SKoT McDonald inspired me with boundless enthusiasm and energy for all things musical. Fellow Robvis lab inhabitants Bruce Backman, Mike Robbins and Dave O'Mara made a computer lab far more enjoyable to be in than I could have imagined.

Peter Kovesi, SKoT McDonald and Matt McDonald[2] are legendary individuals who perform amazing feats of proofreading at short notice in the face of my muddled English. The quality of the reading of this thesis is entirely due to them, and the lack of it entirely due to myself.

My parents Dorothy and Peter Smith deserve a huge thank-you for much needed

[1] `http://www.cs.uwa.edu.au/~skot/compmuse`

[2] No relation, just nearly as prolific as Smiths.

encouragement and support and a big thanks to my non-academic friends and flat-mates who bore with me through the endeavour. Finally, this thesis is dedicated to my Grandmother, Dorothy Blackwood, in her 95th year, for first providing the encouragement and wherewithal many years ago for me to become involved in music and later the inspiration, demonstration of resiliance, resolve and faith; both in the worth of music, and of patient endeavour.

Contents

List of Tables

List of Figures

Chapter 1

Music and AI — Mutually Beneficial Research Tasks

"It is very simple. If you consider that sound is characterized by its pitch, its loudness, its timbre, and its duration, and that silence, which is the opposite and, therefore, the necessary partner of sound, is characterized only by its duration, you will be drawn to the conclusion that of the four characteristics of the material of music, duration, that is, time length, is the most fundamental. Silence cannot be heard in terms of pitch or harmony: It is heard in terms of time length. It took a Satie and a Webern to rediscover this musical truth, which, by means of musicology, we learn was evident to some musicians in our Middle Ages, and to all musicians at all times (except those whom we are currently in the process of spoiling) in the Orient."

John Cage "Defense of Satie" [72, pp. 81]

"Motion is the significance of life, and the law of motion is rhythm. Rhythm is life disguised in motion, and in every guise it seems to attract the attention of man: from a child, who is pleased with the moving of a rattle and is soothed by the swing of its cradle, to a grown person whose every game, sport and enjoyment has rhythm disguised in it in some way or another, whether it is a game of tennis, cricket or golf, as well as boxing or wrestling. Again in the intellectual recreations of man, both poetry and music — vocal or instrumental — have rhythm as their very spirit and life. There is a saying in Sanskrit that tone is the mother of nature, but that rhythm is its father."

Hazrat Inayat Khan, "Rhythm", from "The Mysticism of Sound and Music" [70].

1.1 AI and Applications to Music

A significant problem with existing computer systems when applied to domains of music such as performance, composition and education are their extremely limited models of human musical knowledge and endeavour. These problems are noted by Stephen Smoliar [166] and well surveyed by Curtis Roads [145]. This has resulted in computer music systems which will function satisfactorily in limited domains of musical expertise but are easily "broken" in the face of unexpected inputs. These unexpected inputs are typically the result of human improvisation or ingenuity in interacting with a machine.

In attempting to construct a computer interactive performance system which is capable of interacting in a performance situation, such as playing in an ensemble or improvising with a human performer, the ability to respond to novel inputs becomes a necessity [149, 91, 160]. Even in non-realtime music applications, there is the need for better representations of music to make the system commands more intuitive, by making them correspond more closely to musical concepts and manipulate more meaningful musical data objects than is the current practice.

1.1.1 The Value of Formalisation of Musicology

This thesis constitutes an enquiry into the degree to which musicological theories which have been proposed can be rendered into formal models and tested. These formal models provide a "runable" theory of rhythmic structure allowing one to systematically test theories which have been experimentally determined from music psychology or those produced from more traditional music theory (codification of performance practice).

The limitations of current computer music systems can be seen in a wider context to be the result of attempting to produce a descriptive or artifact based model of music. This models the artifact, the score, or the recording, with only implicit consideration of the human cognition behind the musical material. The alternative approach is to construct ethnographic or process models. This has stimulated research in cognitive musicology, modelling the composition or performance processes of music computationally [139, 81]. Otto Laske et. al has engagingly argued the value of building computational models of musical intelligence:

> "At the very least, they [AI models] show researchers what, in music,

> does not yield to rule formulations, requiring perception-based, rather than logic-based, constructs. Knowledge-based systems are thus exploratory devices and tools of criticism, rather than definitive accounts of knowledge.
>
> ... As a scientist elucidating musical action, the modern musicologist is a humanist fast becoming a knowledge engineer in the service of anthropology.
>
> ... Musicology, like many of the humanities, has remained a predominantly hermeneutic science, focusing on the meaning of music structures for idealized listeners and on some vague collective spirit (often tinged with national colours).
>
> ... much of the research program of artificial intelligence is a reformulation of the failed research agenda of subjectivist philosophies from roughly 1450 to 1950 (Nicolaus of Cusa to Theodore Adorno).
>
> ... The main deficiency of subjectivist approaches to modelling *reason* as *intelligence* lies in the fact that human *reason* is cut off from human *action*, and is simultaneously viewed as the agency that controls action (This is the legacy of Descartes and Kant).
>
> ... we see the real challenge of AI and music in establishing cognitive musicology as an action science.
>
> [The discipline of AI and Music] ... focuses, not on intelligence, but on knowledge, which is a much broader notion; more specifically, it focuses on musical knowledge as an agency for designing musical action (theory-in-use), rather than on an agency supposedly understanding some sounding reality "out there"." (my emphasis)[81, pp. 19–24]

An action science is geared to understanding the theory-in-use of actors (theoretical musicology) and to improving the way in which the actors act (applied musicology). Therefore formalizing and implementing espoused musical knowledge, then testing it in performance situations has a critical value in demonstrating what is not understood about musical action, as much as what is.

1.1.2 Music's Value to AI

Music is detached of inherent meaning from its materials. We cannot speak of the "meaning" of a chord or a rhythm in the same sense we speak of the meaning of a

word or image. This separation offers music as a prime candidate for AI research. Listening to music can be considered as thinking in, or with sound, and organising sound. At the same time, knowledge of music is concrete and deeply rooted in the physics of the actual sounds themselves. Musical cognition *emerges* in music due to its serial nature. Language or speech about music fails to reveal musical processes. These qualities argue the value of studying music as a non-verbal knowledge representation which forms a "narrow" and "deep" problem domain. Marvin Minsky has convincingly argued [116] that these characteristics offer music as a work bench for knowledge representation and AI techniques.

1.2 Multiresolution Analysis of Rhythm

Modelling the human perception of performed musical rhythm offers many insights into the psychology of time perception [55], quantification of musical theories of performance and expression [81], and non-verbal artificial intelligence knowledge representation [116]. This thesis describes an approach of representing musical rhythm in computational terms and then analysing it using multiresolution techniques. The analysis results are then applied to an interpretation task—foot-tapping to performed rhythms. The output of this foot-tapping task can be an accompaniment to the original rhythm which can be audibly verified for its accuracy and musical appropriateness.

By considering rhythm as a low-frequency signal, wavelet signal processing theory and analytical techniques can be applied to decompose it and reveal its spectral components. From these components an executable theory can be constructed (in the form of a computer program) of a listener's perceptual processes. The extent to which a decomposition provides a musical handle to aid in machine understanding in all rhythmic cases is a question which is addressed in this thesis by experimentation. The extent to which rhythm analysis reveals knowledge about human mental processes is addressed in anthropological terms in the next section.

1.3 Anthropology of Computer Music Research

Models of musical time must address the issue of cultural specificity. Comparison of Western and non-western musical behaviour and perception can be used to further distill elements of universality of rhythm perception. Most music studied and reported in music psychology literature is within the bounds of traditional Western musical thought, and notions of metricality. While this does indeed cover a wide range of possible contemporary art and popular music, a model built using such research is implicitly constrained by the degree metricality can adequately represent music not conceived within the theoretical paradigm of meter, such as some avant-garde Western music, and non-western music. Understanding the degree to which a multiresolution approach can address such genres provides for worst case testing of the concept.

1.3.1 Enculturation of Music

John Blacking has proposed that music is a result of a synthesis of cognitive processes of a particular society, with processes of biological origin [117]. In psychological terms, societal knowledge as external actions is *internalised* to become internal actions, communicated through semiotic mechanisms to convey meaning to the individual. Enculturation of the individual is argued by Moisala as a two-stage process, initially by perception of sound in interaction with the outside world, and then by a process of organisation of that internalised knowledge within *intrapsychological* categories [117]. Indeed, Moisala argues that in order to understand musical cognitive processes it is necessary to study musical practices and performance within a cultural context. Not merely the auditory result, but the spatio-motor and theatrical elements in the production are essential in communication of meaning.

There is then always a question hanging over any research to the degree that investigations reveal what portion of perception is culturally informed and what processes are universal. This universality may be either from biological processes or cognitive constructs which are from a cultural source which is so fundamental to societies, that it is a common "wisdom". Any model which proposes to use neurologically influenced architectures (i.e neural networks [182]), must clearly identify the degree to which musical knowledge is universally coded, versus culturally constructed, if there is an attempt to match computational models against cellular

recordings and findings from neuroscience.

In order to build robust functioning computational models of musical rhythm it is important to review how rhythm is conceived in other cultures and the process of reception of new music into an existing culture (from another culture or from within). This argues for the need to embody any computer system with a priori context or otherwise train the system on a corpus of music.

1.3.2 Rhythms of New Music

While music theory must implicitly draw on perceptual constraints, from a modernist perspective, the enunciation of a theory has given composers mental models with which to conceive new music, driving performers and listeners to develop new modes of listening. This has an impact on the degree to which psychological models of listening are indeed inherent. The intentional convergence between expressive timing and phrase structure, reflected in proportional rhythmic notation[1] in contemporary Western art music, indeed calls into question the role and separability of expressive timing [15].

Minimalist experimental pieces such as sound installations and ambient music by LaMonte Young [36] and more popularly, Brian Eno [176], are examples within Western music of composers/performers perhaps intuitively refocusing the listener's perception to depend on longer auditory memory in preference to typical rates of beats which fall within short term memory. The engagement of a different auditory rate may well explain the distinct restful or contemplative mood that such music can bring. At another extreme the monumental polyrhythmic player piano studies of Conlon Nancarrow [46] push to the limits the listener's comprehension of the composers intention. It can well be surmised that the listener is distilling a subset of the rhythmic information presented, interpreting those streams of sound which the listener's segregation by melody and timbre [10] highlights. Purposefully non-rhythmic music, such as the aleatoric compositions of John Cage [13] or some forms of free improvisation [3], challenge existing notions of musical organisation, but as Cage has recognised, the overall structural form of a performance remains as a coherent whole.

[1]The horizontal distance on the stave between notated beats indicates directly the time between events. Works by Stockhausen, Ligeti and Boulez have all adopted such a notational convention [49, 15].

1.4 Thesis Structure

Suitable input representations must be devised in order to construct a computer system to model the perception of some aspect of music. Considerable music psychology literature has identified the complexity of tonal representations and the interrelated influence that tonal expectations have upon rhythm and phrase structuring and vice versa [78]. To limit the problem domain, the task of modelling the interpretation of rhythm has been adopted. This forms a domain which is of itself phenomenologically complete, in that music constructed entirely from indefinite pitch percussion can be listened to and appreciated [49, 13]. While this reduces the number and semantics of objective perceptual cues to be considered in computational models, there is considerable complexity in the phenomenon of musical rhythm.

The complexity of musical rhythm and the variable successes in existing models is a strong argument for developing new methods of representation of rhythm which reflect its perceptual features. Chapter 2 surveys music psychology, musicology and ethnomusicology literature illustrating the layered, hierarchial nature of rhythm as conceived and performed in Western and non-western musics. The hierarchy of musical time and the effect of expressive timing has a natural description in terms of time-varying frequencies. In Chapter 2 this perspective is explored in depth.

An analysis technique which has shown considerable success in analysing time varying frequency signals is the continuous wavelet transform (CWT). Such analysis has shown its worth in analysing auditory signals. Chapter 3 investigates the ability to analyse *rhythmic* signals using such approaches, particularly the degree to which such a conception matches listeners' perception of rhythm detailed in Chapter 2. The purpose behind the analysis is to reveal more detail of the signal than that available from the time-domain representation before building cognitive models. With a clearer representation of the signal it is then possible to construct time-frequency based interpretative models.

The capabilities of multiresolution analysis when applied to rhythmic signals are demonstrated on a corpus of test rhythms in Chapter 4. While these are not the only rhythms that have been tested with the analysis, they are representative of the results obtained in all cases. The mathematical proof of decomposability using the continuous wavelet transform does not guarantee the constituents will meaningfully reflect principal rhythmic attributes. This chapter assesses experimentally the generality of the approach, rhythms exhibiting meter change, agogic accent, ritard

and acceleration, and several other forms of rubato are analysed. Both synthetic rhythms and well known rhythms used by other researchers are tested.

Having investigated and interpreted the results of rhythmic analysis manually, several related approaches to automatic interpretation of the analysis are described. Wavelets allow us to review the original data from a different time-frequency perspective and begin to propose cognitive models. Chapter 5 investigates a now well defined computer music "problem" of foot-tapping. This problem is approached using the analysis results of Chapter 4. The results demonstrate the representative power of the multiresolution approach and the benefits of an interpretative measure constructed from such. The outcomes of the research, new contributions, and future applications are assessed in Chapter 6.

Chapter 2

The Hierarchial, Multiresolution Character of Musical Rhythm

> "To adequately portray rhythm, one must shift from descriptions based on traditional acoustic variables to one based on diverse interactive levels. It is clear that the overall patterning of the acoustic wave brings about the perception of a rhythm, but the rhythm cannot be attributed to any single part of the wave ... It is the independent production of the various rhythmic levels that allows the elasticity, the *rubato* of music, as well as the independence of stress and accent in speech and the independence of meter and grouping in music. In the same way, it must be the parallel perception of these levels that allows for the perception of rhythm."
>
> **Stephen Handel "Listening"** [55, pp. 458].

Inter-related effects of different dimensions of music impinge on its perception. It will be shown in this chapter that certain dimensions of music are interpretable even in the face of impoverished attributes from other dimensions. That is, there are dimensions of music which are structurally significant enough to warrant, and can withstand, independent investigation. These dimensions include those that have classical definition within music theory, most notably pitch, rhythm and dynamics (amplitude). The dimension of rhythm is investigated here by reviewing findings of music perception research and music theory.

Clarke defined rhythm as: "the grouped organisation of relative durations without regard to periodicity" [14]. Dowlings definition is: "A temporally extended pattern of durational and accentual relationships" [35, pp. 185]. The definition from

Parncutt [130] is that musical rhythm is an acoustic sequence evoking a sensation of pulse.

Research in rhythm perception has been marked by complexity and competing theories, which reflects the complexity of our temporal perceptual processes. Context sensitivity, the interrelationship of timing and melody, and an absence of invariant perceptual features are contributing factors to the complexity of rhythmic analysis [55]. Multiple perceptual systems are posited as involved in temporal and rhythmic processing and produce a multidimensional perception.

Both the perception and the production of rhythm are demonstrated here as processes that occur over a wide range of time scales. The context for listening to a rhythm is created by the interrelationship between perceptions of inter-onset intervals over multiple time spans. The production of a rhythm by performers involves a conception of a beat to be performed within an intended context of impending events. The performer conceives the rhythm as an end result of a number of parallel intentions in time, reflecting knowledge of the rhythmical context and pace of the performance.

The aim of this chapter is to elucidate the essential structural information which is communicated to a listener. This is due to the interplay of expression against the structural base in relation to one or more metrical contexts. Musical meaning and intention are communicated from the performer to the listener by relating timing deviations to tension/relaxation principles as proposed by theorists such as Meyer [112] and Sloboda [159].

2.1 Timing Behaviour and Constraints

Fundamental to any conception of rhythm is the perception of timing by the listener. Musical time perception is bounded by a listener's capability of perceiving audible events. Rhythmic definitions differ between authors; drawing from Lerdahl and Jackendoff's view, a *beat* can be defined as a durationless element, a time point [84], typically in relation to a temporal structure, such as a meter (Section 2.2.4). The *Inter-Onset Interval* (IOI) between onset times of audible events measures timing intervals. Audible percepts can be distinguished between stimuli of short sounds, such as impulses or clicks, and of sustained tones [190]. As will be seen in Section 2.2.1 on accentuation, these differences can be reviewed as two extremes of dimensions which cause accentuation. Fundamental to either case is the relative timing of the

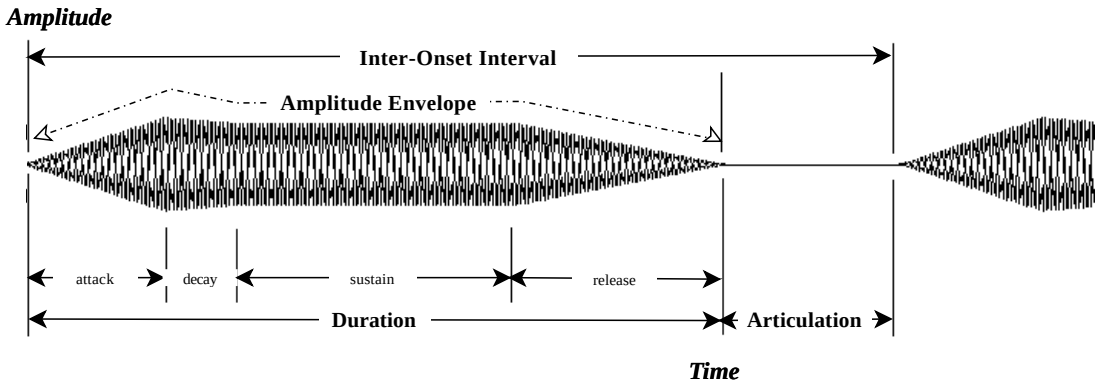

Figure 1: A time/amplitude plot of a succession of two notes, both with a fundamental frequency of 440Hz (concert pitch A), with a piano-like amplitude envelope and a pure sinusoidal timbre (no harmonics). Also shown are the timing terms used to describe musical rhythm.

IOIs between events. Other common notions and terms used to describe the timing of musical events are displayed in an amplitude-time graph of Figure 1.

Conceptions of time enable two notable behaviours: synchronisation with a perceived rhythm, and the notion of the present. These behaviours are dependent on low-level auditory processes including neurobiologic clocks and masking effects. Reviewing the literature on these behaviours and auditory processes informs musical rhythm modelling; however, studies in auditory time perception have often used isolated, non-musical stimuli. The absence of a rhythmic context may therefore be skewing reported results towards the limits of perception rather than common performance.

2.1.1 Synchronisation

As a rule of perception across modalities, subjects (i.e. listeners) react following a stimulus, yet synchronisation produces responses at the same time as the stimuli, so the temporal interval itself is driving the response. Regularity seems to be less fundamental to synchronisation than a listeners ability to anticipate and predict, such that accelerating or slowing rhythms can still be synchronised (at typical rates). For regular patterns, synchronisation can be established very quickly from the third sound on and synchronisation of repetitive patterns is achieved from the third pattern

on [41]. Developmental research indicates that perception of synchrony is potentially biological; children as young as 4-months preferred synchronisation between a visual stimulus and accompanying auditory cue in preference to unsynchronised stimuli [170, 35, pp. 194]. Children's earliest spontaneously performed songs have steady beat patterns within phrases [35, pp. 194]. Thus a model of musical time must be capable of generating synchronisations after two or three beats and model expectation.

2.1.2 The Subjective Present—Our Concious, Ongoing Experience

Subjective present is a term characterised as "the feeling of nowness" [134, 155], a span of attention, a window over time, or the interval of temporal gestalt perception [41]. It is considered the interval where all percepts and sensations are simultaneously available for processing as a unit [130]. The subjective present has been argued by Dowling and Harwood to be the perceived sense of sensory or *echoic*[1] precategorical acoustic memory, a brief store of unprocessed auditory information [35].

Time-spans of the Subjective Present

Evidence presented by Cowen suggests that auditory stores are integrating buffers of a continuous stream of data, rather than a discrete, gating buffer. The integrating period is proposed to be limited to the first 200 msec, with a relatively constant decay in memory recall of events from 500 msec delay and longer (see Cowen's comparison of results [19, pp. 353]). This integration creates non-linearities between the initial processing of stimuli and processing for longer retention of the events.

Seifert has proposed that in order for a musical phrase to be perceived as a structured entity, the total length must remain within the time span of the subjective present [156, pp. 174], citing the fact that cycles of rhythms in African and Arabic music are roughly 2800–3000 msec duration. From this, Seifert has hypothesised that all repetitive (or perhaps all expectable) time patterns must lie within the bounds of our subjective present [156]. Woodrow [190] reported a relatively constant just noticable difference (JND) discrimination of single intervals bounded

[1] In comparison to visual *iconic* memory.

by short audible clicks over a range of intervals from 200 to 1500 msec. As noted by Dowling and Harwood [35, pp. 185], this JND was significantly higher reproducing isolated intervals than when discriminating intervals in the context of a repeating beat pattern.

Dowling and Harwood have reported the window's size bounding the perception of the present ranges from 2 seconds to rarely more than 5 seconds. A maximum of 10–12 sec of present is only achievable by "chunking" (i.e. grouping) long sequences into sub-sequences [35]. For the purposes of his model Parncutt has surveyed literature to estimate the maximum echoic memory length to be 500 msec and the subjective present as 4 seconds. He reports experimental results estimating subjective present ranging from 2 to 8 seconds. The perception of a sense of pulse is argued by him to be limited by the time span of the subjective present [130]. The span of subjective present is dependent on the IOI: "faster the presentation rate, the shorter the memory span" [130, pp. 451]. Yako has likewise argued for conceptions of subjective present weighted by their location over hierarchies of time [191].

Short and Long Auditory Stores Within the Subjective Present

Within the single integrating period of the subjective present, several shorter integrating periods occur. Seifert conjectures two levels of cognition, the perceptual and the cognitive. The former is automatic, forming a pre-cognitive integrating function. In his view, cognitive level processing is not automatic and is under conscious control [156]. His definition of cognitive processing may differ from the typical definition, but what does seem clear is that in attending to auditory sequences, there is a distinction between material which can focused apon using learnt knowledge and that for which the processing is automatic. Bregman [10] describes these as primitive and schema-driven auditory scene analyses, discriminated by the latter's requirement for attention and concious control.

Cowen has surveyed evidence of two different auditory memories, a Short Auditory Storage (SAS), "a literal store that decays within a fraction of a second following a stimulus" [19, pp. 343] and a Long Auditory Store (LAS), lasting several seconds. While described as storage, it is more probable that patterns of neural activity over timespans produce "levels of processing" [19, pp. 363] from which arises the functional equivalent of time limited storage.

Short Auditory Store (SAS)

Short term auditory memory is more directly representative of the original stimulus than LAS, and time-limited. Cowen compares a number of experimental findings measuring SAS temporal integration and decay, concluding that SAS is between 150–350 msec constant duration from stimulus onset, experienced as a sensation, and consisting of a recency biased average of the spectral components of the presented sounds. When interference with the SAS occurs from distracting stimuli, that interference is unable to be prevented, resulting in the loss of existing memory of events, a phenomenon known as *masking*, described in Section 2.1.4.

Measures of the persistence of an auditory stimuli indicated sounds shorter than 130 to 180 msec (depending on visual or auditory cues) were judged by subjects to be of equal length to sounds actually of that duration [19, pp. 343]. Such minimum time measures suggest some form of constant integrating process occuring over a timespan of the SAS duration.

Long Auditory Store (LAS)

Long auditory store (LAS) within the subjective present is summarised by Cowen as lasting a maximum of 2 to 20 seconds or more, experienced as a memory of features of a sound sequence, most probably from SAS, and stored as partially analysed input. Stimuli interfere with previous stimuli only partially, and total masking does not occur, unlike SAS. Estimates of the duration of long auditory memory has varied across published studies, possibly as a result of varied quantities of information inherent in the stimuli aiding recall. From Cowen's review of these, there does however, appear to be a trend of a rapid decay of storage in the first 2 seconds, with a slower decay out to at least 10 seconds. In contrast to SAS, at LAS periods, no minimum persistence effect has been reported.

2.1.3 Hierarchies of Time

Seifert cites research supporting the theory of mental "time quanta" and pulse [130] and describes Pöppel's taxonomy of elementary time experiences [134] with respect to rhythm perception as a means of describing levels of such time quanta:

- Events consisting of short term sound bursts within 0 to 2–5 msec are perceived as simultaneous, indistinguishable (even with different loudnesses, but same

duration), as a single event.

- Events from 2–5 to 40 msec apart can be distinguished, but no order relation can be indicated.

- Events above 30 to 50 msec apart can produce an order relation (i.e order between the events can be distinguished).

In Seifert's description of Newell's time constraints on cognition, identified as different temporal "bands", the cognitive band is quoted as: " ... the apparatus necessary to go from neural circuits (the top level of the neural band) to general cognition that requires four levels. It takes us 10 msec up past the level of immediate external cognitive behaviour at [approximately] 1 second" [155, pp. 291]. Neural circuits are claimed to act within 10 msec, and cognitive behaviour to occur within approximately 1 sec, resulting in 100 steps or time periods to produce cognitive behaviour. This sets strong restrictions on the architecture used for cognitive modelling. According to Newell, the real-time constraint on cognition is: "only [approximately] 100 operation times (two minimum system levels) to attain cognitive behaviour out of neural-circuit technology." [155, pp. 291].

Seifert argues that 30 msec lower bounds are to be expected for rhythmic discrimination abilities, due to a similar performance in discriminability between closely occuring auditory events. However, when considering expression, particularly rubato effects (including phrase final lengthening [97, 98]), and accelerando/rallentando (tempo deviations), listeners discrimination abilities may well be quite different as they are then judging deviation times—as slight as 5–2 msec, (200Hz–500Hz) from a context of beats falling at a fundamental frequency on the order of 3 sec period (0.3Hz).

Handel estimates a 50 msec lower bound on the IOI between events to be perceived as a sequence, rather than a continous tone [55]. The percept of a continuous stream may well result from auditory persistence discussed in section 2.1.2. Handel estimates 1.5–2.0 secs to be the longest IOI before the sense of repeating sequence changes to a sense of isolated events.

From the above literature, one can conclude that listeners have absolute and relative memory limits. Within a short span of a few seconds which constitutes a perception of subjective present, a number of temporal bands exist, rather than a continuous equally perceivable range. These ranges are summarised in Section 2.4.2.

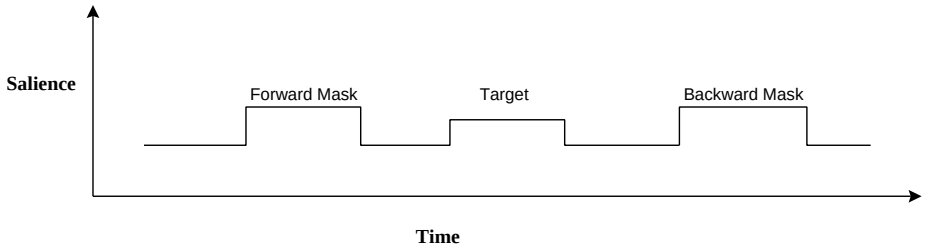

Figure 2: Illustration of the reducing effect on perceptual salience of a target tone when in close temporal proximity to either a forward or backward masking tone. The masked sound's intensity will be perceived as reduced, even to the point of it becoming imperceptible.

These limits influence the process of grouping temporal events, establishing limits a computational model should address.

2.1.4 Masking

Auditory masking is the phenomenon of a sound modifying a listener's perception of another sound. Masking diminishes the perception of one signal due to the temporal or spectral proximity of a second (see Figure 2). Masking occurs most strongly at short delays, decreasing to an ineffective degree past approximately 200 msec [19, pp. 346]. Massaro demonstrated backwards masking [96, 19], finding that when presenting listeners with two short duration sounds, the preceding sound (the *target*) can go unnoticed when followed in rapid succession by the second sound (the *mask*). Forward masking occurs where the decaying trace of the earlier mask sound can affect detection of the target sound [19].

The effects of masking suggest the existence of some form of SAS which holds auditory events as a trace across a short time period of 200 to 300 msec. Cowen has proposed that forward masking occurs from the persistence of the memory of the mask and accordingly, backward masking interferes with the detection of a sound by interrupting the auditory persistence in SAS arising from the earlier tone's duration. For total masking to occur, rendering the target tone inaudible, the mask must also be of longer duration or more intense than the target [62, 19].

Todd has proposed that summation of echoic memory peak responses provides a mechanism to model such masking. Backwards masking is proposed as interrupted temporal summation and forward masking as incomplete recovery from adaption.

The end-product of the interrelation of temporal integration, adaption and "enhancement" creates accented events if these processes collectively increase total neural activity [103, pp. 41].

2.1.5 The Neurobiological Basis Of Rhythms

There has been a long history of investigation into the degree of association between other body functions and temporal perception. Early research unsuccessfully attempted to find a direct connection with walking pace or with the period of word utterances [190, 41]. Early childhood motor actions such as sucking and rocking have periods of 600 to 1200 msec, and 500 to 2000 msec respectively, which fall within the range of spontaneous and preferred rhythmic rates [41]. It may be that biological processes such as breathing, walking or heart beats form underlying cues for qualitative judgements of duration. As Woodrow has noted, and musical pedagogy commonly adopts, the act of counting to oneself is a common method of accurate quantitative estimation of time which can be used to extend time estimation into periods of several minutes [190, pp. 1235]. Of course, this is using a cumulative estimation of periods falling within the bounds of SAS.

Several researchers have proposed the presence of internal clocks. The use of an internal clock can serve to plan when a new action is required, to act as a temporal goal. Shaffer studied the performance of a skilled pianist in varying a polyrhythm in two hands with respect to the tempo of both hands together, and independently [157]. The pianist's ability to perform such variations suggest separate time-keeping levels. Handel suggests the existence of separate clocks for each hand together with a higher-level clock which can *entrain* the lower-levels and provide reference timing [55]. It is perhaps more feasible that we have a number of clocks which can be assigned to be, or are intrinsically, dedicated to specific anatomical motor control and also function to provide clocks which allow for rhythmic planning and prediction.

As Seifert, Olk and Schneider [156] note, there is a strong connection between action (motor behaviour) and perception, and a motor theory of perception is proposed by them as the most suitable model of rhythm perception, with the understanding that this does not necessarily imply a relation to neurological functions. Todd [102] has taken issue with the neurobiological clock proposals, but has proposed that biological processes do contribute to absolute time constraints which produce preferred pulse rates and therefore mediate rhythm perception. Todd's model proposes

a motor basis to rhythm perception, thus amalgamating production and perception tasks. He proposes a body sway rate of 0.2Hz (5 seconds interval) and foot tapping frequency of 1.7Hz (600 msec period) as centres of maximum pulse salience. Further, he controversially proposes the vestibular system, normally attributed to providing the body's sense of balance, as responsible for the perception of auditory rhythms. However both foot-tapping and body sway rates are derived from production behaviour, not perception aspects. It is unclear how such proposed biological biases as these would still allow such a wide choice of tempo behaviour and fluid shift between tempos that is seen in musical rhythms.

2.2 Principal Rhythmic Attributes

Certain properties of rhythm are explicitly represented in Western music theory: tempo, relative durational proportions of events and rests, and meter. In some recent Western music, grouping relations may be notated by slurs [15], however most grouping and other properties emerge from continuities or discontinuities between elements, and interactions between a listener's sense of timing, pulse and accentuation. Both explicit and emergent dimensions of rhythm are now detailed, noting their character and interrelationship.

The dimensions of musical rhythm are reviewed here to characterise musical processes which any computational model must address. Rhythmic information is more fundamental to music cognition than pitch. Early research showed familiar tunes can be recognised by their rhythmic patterns alone [35, pp. 179]. Rhythmic information dominates over pitch in multidimensional scaling tasks to determine primary stimulus dimensions.

The use of multidimensional scaling of similarity judgements, comparing pairs of rhythmic stimulus patterns, has produced dimensions closely matching a layered rhythm model. These dimensions correspond to differences in meter and tempo, accent of the first beat, patterns of accents and durations, variation versus uniformity and rigidity versus flexibility. Dimensions of affective meaning were also prominent: excited versus calm, vital versus dull; and character of movement: graceful versus thumping, floating versus stuttering [55, 41]. Multidimensional scaling has also been applied to combined melodic and rhythmic patterns. Major dimensions were 2-element versus 3-element patterns and inital-accent versus final-accent. Melodic contour was only the third significant dimension in similarity judgements. Rhythm

is a more distinctive parameter and important in music cognition [35]. Performers' actions are (perhaps unconciously) intended to communicate (or create) these dimensions and factors.

2.2.1 Accentuation

Accentuation produces difference between musical notes, distinguishing accented sounding events from temporally adjacent ones. Effectively, the establishment of difference between events allows extension of the perception of auditory processes over timescales longer than the SAS. Tangiane has argued that a rhythm only occurs when a periodic sequence can be segmented into groups [177]. Fraisse asserts the basis of rhythm is the ordering in time of temporal relationships between events, rather than the notion of rhythm arising from patterns of accentuated beats. Evidence for this comes from variation between listeners in identifying which beats are accented sufficiently to indicate a *downbeat* [41].[2] Parncutt measured subjects tapping to *isochronous* patterns[3] and also found wide variation in choice of downbeat, the variations included tapping at the notated rate (the tactus) but with a phase shift [130]. As will be shown below, temporal and accentual influences on the perception of rhythm are interrelated [41, pp. 151].

Lerdahl and Jackendoff have distinguished accents into *metrical*, *phenomenal*, and *structural* types according to their effect on groups [84]. Metrical accents occur where the emphasised beat is related within a metrical pattern (repeating regular accentuation of beats). Phenomenal accents are considered to exist at the musical surface, emphasising a single moment, enabling syncopations (accents out of phase to the underlying meter) to be perceived. Structural accents are defined as "an accent caused by the melodic/harmonic points of gravity in a phrase or section—especially by the cadence, the goal of tonal motion" [84, pp. 17]. A structural accent is therefore perceived with relation to the unfolding phrasing and structure of the music longer than a measure's length, whereas a metrical accent is perceived within a recurrent short time span defined by the meter.

Accentuation is achieved by objective differences between sounding events, enabling the grouping of sounds in time. As will now be described, a listener will also *subjectively* accent sounds which are in fact isochronous, in the absence of objective

[2]The downbeat is the first beat of a metrical pattern and usually perceived as accented.

[3]Sequences of sonic elements having identical interval, pitch, timbre and intensity. Fraisse uses the term rhythmic cadence [41].

accents.

Subjective Rhythmisation

Subjective rhythm is a historical term used to describe the grouping of isochronous pulse trains into twos, threes or fours. The first element of the group is perceived as accented, and the interval between last element and the first element of the next group is perceived as lengthened [41]. In modern terminology, the term subjective metricality is now more appropriate [80]. Subjective rhythmisation evokes a sense of pulse whose period is longer than that of the stimulus [130, pp. 421].

The relative length of a silent interval following a tone in equitone sequences is a determinant of the perceived accent on that tone. Povel and Okkerman [137, 41] varied both the first or second IOIs between tone pairs in otherwise isochronous (equitone) sequences. They sought to determine the interval times that the accents would be perceived on the first or the second tone in the pair as a subjective rhythmisation. When the interval difference between pairs of tones is short, the accent is judged on the first tone of the pair, as the interval increases past 220 msec, the accent is more often heard on the second tone of the pair.

Their second experiment sought to determine if the accent on the first tone was a result of perceived grouping or "an orienting response to that tone". This orienting response may have occured from the fact that a long interval preceded the first tone [137, pp. 568], conditioning listeners to the phase of the rhythm. Only small differences in results to the first experiment were found when preceding the very first tone of the stimulus sequence with a longer tone, and the orienting response was concluded not to be completely responsible. The third experiment sought to differentiate the effect from "energy-integration", which was hypothesised to be the result of overlapping the decay of each tone with the attack of the next tone. Even though there is a slight effect by increasing the articulation of the tones (onset to offset interval), it was not considered to significantly modify perception. The fourth experiment had the subject adjust the strength of the first tone until no accent occured on the second tone. An increase of up to 4 dB was required to balance the interval-produced accent, showing that the interval-accent is a robust phenomenon. The experiment also showed the interval length was proportional to perceived accent strength.

Presentation rate made little difference to the subjective rhythm. Outside 115–1800 msec intervals, when the two events are no longer perceptually linked, subjective rhythmisation becomes impossible [41]. Parncutt found grouping of isochronous pulses into fours in preference to threes was general and independent of tempo in a tapping task [130]. This is argued by Parncutt to be the result of more consonant pulse sensations (strata) falling with the existance region of pulse sensation when grouping by fours (at frequencies of 1/2, 1/4 and the pulse rate), than by threes (only at 1/3 and the pulse rate frequencies). Subjective rhythmisation demonstrates that the process of grouping temporal elements into longer term structures will occur even when not supported by objective differences. This suggests the temporal intervals themselves are responsible for accents and determination of rhythmic structure.

Objective Accentuation and Rhythmisation

Table 1 shows a summary of common objective differences introduced between sounds by a human performer in order to induce grouping. The two most prominent accentual forms are intensity and duration. Intensity is commonly a direct increase in amplitude of the produced sound wave, but the nature of musical instruments is such that increases in loudness are typically accompanied by change in timbre. Plucking a string harder will change the spectral character of the sound as well as its amplitude, with similar interactions occuring when performing on percussive, wind and bowed string instruments. Thus, the musical performance concept of using dynamics to convey accents is, in fact, a multi-dimensional percept in the mind of the listener, and of course, the performer. Such a synthesis of perceptual dimensions constitutes a behaviour a computational model must address.

Timing Accent

Timing intervals can function both as accents, and as notable pauses between groups. Interval lengthening will cause grouping with the interval demarcating the boundaries of one group and the next. When the lengthening is only slight, the accent is perceived as being on the beat following the longer interval. When the lengthening is large, creating a *pause*, the accent is perceived as on the beat prior to the longer interval [41, 15]. Woodrow has shown that once can change a listener's perception

✳ Lengthening of an IOI between two events.
✳ Increase in intensity.
✳ Relative intensity profiles between beats.
✳ Variation in articulation (legato/staccato).
✳ Change in pitch or at extrema of pitch trajectories.
✳ Sources and destinations of harmonic progression.
✳ Difference or change in timbre or instrumentation.
✳ Onset synchrony between voices of same instrument.
✳ Onset synchrony between voices of different instruments.
✳ Density of events in time spans (fast runs, trills, tabla fills, grace-notes).
✳ Phrase final lengthening, rubato effects, deviations in time.

Table 1: Common objective accents used in performance.

of a rhythm from trochaic to iambic[4] by shortening the IOI following an initially perceived-second soft sound to make that soft sound be perceived as leading the group [190]. Slightly lengthening the IOI following a sound conveys an impression of increased intensity, forming an *agogic accent*; and reciprocally, intensifying a beat creates the perception of the sound having a longer IOI. As well as onset to onset (IOI) time, the onset to offset time, or in musical terms, the *articulation*,[5] of a sound is well known to create an impression of increased loudness [19]. Clarke proposes that articulation only acts as accentuation, without a direct impact on the rhythmic structure [15].

Lerdahl and Jackendoff's series of "metrical preference rules" seek to codify the location of accents using interactions between intensity and duration [84] (see section 2.2.4). Their preference rules attempt to account for listeners' placement of accentuation on beats and suggest which assignments are most musically appropriate within bounds of individual choices and plausible differences [55]:

[4]Traditionally, common rhythms have been described using ancient Greek terms of rhythmic "feet" associated with the pacing of poetry describing the order of accentuation [41, 35, 112]. An *iambic* rhythm describes a pattern (typically repeating) of 2 syllables, the first unaccented, the second accented. The *trochee* is a rhythm of 2 syllables, the first accented, second unaccented; the *anapaest*—3 syllables, 2 unaccented, followed by an accented one; conversely the *dactyl*—3 syllables, first accented, then 2 unaccented. The *amphibrach* describes groups of 3 syllables, with the accented syllable between two unaccented.

[5]The term articulation is often broadly used or misused to describe rubato. In the course of this thesis, it will refer to the onset-to-offset time interval.

- Strong beats fall on elements with higher intensity or longer duration.
- Strong beats fall at the beginning of intensity changes (crescendo/decrescendo).
- Strong beats fall at the beginning of changes in articulation.
- Strong beats fall at the beginning of slurred notes.

Preference rules were proposed by Povel and Essens [136] concerning identical elements separated by different length silences:

- A strong beat should not fall on a rest.
- A strong beat should fall on the first or last element of a sequence of identical elements.
- A strong beat should fall on the same position in repeating phrases.
- Strong beats should occur in two beat or three beat meters.

The interaction between duration and intensity and their interchanging roles thwarts interpretations of rhythm built purely from either durational or accentual percepts. However, there is variation between the degree of effect these dimensions have. Fraisse has surveyed research [41] showing durations were less varied than intensity accents in performances of repetitive rhythms, with durations varying between 3–5%, whereas intensity based accents were varied between 10–12%. Durational accent is also observed by Parncutt in his experiment testing metrical accent perception [130]. However, he reported a contradiction to the rule of an accented event preceding a longer IOI, finding for the case of listeners tapping to a march rhythm, the IOI preceding an event had a stronger accentual effect than the IOI following the event. Given the conformance of listeners to expected results for nearly all other rhythms tested, experimental error seems unlikely, this tends to suggest codifications of accent placement on the basis of durations may be missing aspects of structure.

Pitch interactions with rhythm perception

In describing principal attributes of musical rhythm, it has been assumed there is a separability between those dimensions and pitch. A more accurate characterisation

of the situation would be that these dimensions are partially coupled. Krumhansl's review of interactions between melody and rhythm noted: "Clearly, both aspects influence music perception and memory, but it is unclear whether they are truely interactive or simply have additive effects. A number of studies show that rhythmic structure influences judgements of pitch information" [78, pp. 297]. The recognition of a metrical melody is facilitated by the correct recognition of the melody's meter and downbeat [131, pp. 150]. The meter and downbeat is considered as a pre or co-requisite for the recognition of the melody.

Jones and others [68], manipulated accent structure and obtained effects on pitch recognition. They proposed listeners allocate attentional resources over time (elaborated in [66]). They found poorer judgement of change in melodies when presented in a rhythm different from a reference rhythm, and when melodies were presented in rhythms which received local increases in their IOI's. The later case rendering incompatible rhythmic grouping at points either between or within melodic groups [68, 65]. The results indicated rhythm is functioning to direct attention at specific timepoints, aiding discrimination. Palmer and Krumhansl [125, 126] found "pitch and temporal components made independent and additive contributions to judged phrase structure. However, other results suggest interactions between temporal and pitch patterns" [78, pp. 297]. Lerdahl and Jackendoff's metrical preference rules concerning variation in pitch have proposed that:

- Strong beats fall at large changes in pitch.
- Strong beats fall at changes in harmony.
- Strong beats fall at cadences.
- Strong beats tend to fall on lower pitches.

Handel has reviewed interrelationships between rhythm and pitch. He has suggested that highest pitches in a sequence tend to be perceived as accented. In alternative rhythmic contexts, the least frequently occuring pitch is likely to be perceived as the accented element. Another candidate element for perceiving as accented is the pitch forming the local maxima of a rising then falling melodic fragment. Alternatively, the element following the local pitch maxima can function as the accent when it forms the start of a melodic contour [55, pp. 388]. The confluence of melodic (e.g. first note succeeding pitch jumps) and temporal accenting

(succeeding rests) will lead to varying perceived strengths of the beats. Coincidence of accents produces a strong beat and an emergence of meter. Points where melody and timing accents do not coincide results in weaker beats which are of an irregular pattern, and a meter does not emerge.

In summary, the dimensions of pitch and rhythm are clearly interacting, but there are they are psychologically separable—we can perceive them separately, but they interact as they build up our multidimensional perception of music. This produces a need to isolate studies of rhythm to an interrelated set of perceptual features. To limit the problem domain for a computational approach, the use of percussion music has advantages. Percussive tones are often inharmonic, creating complex pitch implications which avoids the overlearned grouping from melodic/harmonic tonal cues.

2.2.2 Categorical Rhythm Perception

Evidence has been summarised by Fraisse [41], Povel [135], and Monahan (reported by Dowling and Harwood [35, pp. 187]) of patterns in 2:1 ratios between elements as being easy to perceive and reproduce. Analysis of examples of Western classical music by Fraisse showed 80–90% of notes were in 2:1 ratio between note durations, with the longer of the two durations in the range 0.3–0.9 sec. This spans the 600 msec interval prefered pace (see Section 2.2.6).

As noted in section 2.2.4, rhythms tend to be categorised into subdivisions of small primes, typically 2 and 3, in a similar vein to tuning systems construction on small prime limits [189, 132]. Sloboda has identified behaviours which are suggestive of a categorisation (i.e quantization) of the duration of notes into the subdivisions of multiples of two or three. He cites the inability of performers to imitiate another exactly, the extraction of structure from rubato passages, and the difficulty of perception of metrical deviation (under a threshold) as examples of categorisation. Experiments by Sternberg and Knoll [172], also described by Sloboda [159], showed skilled musicians were unable to accurately reproduce or perceive rhythms which were non-standard subdivisions of the beat.

In tapping experiments, first performed by Fraisse [127, 41, 55], with later support by Povel [135], subjects simplified a variety of complex ratio rhythm intervals.

In reproducing a rhythm, listeners reduced to just two interval durations: short (between elements, 150–400 msec) and long (between groups, 300–860 msec), demonstrating a preference for 2:1 ratios.[6] This categorical preference in production or *metrical categorisation* also appears in statistical regularities of IOIs as notated in scores of Western composers from the Baroque through to Modern eras. Frequency distributions of intervals notated showed that just two durations were most frequent, typically a crochet and quaver, forming either 1:1 or 2:1 ratios between IOIs, with the shorter IOI the more frequent [41]. The preponderance of 2:1 duration ratios suggests there are few relative levels of metrical hierarchy formed if IOI is the only metrical cue. The candidate is chosen on the basis of economy of perception, favouring simple duration ratios.

Sloboda has argued that categorical rhythm perception does not exclude perception of finer temporal structure, but he argues that it produces changes in "quality" in the same manner as slight tuning variation produces the psychophysical percept of "roughness". This seems hard to accept when well trained musicians are able to repeatedly reproduce their subtle variations in timing [157, 97] to demonstrate rubato, whereas Sloboda's listeners are distinguished between the majority that perceive timing variation as simply a quality and others (i.e. musicians) who are able to perceive it (by implication) in more structural terms.

The production of jazz "swing" rhythms in group scenarios are characterised by highly accurate deviations from metrical time locations are possible by master players [2, 139, 17]. Desain and Honing's positive results for a quantizer that stretches over several time contexts [25], would appear to demonstrate the contextual basis of the categories, and their inability to simply be reduced to nearest immediate neighbourhood operations. It seems more likely that a "swung" rhythm is structured as phase shifted inharmonic partials of lower frequency metrical strata, but that this does not exclude the metrical stratum's perception or production. Rather, this produces a rhythmic richness and tension by the counterplay between the implied meter and the stated events.

2.2.3 Grouping

A general rule common to all perception is that elements tend to be placed in equal size groups larger than two elements [55]. Grouping in rhythm is the assignment

[6] This would be notated ♩♪ which forms a typical galloping rhythm when repeated.

of temporal structure to an auditory stream, and is considered responsible for the concept of musical phrases or motives. Grouping appears very early in life, and is seemingly a spontaneous behaviour [41]. From the effects of subjective rhythmisation, grouping will occur even when not supported by objective accents, suggesting grouping is an independent process reliant on timing, with objective accents confirming and increasing the perceptual strength of grouping boundaries.

Forming groups of 2,3 and 4 elements is significantly easier than groups of 5 and 7, suggesting that relative timing, economy of attention and representation [55], and limits of auditory memory, rather than accents alone, are contributing to group organisation. These longer rhythms tend to be grouped as subgroups of 2 and 3's, suggesting that hierarchies of grouping are spontaneously organised [35]. Fraisse has estimated an absolute maximum of 25 successive sounds that can be perceived as a unit [41].

The duration of groups are limited: "One can perceive groups of from two to six sounds that correspond to the boundaries of our immediate memory or of our capacity of apprehension" [41, pp. 157]. Fraisse reports an interaction between the number of elements and their frequency of presentation, proposing the subject attempts to strengthen the unity of the group when the number of elements to be perceived is larger. This corresponds to the limits of subjective present reported by Dowling and Harwood [35]. Thus the tempo of presentation affects interpretation, rapid sequences caused listeners to place longer runs at the end of patterns, slow sequences caused listeners to begin the pattern with longer runs [41]. The number of chunks able to be held and processed reported by Dowling and Harwood agrees with the general estimate of 7 ± 2 item immediate memory capacity by Miller [35, 115].

Two interacting principles of grouping have been identified: the run principle and the gap principle [41]. The run principle proposes that the longest run of similar elements will begin a rhythmic pattern. Listeners tend to group sounds of the same intensity together. Runs of different intensity (or pitched) sounds will be organised so that the longest runs are placed at the beginning or end of the pattern, never in the middle. Objective accents are situated most spontaneously at the beginning of the pattern [41, pp. 159]. Lerdahl and Jackendoff's grouping preference rules suggest that similar elements (identical, alternating, ascending/descending progressions) will form a single group, with boundaries formed between dissimilar elements [84]. Graduated parameteric changes induce a *directed motion* or tension/relaxation towards some goal [15]. Povel and Essens' rules proposed that groups are composed

of elements close in time, that rests (silent durations) partition elements, and that slurred elements[7] will tend to be in the same group [136]. These last two proposals are examples of the gap principle.

The gap principle proposes the longest interval terminates the pattern. Vos identified three grouping preferences associated with timing accent and the gap principle: Tones separated by short intervals are perceptually grouped together, long intervals will preceed the first tone of a group (pauses in Fraisse's terms, described in section 2.2.1), and tones of a longer IOI are perceived accented, while short IOI tones as non-accented [187]. Where a rhythm demonstrates the gap and run principles with no contradictions, the pattern can be reproduced more easily than in contradicting examples [41, pp. 169].

Higher-level grouping structures are bounded by repetitions of, or discontinuties between, events [15]. The most predictable group sequences are the easiest to organise and respond to. Listeners' ready identification of repetition allows them to lock onto structurally simple patterns and thereby form groups. Woodrow has noted that regularly recurring intensity accents will produce a grouping where the intensified beat forms the downbeat, while a regular longer duration will group the longer duration beat as the end of a group [190, pp. 1233], another example of the gap principle. Fraisse has observed the impact that priming effects of the first perceived pattern will have on a stream of events. The initial pattern imposing its structure on interpretations of later patterns [41, pp. 162].

A single change in an established rhythm affects the entire percept of the grouping. Subjects will spontaneously *reorganize* isochronous tone sequences to conform to the run principle [55, pp. 407]. Studies of arhythmia indicate listeners will assimilate rhythms towards an economy of perception, deforming a complex pattern during reproduction to simplify it [41, pp. 167]. Where modes of accentuation are in contradiction, modifying the strength of the competing accents typically leads to reorganization such that the most salient accents form downbeats and conform to run and gap principles [55, pp. 390].

Pitch relationships create a multidimensional range of element variations which contribute to determining group boundaries, using enculturated tonal relationships (for example cadences) to create between-group and within-group (as measures of

[7]Slurs are a notational device to concatenate durations, producing long IOI elements that often cross metrical boundaries, Povel and Essens are thus proposing groups will extend past measure lengths.

similarity) relations. Competition occurs between these relationships to determine the final grouping, with the final arbiter being the individual's preference. Thus a reorganization of grouping of a musical sequence will occur not only with respect to the temporal principles described earlier, but with the melodic context. The first performed element can thus be heard either as a downbeat or anacrusis[8] depending on it's role within the melody which will only become clear after the performance of some portion of the sequence.

Parncutt has identified both serial and periodic grouping [130, pp. 411], the former describing the concatenation of motives, the latter to describe the form of grouping of periodic accentuation, that is, the meter.

2.2.4 Meter and Pulse

The Functional Role Of Meter

Meter is the occurance of regular subjective or objective accentuation, whereas isochronous beats only produce a sense of periodicity or *pulse*. A feeling of pulse is necessary for meter or rhythm, but requires a sense of differentiation of beats for these higher structures to arise. The regular alternation of perceptually weak and strong beats produces a sense of meter [127], which Dowling and Harwood have described as "the most basic level of rhythmic organisation" [35, pp. 185]. Clarke describes this as "[playing] a crucial role in determining the stability of detailed rhythmic groups" [15, pp. 221].[9] Meter mediates rhythmic interpretation, measuring time for the anticipation of forthcoming events and aiding organisation of equivalence classes of time points. These are similar to octave invariance of pitch, the first beat of the measure (downbeat) assumes a functional equivalence across time, aiding similarity and memory judgements [127].

Much of popular ensemble music, i.e. Jazz, Rock and Folk, is characterised by the meter being explicitly stated by performers in a rhythm section. In comparison, solo performances or many examples of Western classical art music typically only imply the meter, it being imagined in the mind of the listener without being heard.

[8] An initial upbeat, one or more notes preceding the first downbeat [84, 40].

[9] Meter is notated in Western music as a number of beats of a given duration within one periodic pattern of regular accents—the *measure*. For example, waltz meter, $\frac{3}{4}$, describes three beats each of a "crochet" or "quarter note" duration composing a measure group. A crochet can only be assigned a physical time duration according to a specified tempo rate, typically specified by the composer in beats per minute (BPM).

According to Sachs [153] and Meyer [112], in examples such as plain chant, "rhapsodic fantasias of many different cultures", or recitativo secco, pulse and meter do not exist at all [112, pp. 103].

However, as Frigyesi has noted [43, pp. 59], rhythm described as "free", "lacking clear beat" or her term, "flowing", is significant in non-western music, yet is rarely ever totally lacking any metric orientation. Musical forms such as the opening invocatory/introductory Hindustani *alap* [169] or Maghreb/Levant *taqsim* [153, 184] sections often create a feeling of "a latent, slow beat behind the rhythmic flexibility and a clear sense of periodicity" [43, pp. 66]. Likewise, such "free rhythm" sections may be characterised as disjoint periods of meter undergoing rubato between periods of single notes near the limit or beyond the subjective present.

In the music of India, the western concept of beat is matched by *matra*—the basic time unit—being the minimum time for one spoken syllable [169]. A given *tal*—a cycle, the meter or measure—is the grouping structure imposed on beats and is typically less than 20 beats in duration but may be theorised longer. The tal is subdivided, often in uneven and unequal concatenations. A tal indicates a metric grouping and also a collection of rhythmic patterns, specifying a hierarchy of rhythmic structures, of accent patterns, and of filled and empty beats. The occurance of frequent silences at beat times means the rhythm of the tal can be more implied than explicit and "aesthetic interest lies in the degree to which the musicians and their audience can continue to follow the tal accurately across passages of elaborate rhythmic improvisation in which it is only hinted at." [35, pp. 199]. Here a strongly conceived mental representations is required to provide musical meaning, in the face of little confirmation.

The assignment of beat strength, the grouping according to strength, and the effect of meter on perception of temporal structure must all be accounted for by a computational theory of rhythm.

Interaction Between Meter and Grouping

While the concept of auditory grouping has it's genesis in visual psychology anologies, meter is a concept evolved within music theory, having characteristic notation and typical forms, especially $\frac{\mathbf{4}}{\mathbf{4}}$. However, meters which extend beyond theoretical conventions do exist, such as conceptions of hypermeter proposed by Rothstein (noted by Parncutt [130]) as a *perceived* meter extending over time periods longer

than that notated within a measure. The notated meter of a rhythm may not necessarily be the meter the listener perceives.

Meter and grouping structures interact but may not always remain in phase, so that important notes within a group may fall on metrically weak beats, with the interplay between meter and grouping ultimately determining the rhythm perceived [55], [112, pp. 103]. Meter will effect the perception of the entire sequence [55, pp. 396], Povel demonstrated that playing 12 element sequences with different meters made identification of identical sequences impossible [136]. Handel demonstrated difficult rhythms not conforming to natural (i.e metrically typical) accenting were heard in terms of element grouping, rather than as a meter with a timing interval.

Clarke studied interactions between rhythm and meter by testing subjects' perception of a fixed rhythmic structure located at different starting points in two different meters. The pitch material of the stimulus was first comprised of tonal melodies, which were found to interact with the rhythm, and then with atonal melodic lines. The results were the same, indicating interaction between different musical properties [14, pp. 212]. He analysed variance between metrical context and note position (independent variables) and IOI as the dependent variable "indicating that the relative timing of notes is affected by metrical orientation" [14, pp. 213].

Tactus

Listeners tap to, and direct attention towards priviledged beats, and therefore a subset of metrical levels, in a sequence. *Tactus* is a renaissance term adopted by Lerdahl and Jackendoff [84] to describe the rate (tempo) and most salient hierarchial level at which the listener will tap their foot in accompaniment to a rhythm. Desain and Honing have described the tactus as the "level of metrical structure where beats pass at a moderate rate" [28, pp. 145],[26]. As described in section 2.2.6, on preferred timing rates the tactus will centre around a period of 600 msec. Parncutt found that requiring subjects to tap on their perceived downbeat of a number of percussive rhythmic patterns did not always result in subjects choosing the downbeat conforming to music theory. Subjects would often choose the theoretical meter, but phase shifted, such as choosing the fourth beat rather than the first in a measure of a swing rhythm [130, pp. 422]. These findings reflect that listeners are selecting the tactus as a frequency, from the possible periodic candidates which arise within the rhythm.

The Role of Repetition in the Formation of Meter

As noted in Section 2.2.3, repetition is a prominent aid to grouping and likewise repetition of sequences is a strong cue to the formation of meter. Longuet-Higgins and Lee [88], and Lee [82] have described the role of repetition of melody or rhythmic pattern, perhaps subject to pitch or temporal transposition, in the formation of meter. The nature of repetition in meter induction is an argument for a periodicity based representation. The role of repetition in determination of tactus of rhythms was demonstrated by Parncutt to not influence pulse salience [130].

2.2.5 Polyrhythms

In examples from African, Indian and Indonesian music [35], and examples of Jazz performances [40], rhythmic structure is commonly organised in layers involving patterns that do not subdivide from a single conductor's beat, rather they are composed of simultaneous lines in non-simple relationships. Handel has defined a polyrhythm as "the simultaneous presentation of two isochronous patterns that do not share a common denominator" [55, pp. 405]. African polyrhythmic organisation has been characterised [71, 156] as comprised of the following concepts:

- Additive rhythm.
- The concept of off-beat.
- Hemiola,[10] both horizontal and vertical.
- Cross-rhythms and inherent rhythms.
- The "standard pattern" or bell-line.
- A standard pattern in the form of timbral pattern.
- Motor pattern.
- Transaction—the specific polyphonic texture created by interlocking parts.

[10] Three beats in place of two or two beats in place of three [69].

The Bell-Line

Several cultural forms of African polyrhythmic drumming (for instance the Eve of Ghana) is characterised by the use of an asymmetrical "bell line" performed on a high pitched bell acting as a time-keeper [93, 35]. The use of this bell line and the difficulty of adequately notating or otherwise representing such polyrhythmic performances illustrates that the concepts of single meter, tempo and tactus are not universal. This clearly has implications for biologically inspired (e.g. connectionist) computational models in that these musical ideas are derived from a less simple underlying representation.

It is worthwhile to consider the counter-argument that a bell-line is a culturally unique derivative from underlying universal isochronous representations. Ethnomusicology in the field has determined that the bell-line is the intentional focus for performers in this musical form. It remains to be tested that indigenous performers will *always* be able to arrive at an isochronous re-interpretation of a recording of their performance, which would be the necessary proof of universality.

Other forms of African drumming (for example, the Kasena of North Ghana) do not use the bell line while exhibiting similar levels of complexity of pattern, this is argued by Koetting [35, 71] to suggest an abstract beat pattern exists implicitly in the structure of the performed rhythms. Clearly there is a common concept shared between the musicians which allows ensemble performance to produce the polyrhythm as a shared gestalt effect.

Analysing Polyrhythms by Fastest Pulse

The concept of fastest pulse is argued by Koetting [71] and Seifert [156] as far more suited (least dogmatic) to explaining this music rather than using Western concepts of rhythm, meter proportions and accents. Fastest pulse is considered to be the beat with the shortest duration in the music considered, but this is too coarse in consideration to our perceptual abilities illustrated in Section 2.4.2. Seifert proposes a pace-maker or clock as the central concept of rhythm perception and therefore implicitly declares a discrete structure to rhythm perception, in terms of the fastest oscillation of the pace-maker.

Fastest pulse allows "a better understanding, modelling and notational representation of African music as it sounds and has been recorded in the field" [156, pp. 180]. While this may describe the process of creation and can suit the needs of

analysis, it is also cautioned by Koetting [71] that fastest pulse does not seem to describe how African timing is perceived. What it does suggest is the upper bounds on a rhythmic sample rate, which Seifert concludes as 33.3Hz (30 msec interval) [156, pp. 175]. It also suggests the applicability of a time sampling approach to music which does not conform to a single meter.

The coincidence of important notes and strong beats is conjectured to aid in attentional efficiency, allowing plans to be made for the forthcoming beats which rationalises the limited attention which can be given to the task [64]. Handel has suggested from his research using isochronous, dissonant polyrhythms that one of the rhythmic lines, or every second element, is nearly always perceived as the meter tapped to [55, pp. 404]. He found that tempo (see Section 2.2.6), pitch and instrument/line intensity, and the particular combination of relative rates of each line would influence which line was selected as the tactus. Likewise Olk and Schneider found that Western listeners attending to synthesised typical forms of African polyrhythms used a sub-pattern or line which is repeated to determine the length of the pattern. They assumed the following signal features were important for non-African listeners:

- Detectability of periodicity inherent in a given sound sample.
- Length of the period defined in quavers.
- Complexity of patterns played.
- Number of distinct instruments/voices.
- Relative density of events per time unit.

These finding may have been biased from the culture of the subjects. Conversely, Koetting has argued underlying beat patterns in Kasena drumming are abstracted from any performed line [35, 71]. In the Olk and Schneider study, they found only those subjects with a background in African music attended to the traditional time keeping instrument, the bell-line.

2.2.6 Tempo

Tempo is a musical term describing the rate of presentation of beats. Effectively, the concept of tempo considered in Western music is the presentation rate of the

tactus. That is, Western notions of tempo are simply the rate of a structurally important rhythmic strata. In polyrhythmic music there does not exist the same sense of single tactus, although the rate of fastest pulse (where it can be derived), or average density of events per second can be used. However, regardless of culture, it is possible to understand tempo as a presentation rate of a perceptually significant level of beat. Examination of spontaneous, preferred and minimum perceivable tempos establishes absolute time constraints on rhythmic processes. These constraints provide measurable behaviours for automated rhythm perception.

Spontaneous and Preferred Tempo

Humans produce natural, spontaneous rhythms which are "a fundamental element of human motor activity" [41, pp. 152]. Spontaneous rhythms have their own ranges of tempo, which Fraisse asserts as ranging from 200 msec interval (5.0 taps/sec) to 1.4 sec (0.7 taps/sec), with the most representative value at 600 msec (1.1 taps/sec) ([41], see also [55, 35]). Spontaneous performed tempo can be distinguished from a listener's preferred tempo, but both have similar rates of approximately 600 msec. Maximum accuracy in timing judgements, and reproduction with minimum over or under-estimation also occurs at around 500–600 msec IOI [55, pp. 385]. Spontaneous rhythms of identical twins are very similar, whereas fraternal twins differ by degrees similar to that of unrelated subjects, suggesting a biological basis for spontaneous rhythm [41, pp. 153]. However, the variation displayed among listeners is too dynamic to imply a single rhythmic constant generated by a simple biological clock.

Relationship of Preferred Pace With Tactus

Parncutt [130] found that the selection of tactus by a group of subjects for a group of thirty-six tempo and rhythmic pattern combinations varied widely. The variation of the tactus selected increased with rhythmic pattern complexity. He found a distribution of preferences with a mean of 710 msec IOI and a standard deviation which corresponded to an interval of 420–1190 msec. These intervals correspond well with preferred tempo from other literature as compared in section 2.4.2.

Subjects were measured tapping to isochronous pulses at varying tempos. It was found that preferred pulses will gravitate towards a moderate tempo. Noting the tempo dependency of tactus, Parncutt has formulated an "existence region of pulse

sensation" model, defining a range of periods within which isochronous sequences are perceived as musical. Parncutt goes further to define that the existence region is a Gaussian weighting function centered over the moderate tempo rate of 600 msec IOI and is symmetric with respect to the logarithm of the pulse period between 400 to 900 msec [129]. Therefore the closer to a moderate tempo the stimulus is, the more salient the pulse sensation.

Structure and Tempo Interaction

Clarke's investigations of tempo in performance have shown an interrelationship between structure and tempo [14]. Music tended to be grouped into fewer units at higher tempi, with slower tempo aiding segmentation at points of structural boundaries due to discontinuities in pitch and duration and structural parallelisms. Clarke argues this tempo dependent grouping as due to the limitation of the subjective present. This limit enforces segmentation by the performer and thereby causes the subdivision of larger groups at points in accordance with structural properties of the music.

Predictability of a stimulus influences the perceived pace, the more varied it is, the shorter it is perceived [35], this is also known as the *filled-interval effect* [67]. Studies of time discrimination and reproduction of isolated intervals determined a just noticable difference (JND) ranged between 5%–10% [190], however temporal judgements for tapping to a regular beat (Povel [135]) determined JNDs of 2–3% at 600 msec interval, with continuation close to this capability after the beat was stopped. This was independent of musical training.

Povel has suggested a steady beat pattern is the cognitive framework used by a listener to structure musical time and produce precise rhythmic patterns, with the metrical structure functioning as a schemata. This would appear biologically determined: "The dual structure of underlying beat and superimposed rhythm is fundamental to the cognitive organization of music from very early ages." [35, pp. 186].

Povel [135] found categorical beat assimilation towards a 2:1 ratio of IOIs when listeners attempted to tap to a variety of stimulus ratios between 4:1 and 5:4, with more complex rhythms being poorly imitated. He proposed rhythmic encoding occurs in two stages: First a search occurs for a regular "beat framework" within the preferred pace. This search is lower bounded by the IOI of 1.5 sec or 40 BPM

tempo. Subsequently the rhythm is subdivided either equally or into long/short durations of 2:1 ratio.

Dowling and Harwood [35] and Handel [55, pp. 404],[56] have reported that in attempting to tap to complex polyrhythms, listeners typically tapped at the cross-rhythm (combination of the rates) at slower tempos (3 secs/measure), and selected one of the component rhythms at higher tempos (1.6–1.2 sec/measure). Listeners rarely followed component beat patterns with IOIs greater than 800 msecs apart. The prefered pace appears to influence the subjective rhythm, causing listeners to shift attention to component rhythmic lines with shorter intervals when they encounter a slower tempo. The result is to attempt to remain close to the preferred (600 msec) rate. As tempos increased, listeners chose either the cross-rhythm or a component pattern which was closest to the preferred pace of 600 msec IOI.

2.2.7 Expressive Timing and Rubato

Expressive timing is an evolved performance practice known by the Italian term *rubato*—literally "robbed time". It is also termed 'micro-tempo' in computer music and 'local tempo' in music psychology. It is effectively a local tempo change from event to event as the piece progresses. While it has an emotive character, especially in the style of Romantic music where it has its extremes, expressive transformations on a canonical rhythm derived from a score are intended to highlight the grouping structure [15, 26, 55], [28, pp. 145]. The expressive timing variations alter the structural relationships between beats at different levels, within the bounds of the abstract structure, which the expression (intentional deviations from metricality) seeks to accentuate.

The expressive deformations from metricality produce distinct effects, which depends on a listener's enculturated experience. Bengtsson and Gabrielsson [4] showed that time intervals needed to be uneven in order for listeners to perceive them as musically "correct". Using listener preferences for alternative timing deviations of the same folk tune (the test data), equal preferrence was given to ratios of 1.7:1.0 through to 2.0:1.0 against other ratios outside these bounds. Within these ratio bounds, the rhythmical motion of the piece is affected by the particular ratio chosen. This would seem to be an interplay between categorical perception tendencies to simplify, and a tendency to accelerate or ritard the rhythm as a rubato.

Forms of expressive timing can be identified as ritard/accelerand behaviour,

phrase final lengthening or more generally structural/metrical contexts. Traditionally these have been modelled as tempo curves. These concepts are now considered.

Ritards

Common musical theory has adopted a number of Italian terms to describe forms of expressive timing, among them *accelerando* and *rallentando* as standard structural elements describing common monotonically increasing or decreasing rate of beats. The term *ritardando* describes the characteristic slowing of a piece, particularly at the concluding measures.

Kronman and Sundberg have used regression analysis to model the final ritardando in musical performances as explicitly " ... motion under constant negative acceleration" [77, pp. 1941] (i.e linear tempo decrease). They found this produced a reasonable approximation for the example data taken from performances of Bach preludes ("motor music" where all notes were equal value, i.e quavers [174]). Longuet-Higgins and Lisle [89] also found linear tempo changes within a single accelerando/ritardando produced the most natural sounding rubato.

This approach of relating ritard to physical motion is reflected in the musical terminology above and is popular in research [104, 38], but has been recently critiqued by Desain and Honing [32]. While they complement musical models based on human body functions and constraints, they caution against physical motion as a model as it does not reveal the underlying mechanisms. "A good approximation is not neccessarily a good explanation." [32, pp. 459].

They demonstrate final ritards are dependent on the *global* tempo, and argue the dependency of ritards on the structure of the piece. For this they argue that passages of many isochronous notes allow for deep rubato, while ritards of few notes with more elaborate rhythmic structure will have less radical deviations from metricality. This is necessary not to break the rhythmic categories. This intricate dependency is not reflected by physical motion (linear deceleration) models.

It is telling that given significant research on this issue, there is still considerable debate regarding the ritard behaviour. This suggests that current analysis models are incomplete, and that further development of analysis methods are required to build a better picture of such behaviour. A method of describing rates of events changing over time is required without pre-suppositions of frequency characteristics.

Tempo Curves

The ritard is one example of expressive timing that has been attempted to be represented by a *tempo curve*. Tempo curves are profiles of beat-by-beat time deviation from a canonical metrical grid. Local tempo is represented computationally as an event to event ratio of score time interval to performance time interval

$$L = G \times \frac{S}{P}, \tag{1}$$

where L is the local tempo at the time of each beat, G is the global tempo, S is the score time interval and P is the performance time interval [28]. Thus a note performed longer than in the score will be a value below the global tempo, and a note shorter, a value above. In computational representations of the tempo curve, points are connected by straight line-segments, or alternatively by spline interpolation (to create smooth transitions between tempos).

Desain and Honing's entertaining "Tempo Curves Considered Harmful" series [28, 26, 27] made important recommendations regarding the limitations of tempo curves. They argue a tempo curve conveys a false impression that time can be abstracted to become independent from the events that mark it. This fallacy is demonstrated by attempting to apply the tempo curve from one piece to a related piece (such as a variation on a theme). Representations of expressive timing using tempo curves miss the essential link with the underlying musical structure—that expression can only function with respect to a structural base. Desain and Honing make the point that systems using tempo curves have become ubiquitous, but have propagated the erroneous assumptions that they make musical sense, that they are useful computational rhythm representations, and that they are a mental representation held by the listener or performer.

Tempo Dependency Of Expressive Timing

In the same manner as their critique of ritard models (Section 2.2.7), Desain and Honing demonstrate the problem of tempo curves being unable to be applied from one performance to another performance of the same piece recorded at a faster global tempo.

In a further study [30], Desain and Honing produced experimental evidence contradicting an earlier study by Repp [143] which argued that expressive timing scaled

proportionally with a piece played at various tempos [30]. They showed that for a piano piece performed at a range of tempi chosen by the performers within their skills, a significant interaction between tempo and IOI existed. This was achieved by measuring the IOI of grace-notes, and separately, by the correlation of expressive timing profiles across tempi. The conclusion was drawn that expressive timing was non-proportional to tempo, contradicting the earlier hypothesis of expressive timing being relationally invariant to tempo. Desain and Honing have not yet proposed a model of non-linear relation between tempo and expressive timing.

Phrase Final Lengthening—Expression from Structure

Measures at the ends of groups (phrases) also tend to be extended by a short ritard and are termed *phrase final lengthening*. Clarke analysed the variance between IOI as the dependent variable and metrical context and note position as the independent variables. He found "significant interactions between metrical context and note position, indicating that the relative timing of notes is affected by metrical orientation" [14, pp. 213]. He identifies three principles (which may interfere or promote each other) causing this interaction:

- Metrical strength accounts for the most significant variance in note length. The stronger the metrical position (its position in the metrical hierarchy), the greater the lengthening of the IOI, a weaker metrical position results in a shorter IOI. The relationship to metrical position may be direct (agogic accents) and indirect (lengthening of notes completing groups).
- Less significantly, notes completing phrases or groups (at varying structural levels) are lengthened.
- Least significantly, notes immediately prior to a structurally important note are lengthened (delaying the onset of the important note).

Palmer studied the effect of interpretation of phrasing on performance of romantic period piano music by comparing timing of expressive performances with the performer's intended grouping structure [124, 123]. Like Clarke, she also found slowing towards the end of each phrase with a characteristic over-corrected return to metricality. In comparing performers who were asked to play "expressively" and "mechanically" (lacking expression), rubato was found to no longer match phrase

structure boundaries on the "mechanical" version (which still contained some performer variation from strict isochrony). By comparing the expressive timing of a performer with the intended phrasing of others, Palmer found only a match between a performer's phrase final lengthening and that same performers intended phrasing, as previously marked by them in the score.

Parncutt's expressive timing model [130] adopts a proportional increase in the degree of slowing in the temporal vicinity of an accent with the strength of that accent. He defines structurally important beats as "events preceding relatively long IOIs, on rhythmically strong beats, at the start of phrases, at harmonic dissonances, or at phrase or structural boundaries" [130, pp. 447].

Todd [97, 98, 109, 99] found that the importance of a phrase corresponded roughly to the degree of lengthing of phrase endings. He also proposed relationships between intensity and tempo that aid in phrase segmentation [100, 103], however there are several musical examples which run counter to this positive association.[11] Bengtsson and Gabrielsson found for the music they examined (a short waltz), that lengthening of phrases occurs where endings of phrases at different structural levels coincide [4]. Timing deviations occur at each structural level and the timing of the performance is the combination of the timing of each level.

Most models of expressive timing have linked the generation of rubato to a single structural entity (immediate event intervals [173], metrical units [16], and phrase final lengthening). Desain and Honing suggest that future expressive timing models should be linking rubato to several structural entities, both surface features and deeper structural entities [28].

Summary of Expressive Timing

Two cases of expressive timing are apparent: *shaping*, occuring over the relatively long duration of groups or measures, and *localised deviations* such as agogic accents or grace notes. While these terms are old [141] and may seem inexact, they can be understood as follows.

Shaping of phrases occurs over more than one beat and can be defined as acceleration based deviations from a canonical beat. These changes of rate (whatever their linear or non-linear character), are such that a grouped phrase of beats undergo

[11] For example, the introductory motif of Heitor Villa-Lobos' Prelude Number 3 is performed with a simultaneous ritard and crescendo. This is opposite to the "faster and louder" association observed by Gabrielsson [45] and adopted by Todd [100].

some modulation of the underlying beat frequency, pushing and pulling beats away from metricality. This modulation of frequency is most apparent when the shaping occurs over motor music, a canonically isochronous series of note events. This also defines the modulation rate to be significantly lower (at least an octave lower) than the beat rate.

Localised deviation or agogic jitter can be considered to be deviations from metrical location which are short term, i.e highly localised, dragging or leading a single beat but not the surrounding beats. This could occur with respect to lower frequency modulation, for instance, all but the subject beat are correctly ritarded. From the perspective of rhythmic frequency this beat deviation occurs at a higher frequency than the canonical beat rate. The reoccurance of the agogic accent at a regular number of beats would create a rhythm frequency at a rate lower than the canonical rate. This can be considered to be grouping using agogic accents.

The issue of culture arises in the consideration of expressive timing, in this case the difference between expressive timing within the tradition of Western music [158], in comparision to Jazz (and less prominently popular music) where the meter is strongly implied by the rhythm section. The notion of *swing* has become a (more-or-less) theoretical concept [139, 17]. However, it seems to demonstrate the same general forms of expressive timing, shaping [2], and localised deviations [139, 6, 7, 8] as Western classical music. On reflection, this appears plausible as jazz has a clear heritage to Western common practice music as much as to earlier Afro-American music.

2.3 Rhythmic Models

2.3.1 Rhythmic Strata

There have been a wide range of representational approaches to musical time, ranging in their degree of association with music theory and performance practice, and in their conformance to results of perceptual research. Recent proposals have explicitly modelled the concept of a hierarchy of temporal levels. Yeston has argued for the conception and representation of rhythm as a hierarchy of *strata* [192], each of increasing timespans, with proportionally decreasing pulse rates. Yeston considers his model to be an analysis of musical structure, rather than as a direct perceptual model. Meter is conceived by him to arise from accents created by the interaction

between hierarchical strata levels [192, pp. 66]. Other researchers have also used terms such as a level of motion, hierarchical clocks [136], a time level [15], level of pulsation [127], rhythmic level [147] and in a wider sense, pulse sensation [130], to describe all levels of time evoked when listening to a musical rhythm, either including or excluding expressive timing.

While hierarchical stratification is a widely held view, it does not seem that all levels are equally salient. Certainly the tactus, typically the rate of beat that corresponds to the notated meter, holds a significant perceptual position. Clarke identifies distinct levels of musical time, demonstrating the levels differ significantly between event relationships within a group, and between groups [15]. He characterises three categorical levels of temporal structure (see Figure 3): a "low" level of expressive timing; an intermediate level of "canonical" relationships between single notes and phrases upto 3 or 4 measures, including meter; and a "high" level of *form*, describing long term group structures of 4 to 8 measures, considered by Clarke to be the functional limit of listeners ability to construct high level temporal frameworks.

Levels range respectively from "lower" strata which are more ornamental, less structural, and of shorter time spans, to the "higher" abstract stratified levels, of longer time spans, having fewer, more significant events, and influencing boundaries between groups of events, metrical structures, durational proportions, and directed motion within groups. Levels are argued to not always be perceptually distinct: "Continual cross-connection and transformation between levels blur their boundaries, and generate a structural network that contains a complex mixture of hierarchical and associative relationships." [15, pp. 212].

2.3.2 Hierarchical Theories of Meter

The common Western musical practice of considering meter as a hierarchy of binary and ternary beats [127] has been proposed by Steedman as a "principle of consistency" in listening [171]. As subjects can only judge two or three tone durations accurately, Handel conjectures that there may be perceptual limitations which explain the reason for the predominance of duple meter in Western music [55, pp. 403]. London has also noted a binary bias to Western rhythm, particularly in dance music, beginning with the baroque era [85]. This is hypothesised to be the result of movement constraints of normal human bipedal motion and is used to explain

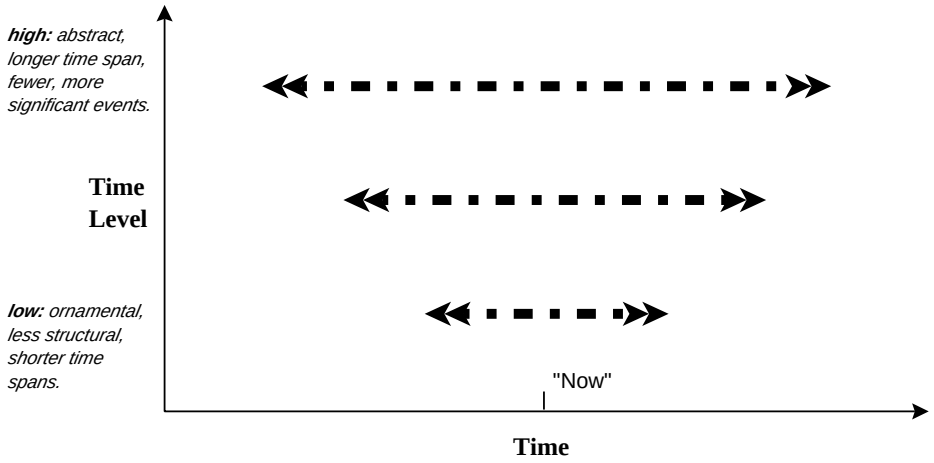

Figure 3: A time level/time schematic diagram of Clarke's proposal of categories of temporal levels in musical time.

the scarecity of triple meter in Western music. However, the argument for a biological motor basis seems less plausible considering asymmetrical and polyrhythmic non-western music.

Lerdahl and Jackendoff have proposed their "Generative Theory of Tonal Music" (GTTM), applying a theory of generative grammars as a descriptive mechanism for musical relations [84]. They propose both a metrical hierarchy and a grouping hierarchy. They propose the meter hierarchy as being responsible for the assignment of importance of beats in short time spans. Over longer time spans, the grouping hierarchy is responsible for the organisation into phrases, themes and sections. The perception of meter and of grouping occurs simultaneously, and it is conjectured by Handel that separation of the two percepts is impossible [55].

The grouping hierarchy describes the organisation of beats into levels of phrases (groups) of increasingly longer timespans. According to GTTM, "meter preference", "grouping preference", and "well-formedness" rules are required to resolve the ambiguity of the meter and grouping hierarchies from many possible interpretations of a given series of notes and intervals. The preference rules attempt to predict the perception of one of the well-formed meter hierarchies and one of the well-formed grouping hierarchies.

The well-formed meter hierarchies are musically appropriate ones, i.e. meter has evenly spaced beats at each level and synchronises with beats at a lower level.

For Western music, beats are equal groups of two or three timespans of lower level beats. The coincidence of beats within meters between different levels produces the perceived strength of a beat. Well-formed grouping hierarchies enforce groups to be composed of adjacent elements, typically with an element existing in only one group at a single level [55, pp. 393]. Lerdahl and Jackendoff's model is informally descriptive, rather than computationally implementable and operates, like Yeston's model, on a canonical musical structure derived from musical notation of a piece.

Enculturated knowledge of meter can be assumed to play a role in its perception. Sufficently abstract "context-free" knowledge of the temporal functions at work in Western music provides a mental framework (or *schema*) to interact with the "context-dependent" knowledge of the unfolding music to produce the conception of meter [127, pp. 730]. This is reflected in Longuet-Higgins' model [87] by an inertia for changing meter in the face of conflicting evidence. Longuet-Higgins and Lee [88] used only relative duration cues and the position of the durations within the bar in their model of beat induction.

Palmer and Krumhansl [127] have demonstrated that the frequency of occurance of events at a given position within the measure (that is, within a *metrical context*), sampled from a range of classical pieces, matched listener choice of goodness-of-fit. The experiment consisted of presenting a probe tone following presentation of a series of "context" beats. The context beats were imagined by the subject to be the downbeat of a meter. The probe tone would then fall on one of the semiquaver locations within the imagined measure. Using their findings,[12] Parncutt [131] has constructed templates of meter and pulse by proposing weightings for each semi-quaver location within a measure. The pulse template is argued to enable the recognition of events as periodic, from a common source, and thus the period. Additionally it is argued to be a universal construct, whereas the meter templates are argued to be culture specific, and constructed for Western music. Matching of a stimulus to a template is characterised in two forms: "jumping" and "sliding"; describing initial recognition and ongoing matching of stimulus to a template respectively. Sliding is said to occur when tracking a rubato, in the case of a metrical template.

Parncutt proposes a model [130] for determination of perceptual saliences of rhythmic strata. Occurance of *consonant* pulse sensations two or three times slower or faster than a target pulse sensation is proposed to enhance perception of the

[12]Noting discrepencies of their findings to the position of the downbeat of $\frac{2}{4}$ from music theory.

target. The salience of a meter is argued to be proportional to the sum of consonant pulse sensations. The meter chosen as the tactus is therefore argued to be the one with the highest salience. Parncutt hypothesises a lower limit of two consecutive events as the point at which a pulse sensation will no longer be salient, unless supported by consonant pulse sensation of higher or lower frequency [130, pp. 445]. This constraint is not included in his model, however. In addition, this may not conform to an assumption he makes [130, pp. 434] that two events are sufficient to give rise to a pulse sensation.

The perceptual salience of each pulse level was derived by experimental evidence from subjects tapping along to experimental rhythms and measuring period and phase of each tap rate [130]. Metrical accent is represented in Parncutt's model as arising from a two stage process, with metrical accents determined from the pulse saliences, which, in turn, are determined from durational (timing) accent [130, 131]. Metrical accent is therefore computed with a linear summation of all pulse period sensations at each time point. According to Parncutt, there is a direct correspondance between metrical accents and temporal categories.

2.3.3 Models of Grouping and Metrical Structure

More than one position exists regarding the relationship between grouping and meter. Questions concern the evolution of grouping during listening, and the role of accents—whether they define meter, or whether meter gives time points an identity independent of objective accentuation [127]. The traditional perspective (Meyer's, for example [112]) is that meter and grouping are intrinsically linked and both arise from the pattern of accents, with groups formed by "affiliation" of unaccented events to each accented event [127, pp. 729]. The grouping may be non-regular while still being within a given meter, and may vary in clarity at each time point. According to Meyer [112, pp. 103 and 147], a meter can be created without also creating a sense of rhythm due to ambiguous grouping. A listener's impression may only be of a sequence of repeating strong and weak beats.

Conversely the GTTM viewed meter as the organisation of time-points resulting from accent, whereas grouping was seen as a separate organisation of events without reference to accent [84]. Lerdahl and Jackendoff proposed a time span reduction (TSR) as a generalized model of interaction of grouping structure and metrical structure, representing the rhythmical structure of the piece. One TSR

well-formedness rule is that there is a most important event in a given time span, chosen from important events in shorter time spans covered by the parent time span. This forms a strict hierarchy of structural importance. This structure can form top-down expectancies of metrical accent that allows for syncopation when objective accentuation contradicts the expectation.

Longuet-Higgins and Lee [88] found the correct metrical interpretation could be achieved by weighting each event according only to its relative duration cues and the position of the durations within the bar. This explicitly distinguishes accentuation from the cues that are commonly thought to be accents, that is, dynamics. Palmer has noted that "duration and intensity cues are influenced in both composition and performance by many factors in addition to meter ... often these cues are ambiguous, interactive, or simply absent: yet the listener still determines the meter correctly. Therefore it is unlikely that sensory cues alone determine the meter" [127, pp. 730]. In the opposite camp, meter is not wholly responsible for accentuation. Palmer and Krumhansl [126], Clarke [14], Jones [66], Jones and Boltz [67] describe conditions producing accent perception, regardless of prior establishment of a meter.

2.3.4 Models of Expressive Timing

There have been many approaches to modelling expressive timing, these are well summarised by Parncutt [130, pp. 447]. Two categories of models can be distinguished: Generative models attempt the generation of expressive timing from a structural description while interpretative models attempt the interpretation of a performed expressive gesture to produce a structural description. Examples from both categories which reflect a multiple time scale model are detailed here.

Generative models attempt to determine deviation of events from a canonical description of the intended structure of a piece, typically some annotated encoding of the musical score, or a more abstract structural description. Clynes composers pulse [16] model determines all expressive timing from metrical units, effectively a two level (measure and beat) analysis. As investigated by Repp [140], meter does not completely describe expressive timing, especially in romantic music where phrase structure is communicated with deep rubato and meter is highly variable in local tempo.

Todd's model uses tempo curves linked to phrase structure [100, 98]. His approach is to look at the local maxima of the tempo curve and relate the strengths of

the peaks directly to the structural boundary strength [97]. From the tempo curves he determines an acceleration/deceleration profile over the length of a phrase. The profiles occur over a nested hierarchy of phrases, combining to produce the final tempo of the entire piece. Todd considers expressive timing in terms of kinematics of motion (see Sections 2.2.7 and 2.2.7) in a 2-D space, with metrical position representing space in terms of beats (discretised time when with respect to a tempo) [104]. He characterises timing as a series of connected trajectories or timing segments. Changes in sign of the acceleration value is used to indicate segment boundaries. The piece then consists of connected linear tempo segments of typically 2 to 7 metrical units.

As an example of an interpretive model, Desain and Vos used partial autocorrelation [24] as an analysis tool of POCO [59], that focused on expressive timing. A similar autocorrelation was also used by Brown [12]. Autocorrelation was aimed at identifying structural levels, rather than microstructural deviations. Deviations from metricality were assumed to stem "from a multiplicative combination of tempo factors at several structural levels and the exact metrical note durations in a the [sic] score" [24, pp. 357]. Assuming the musical structure to be analysed is homogeneous, Desain and Vos searched for periodicities in the tempo curve,[13] with the periods being interpreted as the lengths of the structural components. Limitations with the autocorrelation approach identified by Desain and Vos include that no phase information of the rhythm is retained.

Addressing the tempo curve problem, Honing and Desain's system "Expresso" [60] interprets expressive timing from a performance with respect to human authored structural descriptions and a quantized note list [29]. This last requirement does imply the quantization must correctly deduce the intended score, which shifts the burden of accuracy in expression interpretation onto the quantizer. Assuming a robust quantizer [25], it is possible to create different levels of the structural annotations. Transformations such as tempo change or adding/removing notes can then be made with respect to higher level (longer term, more abstract) structures. These transformations can then selectively respect the invariability or modifiability of other structural levels.

Tanguiane [177] defines rhythmic perception as separated into high and low-level configurations, consisting of tempo curve, and correlated rhythmic patterns of time

[13] See Desain and Honing's reply [33, pp. 113] to Smoliar's criticism of autocorrelation of expressive timing [167].

events, respectively. High-level configurations concern time relationships between rhythmic patterns. He assumes low-level repetitious patterns form recognizable references that are used to track tempo by the listener.

Widmer's research in the learning of expressive timing used the score of melodies and the tempo curves of individual performances [188]. Several increasing longer term structural levels are determined from the score by symbolic matching. Surface patterns, a metrical structure and a grouping structure of a form similar to GTTM [84] are computed. Expressive patterns of dynamics and rubato (from the tempo curves) are then learnt using an "instance-based" algorithm, generating production rules. These expressive patterns are of prototype forms such as ascending, descending, and ascending-then-descending. The production rules are then used to synthesize appropriate dynamics and rubato behaviours at points in a new musical score when the production rules are triggered.

Parncutt's pulse salience model [130] examines a whole rhythmic sequence and is currently restricted to cyclically repeating rhythms. A pattern matching routine models the perception of pulse and produces a set of concurrent pulse saliences. The salience being defined as the probability of the pulse being a tactus. Phenomenal (i.e. objective) accents (including duration accents) contribute to pulse salience using a non-linear "saturation" function. Measures of pulse salience are weighted by absolute tempo ("existence region of pulse sensation"), with 600 msec interval as the maximally responsive "moderate pulse period". In this model expressive timing is suggested to always follow the behaviour of slowing local tempo close to metrical and structural accents.

Todd's auditory-motor basis of rhythm perception [104] is composed of:

- An auditory periphery, simulated with a cochlear model and hair-cell array simulation.

- The "rhythmogram" time-domain process [103], which involves convolving a bank of low-pass one-dimensional, causal filters with the auditory nerve response and then finding peaks in the first derivative of the rhythmogram response. This "carries out a temporal segmentation of the activity in the auditory nerve" [104, pp. 1946].

  The rhythmogram is performed for a number of spatial scales and the resulting multiple channel peaks are recombined to allow searching for coincident peaks

across channels. Extrema which are "temporally coincident" across channels are assigned a single signifier - they are assumed to be a single phenomenon. The result is a hierarchical structure which reflects the effect of phrase shaping and grouping.

- Using bandpass rather than the lowpass filters, periodicity analysis is performed. This produces an association between each event and a number of harmonics. Todd placed the conditions on periodicity analysis that an absolute change of tempo is piece-wise continuous and that the rate of absolute change of tempo is less than one octave per onset. The ratios of the harmonics will be invariant.

- A sensory motor feedback filter (consisting of two peak bandwidths, for foot-tapping, capturing periodicity, and body-sway, capturing gesture) used to promote the metrical harmonics and therefore select the tactus. In Todd's model, these are achieved by boosting the response from the filters representing 1.7Hz and 0.2Hz respectively. The harmonic closest to the foot-tapper resonance (1.7Hz) "will be the one favoured for the tactus" [104, pp. 1947].

2.3.5 Connectionist Oscillator Models

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 [80, 114, 122, 47, 183].

BeatNet [114] consists of idealised low-frequency oscillators of different beat periods which align their periodic impulse ("ticks") with event onsets. Large [80] identifies its inability to deal with timing variation, due to the use of idealised oscillatory units. However, it may be that using sufficient inharmonic oscillators of short-term activation would enable recognition of variable timing.

Large proposes and reports [80] a dynamic system that synchronises a bank of neural oscillators in a range of oscillator ratios to the performed rhythm's IOIs. The stimulus of a "basic oscillatory unit" is a series of normalised impulses (dirac functions) for the onset of each note, without an intensity encoding, derived from MIDI or from acoustic signal amplitude. The output of the unit is non-zero for only a small portion of the period, which creates a region of temporal expectancy—a time period during which a stimulus pulse is expected. The unit adjusts its phase

and period in response to stimulus to minimise an error function measuring the difference between the expected stimulus and the occuring stimulus. Updates to phase and period are proportional to the partial derivative of the error function. This functions to lock its output pulses to the period of the stimulus. The phase adjustment is scaled so it is independent of the oscillator period.

The frequency (period) of the oscillator also tracks using a gradient descent (error minimisation) approach, with limits placed on the period. The unit synchronises its output based on a periodic train of input pulses; to operate in a network, it will encounter stimulus frequencies outside its response region. Frequency tracking is demonstrated to act both as a memory for a periodic stimulus and as stabiliser against timing deviations. The resonance theories described by regime diagrams indicate the allowable beat-periods, according to their sensitivity to timing variation and therefore (it is argued) well-formedness rules of metrical structures [84] can be expressed. Perception of polyrhythms is limited by the number of mode-locks that a system can contain.

Large's oscillator bank is claimed to model the perception of metrical structure, that is, the inducement of the beat (tactus). The network of neural oscillators is intended to be interconnected in a self-organising behaviour. These entrain simultaneously to the periodic components of the rhythm at different timescales. Metrical units are said to manifest themselves as the alignment (or the relative phase) between adjacent levels of beats.

While phase locking is reported in the neurobiological literature, Large's use of oscillatory units is higher than the neuronal level, rather at the level of meter perception, using ratios more complex than 1:1. Large suggests the behaviour of the units mimic the emergent behaviour of a wide range of possible brain structures [80, pp. 202].

Using non-linearly coupled oscillators (one oscillator being the rhythmic performance input, the other the metrical clock), the coupled oscillators have an inertia controlling their stability (coupling strength) at chosen ratios (the dressed winding number) from their "bare winding number".

Due to the delay in reestablishing a phase lock, musical beats can't be adequately modelled using phase locking entrainment alone. Frequency tracking is used to alter the period of the oscillator rather than simply the phase. On removal of the driving signal, the frequency tracking oscillator will retain its last driven period until the stimulus reappears. Entrainment in Large's model consists of perturbing both phase

and frequency of the oscillators only at certain points in the rhythmic pattern. This is said to effectively isolate a single periodic component in the incoming rhythm [80, pp. 190].

Experiments performed by Large are by exposing a small system (six oscillators) of unconnected oscillators to MIDI piano performances. Each oscillator is frequency limited to an octave and a third, there are two octaves of response, with a minimum period of 600ms, maximum period of 2560 ms, distributed as:

$$p_{i+1} = 2^{\sqrt{3}p_i}$$

where p_i is the period of the oscillator.

Large assesses the performance of the oscillators to isolate a periodic component of the rhythm without any phenomenal accent information. The oscillators correspond to quarter and half note timings, however it is not explained why eighth and sixteenth note timings are not stabilised, given there are a large number of dotted quarter, eighth and sixteenth notes, one half note and 9 quarter notes in the reported example. Perhaps this is due to the variability from performance of the timing of the shorter duration notes.

An oscillator unit's preference for a periodicity is an interrelationship between:

- The unit's point of maximum expectancy.
- Spacing of event onsets around the expectancy point.
- Width of the expectancy region.
- Absolute amount of adjustment to phase and period in response to each onset.

Wide response regions produce a preference for simple ratios and narrow temporal response regions allow more complex ratios [80, pp. 203]. It may well be that this prevents stability at shorter timescales in the example reported.

Typical versions of such models are not actually interlocked between hierarchies [80, pp. 198]. This suggests that independent stratified layers of rhythmic times produced by a time-frequency analysis will equally reveal the signal on which the oscillators are adapting to and their behaviour. While the oscillator entrainment approach can reflect beat tracking, it is not clear if the oscillator dynamics can show longer term structural entities.

A problem with neural oscillator models is their assumptions of phenomenal accent to display "appropriate" behaviour. Large concludes: "In summary, modelling the perception of metrical structure is difficult, in large measure because of problems arising from timing variabilty in musical performance ... Entrainment models must have the ability to "pick" component periodicities out of a complex rhythmic pattern in spite of missing, ambiguous, or misleading phenomenal accent information" [80, pp. 186]. The inherent problem is stated—the representation of rhythm as periodicities—but the timing variability is characterised as noise rather than as explicit, non-verbal communicated knowledge.

Desain's decomposable rhythm perspective has similarities to a wavelet approach [23]. 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 time-window within which beats are identified.

2.4 Summary of Findings

2.4.1 Adopting a Multiresolution Approach

All these models implicitly or explicitly represent rhythm by decomposing time into levels, strata, or saliences having temporal periods which differ between each level. These periods have an intuitive hierarchy from arranging them ordered by time extent. Beat duration times (IOI) inter-relate by low value integer ratios. This matches binary and ternary decomposition of the symbolic rhythmic forms of music theory.

These time periods may be seen as the reciprocal of frequencies of events. The ascending/descending arrangement can then form a rhythmic spectrogram, in a similar manner to sonogram representations of audio signals [119, 34], but at rhythmic frequencies (summarised in Section 2.4.2). This suggests a computational analysis approach of exhaustively representing the periodicities which are created by the temporal relationships between beats over multiple timescales. This includes both metrical and agogic times on a continuous scale, the later will form inharmonic ratios to a rhythmic "fundamental frequency".

With the exception of Todd's primal sketch approach [103], earlier models have worked exclusively in the time domain. Autocorrelation approaches identify periodicities but impose an overly restrictive strict periodicity. This is clearly unsuited for rubato and agogic accentuation representation, as these effects are inharmonic to meter and grouping structures.

Oscillator models also imply identification of periodic behaviour, while allowing a (albeit non-obvious) degree of deviation from strict periodicity. Oscillator models attempt structural decomposition and entrainment simultaneously. By considering rhythmic frequency directly allows for an explication of each of these processes. This allows for transient periodicities marked by phenomenal accents to be matched and to compete, to disambiguate sections which are not marked by steady accenting. It is therefore fruitful to investigate frequency analysis methods which enable explicit representation of frequency change over time.

2.4.2 The Rhythmic "Periodic" Table

The levels of musical time discussed in this chapter may be arranged by their associated IOI times reported in the literature. Inspection of the times (arranged in Table 2) reveals a relatively evenly spread distribution of perceptual effects across a range of interval times.[14] While this table incorrectly presents the impression of a common accuracy of interval timing from a wide range of literature, it does function to provide converging evidence of the states of temporal perception.

With the development of common musical practice, standard terms have been used to describe different dance styles. From these conceptions of dances, an appropriate rate is devised and over time codified to a metronome marking (M.M.). However, there are limits to which there can be direct relationship between musical measures of tempo and psychological findings. Pedagogical approaches to the knowledge of tempo, often aim to impress on students that a dance term (i.e. *Andante*) is as much a description of a dance *style*, with connotations which will effect articulation, structure, trills and degree of ornamentation, as a description of a strict tempo rate of musical event presentation.

[14] A hearty thanks to Robyn Owens who first described this table as a "multiresolution analysis of the literature". A coarser, wider ranging and light-hearted taxonomy is also given by Pope [133].

Literature review of time intervals and their perceptual functions				
IOI (msec)	Frequency (Hz)	Tempo (BPM)	M.M.	Comment
10000	0.10	6		Approximate minimum time of Newell's unit task level. Cowens maximum length of Long Auditory Storage. [155, 19]
5000	0.20	12		Centre frequency of Todd's gestural bandpass sensory-motor filter associated with whole body motion (body sway). Near longest limit of perceptual present interval. [104, 102, 35]
3000	0.33	20		Extent of subjective present: The level of temporal gestalt perception phenomenon. [181, 156]
2000	0.50	30		Fraisse's slowest interval for grouping. Dowling's estimate of slowest interval spanning the perceptual present. Cowen's estimate on minimum duration of Long Auditory Storage. [41, 35, 19]
1800	0.56	33		Unable to be predicted and clapped along to accurately. Parncutt's assessment of the longest interval of musically salient pulse sensations. Fraisse's longest interval able to be syncronised with. [131, 41]
1500	0.67	40	Largo	Slowest interval for tactus. Slowest interval for change from repeating sequence to isolated events [84, 55]
1400	0.71	42	Largo	Slowest limit on spontaneous tempo [55]
1000	1.00	60	Larghetto	Newell's approximate fastest interval for the elementary operations level within the cognitive band. Slowest interval of maximal time interval sensitivity. [155, 42, 107]
857	1.17	70	Adagio	Traditional renaissance tactus. [84]
800	1.25	75	Adagio	Slowest interval for easy perception of meter. [55]
710	1.41	84	Andante	Parncutt's mean preferred tapping tempo [130]
700	1.43	85	Andante	Average preferred or spontaneous tempo. [41, 68]
600	1.67	100	Andante	Desain and Honing's approximate tactus rate and preferred or spontaneous rhythm. Todd's centre frequency of his foot-tapping bandpass sensory-motor filter. Maximal accuracy in timing estimation. [27, 41, 104, 55]
500	2.00	120	Moderato	Meter perception threshold, faster than this, perceptual (automatic, universal) rather than cognitive or schematic rhythm induction occurs. Slower than this, knowledge / culture-dependent, schema-based control occurs. [156, 113, 10]

continued on next page

IOI (msec)	Frequency (Hz)	Tempo (BPM)	M.M.	Comment
375	2.67	160	Allegro	Fastest interval for tactus. Collier's highest ratio between triplets and swing rhythms of Jazz drummers. [84, 17]
333	3.00	180	Presto	Maximum of peak distribution of speech modulation frequencies [79]
300	3.33	200	Prestissimo	Maximum time interval sensitivity [42, 107]
250	4.00	240		Longest syllable length. Slowest period of the peak distribution of speech modulation frequencies [156, 79]
200	5.00	300		Fastest interval for easy perception of meter, fastest interval for maximal time interval sensitivity, synchronisation capability and spontaneous tempo. [55, 42, 107, 41, 131]
166	6.02	361		Freund's lower bound on fastest musical motor movements (trills). [156]
115	8.70	521		Fastest rate that subjective rhythmisation is still possible. [41]
100	10.00	600		Lower limit on Terhardts roughness frequency range. [79]
83	12.05	722		Freund's upper bound on fastest musical motor movements (trills). [156]
70	14.29	857		Streaming begins to occur.
50	20.00	1200		Point at which streaming has occured, a single auditory stream is perceived, rather than a sequence of events. [55]
40	25.00	1500		Order relation can be distinguished. Category boundary between "plucked" and "bowed" sounds. [156, 159]
30	33.33	2000		Order threshold: can produce an order relation (i.e order between the events can be distinguished). Proposed abstract fastest pulse, pace-maker or beat clock. [156]
10	100.00	6000		Approximate slowest limit on Newell's neural band. Approximate interval for preemption of the leading voice in ensemble playing. [155, 28]
5	200.00	12000		Fastest interval for the the fastest pulse period. Fastest interval corresponding to the upper limit on Terhardt's roughness frequency range [156, 79]
3	333.33	20000		Fusion threshold fastest interval: simultaneous, indistinguishable (even with different loudnesses, but same duration), a single event. [156]

continued on next page

IOI (msec)	Frequency (Hz)	Tempo (BPM)	M.M.	Comment

Table 2: Literature review of time intervals and their perceptual functions.

2.4.3 Non-causality of Rhythm

Leonard Meyer has noted in his seminal work "Emotion and Meaning in Music":

> "In short, earlier rhythmic groups influence later ones; or, to put it in another way, an established rhythmic process tends to perpetuate itself. *Equally important is the fact that future organization also influences grouping.* Thus the performer will play the first measures of the Mozart theme [of the first movement of his Piano Sonata in A Major] in such a way that its trochaic pattern is clear because he [sic] knows what the organization of the two-measure group is.
>
> It is, then the total disposition of all the musical materials that determines what the rhythmic grouping will be. This is another way of saying that the entire musical pattern will tend to be perceived in the simplest and most satisfactory terms" (my emphasis) [112, pp. 109].

Of course Meyer is not arguing omniscience of the performer: "Often the rhythmic organization is discontinuous, incomplete or ambiguous" [112, pp. 110], but that the intention of the performer is preconceived, possibly below a conciousness capable of self-inquiry or verbal articulation. Todd has also noted that the performer has the intention formed before performance:

> "It is assumed here that a trained musician is likely to be a good candidate for the status of "experienced listener" and, if a particular piece has been practised, is likely to have some kind of global understanding of the piece" [97, pp. 35].

This argues against simply viewing the rhythm perception of a listener in a causal manner, i.e we start from a tabula rasa and the rhythmic organisation builds up as we progress through the piece. This ignores aspects of temporal pattern completion, enculturated knowledge, temporal atoms (preferred interval categories, i.e 2:1, rhythmic feet) and training. There must be a significant retrospective assignment of the role of each event as the temporal context expands during performance.

2.4.4 Hierarchial, Multiresolution Rhythm

In summary, these points can be made:

1. Tempo constraints occur from memory and motor limits. These manifest themselves in preferred rates of rhythms and determine the structural priority of events as integral or superficial to the rhythm.

2. Subjective present acts as an integrator causing non-linearities in the perception of temporal intervals, such that certain rates, and relative ratios of rates of events are priviledged.

3. Objective accentuation confirms structure. Accentuation in all its forms is so prevalent in music that events will be expected to be accented simply from the rhythmic structure. The expectation of accented events from rhythmic structure allows syncopations to be acceptable and objective accentuation to be absent.

4. Meter, tactus and grouping are enculturated concepts that are encouraged by perceptual constraints, but may not be universal in the face of non-western bell-lines, polyrhythms and free rhythms. Non-western notions of rhythm do share many features with common practice, so clearly a uniform federation of behaviours are behind all forms of rhythm perception. Several attributes of these behaviours have been documented here, but the behavioural processes themselves may well be too distributed to be isolated.[15]

5. One attribute which can be identified is the perception of pulse. Mechanisms aiding syncronisation allow detection of repetition, which in the ideal case has exactly equal IOIs.

6. Time is organised into temporal levels. Each level corresponds to an increasingly wider time-span integrating period. The perception of a pulse corresponds to the organisation of a single level.

7. Regardless of privileged forms mentioned in Point 2, humans perceive and produce wide ranges of rhythmic rates. This suggests the ability to organise and adapt temporal levels and to process a large number of temporal levels.

[15]The term distributed is used in the sense of subsymbolic neural representations [182, 152], where a behavioural process may be modelled without isolating the data representation to interpretable symbolic forms.

8. Complex rhythmic structures can be considered to be the result of the interaction of several temporal levels.

9. Expressive timing corresponds both to the modulation of temporal levels and to short term distortions of periodicity. Modulation in the frequency of a temporal level accounts for the acceleration and deceleration in a performed rhythm. Time limited deviations away from periodicity can account for the leading or dragging of the introduction of a note and can account for swing when that deviation is itself applied repetitively.

From this survey of rhythmic attributes, it is clear that there is a complex interaction between event intervals forming temporal levels and the degree of perceptual strength of each event. In the next chapter, the signal processing theory of the continuous wavelet transform is introduced and applied to perform multiple resolution analysis of representations of musical rhythms. It will be shown that Wavelet theory reflects a significant portion of the features of hierarchy, accent and temporal context of musical rhythm which have been described here.

Chapter 3

Multiresolution Analysis of Rhythmic Signals

As has been detailed in Chapter 2, there are compelling arguments for analysing musical rhythms over multiple timescales and modelling rhythm in terms of multiple simultaneous periodicities. In this chapter, the theoretic formalism of a multiple resolution signal analysis using the wavelet transform is introduced. The wavelet transform is an improvement over the Fourier transform, the traditional means of signal frequency analysis, because it is better able to simultaneously localise the analysed signal in both time and frequency domains. Musical rhythm is then recast as a low frequency amplitude and frequency modulated signal. Morlet wavelets suitable for signal analysis are then introduced and their ability to analyse a musical rhythm is evaluated.

3.1 The Fourier Transform

The traditional analysis approach to reveal a time-domain signal's behaviour in the frequency domain is to use the Fourier Transform, which decomposes a signal into a series of *basis functions* consisting of weighted complex exponentials.[1] For the signal s in the continuous time t case, $\hat{s}(\omega)$ is the Fourier transform of $s(t)$ for each continuous frequency ω:

[1]Readers may wish to refer to the excellent text by Proakis and Manolakis [138] for background mathematics to signal processing and Fourier analysis.

$$\hat{s}(\omega) = \int_{-\infty}^{\infty} s(\tau) \cdot e^{-i\omega\tau}\, d\tau,$$

where i is the imaginary identity $i = \sqrt{-1}$. The natural exponential base is representated as e, with the Eulerian identity

$$e^{-i\omega} = \cos\omega - i\sin\omega$$

composed of sine and cosine basis functions, independent of t and thus of infinite duration, so that $s(t)$ is decomposed into a set of scaled basis functions.[2]

The Fourier transform does not convey information about translation of the signal in time, and therefore can not reflect frequency change over time. This issue is addressed by the Short Time Fourier Transform (STFT), a time limiting *windowed* version of the Fourier Transform

$$\hat{s}(t,\omega) = \int_{-\infty}^{\infty} s(\tau) \cdot \bar{h}(\tau - t) \cdot e^{-i\omega\tau}\, d\tau,$$

where $\bar{h}(t)$ is the complex conjugate of the window function. Significantly, $h(t)$'s time scale is independent of the harmonic number $f = \frac{\omega}{2\pi}$, the same window function is used for all harmonic components (see Figure 4). In the discrete version of the STFT, all harmonic components produced by the analysis are assumed to be in ratio to $h(t)$, which is assumed to extend in time to cover the period of the fundamental frequency of the signal. The energy of any components in the original signal *changing* in frequency, that is, not matching the assumption of strict harmonic ratio to $h(t)$ within its time extent, will be distributed ("blurred") across the window.

While several different window functions have been proposed, such as the Hanning or Hamming windows [138], most tend to have the general characteristic shape of a Gaussian probablity density envelope

$$g_\alpha(t) = \frac{1}{2\sqrt{\pi\alpha}} \cdot e^{-t^2/4\alpha}$$

over the basis function. A Gaussian window function has the property that it is invariant between time and frequency domains, therefore producing the best simultaneous localisation in both domains with respect to Heisenberg's uncertainty relation [9, pp. 440], [94, pp. 33]

[2] A corresponding synthesis equation allows reconstruction from the Fourier components back to $s(t)$ [138].

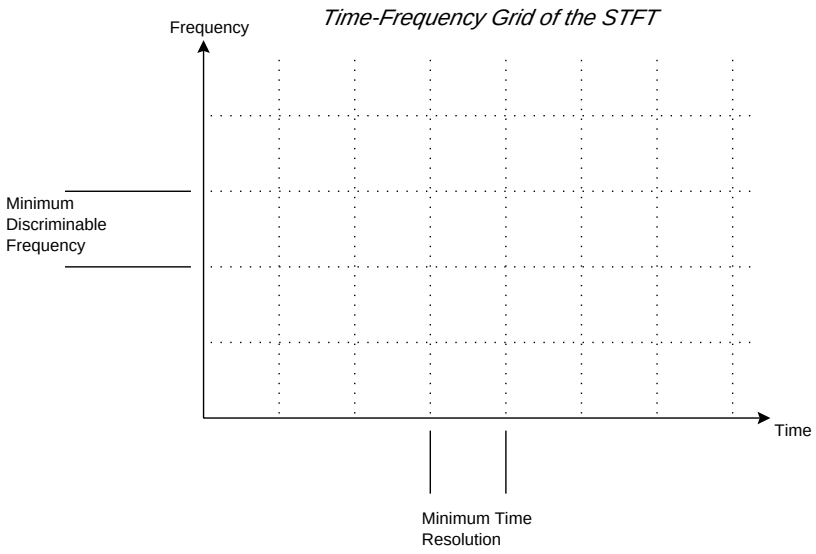

Figure 4: Analysis extents in the time and frequency domains of the Short Term Fourier Transform. Of significance is the same minimum time extent for each frequency band. High frequencies will be capable (due to their short wavelength) of changing frequency within the minimum time extent. This change within the time-frequency "box" will be distributed over the boxes' area.

$$\delta t \cdot \delta \omega \geq \frac{1}{4\pi}.$$

This lead Gabor to propose its use for basis functions which incorporate both time and frequency [44].

3.2 Rhythm as an Amplitude Modulation

Before introducing further signal analysis techniques, it is necessary to cast musical rhythm in terms of a signal capable of being analysed without misrepresenting characteristics of the perception of rhythm as described in Chapter 2. An ideal,[3] perfectly isochronous rhythm can be considered as an amplitude modulation [34] of the auditory frequency ranges. Such a view was first noted by Todd [101]. A simple and fundamental example is illustrated in Figures 5 to 9. A constant tone,

[3]Here the term "ideal" is used in the signal processing sense of a theoretical signal or system that is unrealisable in practice. While an isochronous beat can be produced by machine, a human performer will neither achieve such accuracy, nor as we have seen in Chapter 2, wish to do so, as it would prevent communication of long term structure.

the carrier, comprising a single frequency and therefore energy at a single spectral component,

$$y_c = \sin(2\pi f_a\, t),\ 0 \leq t \leq 1, \tag{2}$$

(in this example, the auditory frequency $f_a = 440$Hz) is modulated in its amplitude by a much lower frequency function, in this simplest case, also a sinusoid, with a DC and phase offset (Figure 5):

$$y_m = \cos(2\pi f_r t + \pi) + 1,\ 0 \leq t \leq 1, \tag{3}$$

where $f_r = 4$ is the 4Hz rhythmic frequency. The modulation is performed by multiplying in the time domain the rhythmic modulator periodic function with the auditory carrier periodic function (Figure 8),

$$y = y_c y_m. \tag{4}$$

The corresponding separate Fourier domain representations of the carrier and modulator are shown in Figures 6 and 7 respectively.

The Fourier domain representation of the resulting modulated signal is that of a convolution of the two spectral impulses of Figure 7, with the rhythmic frequency forming sidebands around the auditory tone (see Figure 9). Significantly, there is no spectral energy at the original rhythmic frequency component due to the convolution.

In our contrived example above, the rhythms have periods that exactly match the Fourier analysis window. Implicitly, there is a rectangular window over the data (see Moore [119] and Proakis and Manolakis [138]). The window and the rhythm frequency have been chosen to minimise the effect of the windowing function in order to demonstrate the spectral characteristics of rhythm without requiring a STFT. Only in this example of a periodic sinusoidal amplitude envelope perfectly matching the analysis period of the Fourier transform, will the rhythmic frequency be a single spectral component. In contrast, an anapest rhythm (Figure 10) will alter the periodic nature of the isochronous pulse. The resulting multiband spectrum will be distributed over the Fourier window (again the rhythm frequency has been chosen to match the analysis window to avoid artifacts), appearing as composed of low

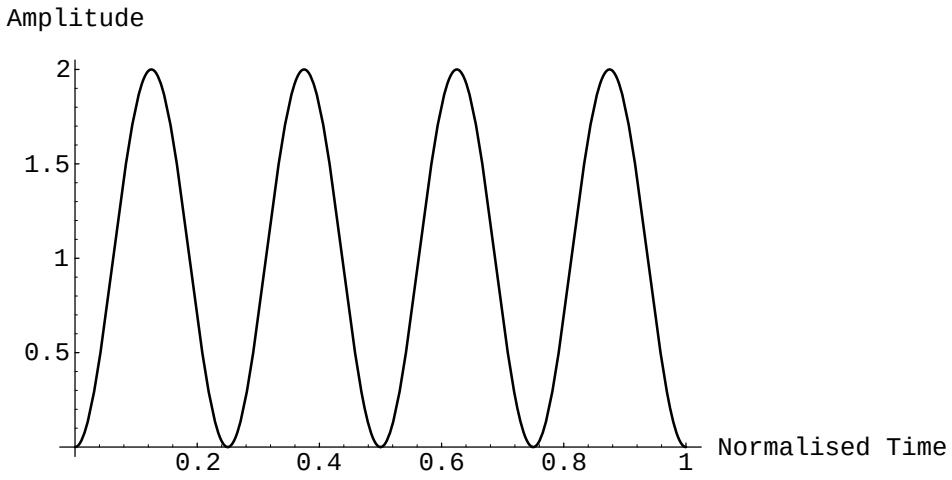

Figure 5: An amplitude function formed by DC shifting a low frequency sinusoid.

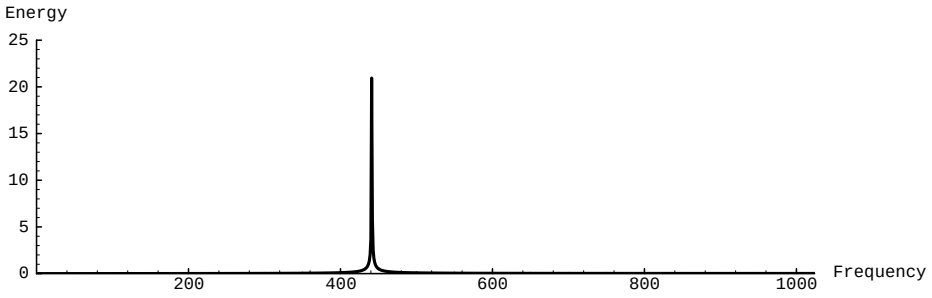

Figure 6: A Fourier transform of the acoustic 440Hz pitch function.

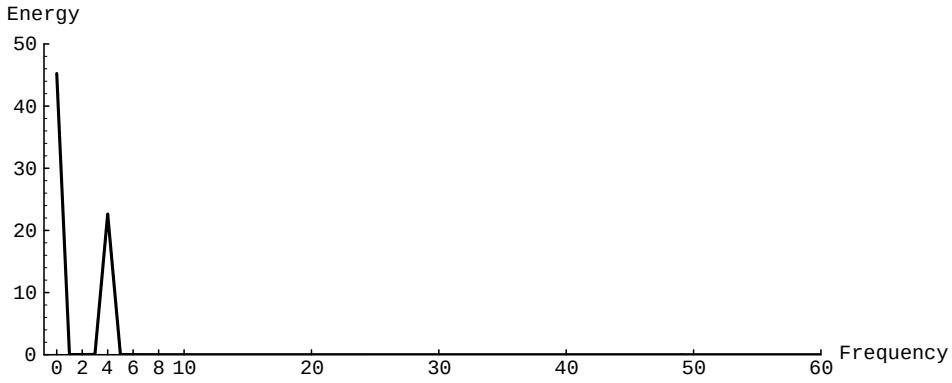

Figure 7: A Fourier transform of the rhythmic amplitude function in Figure 5. The two spectral peaks are at the zeroth component (the DC offset) and the fourth harmonic (the rhythmic frequency).

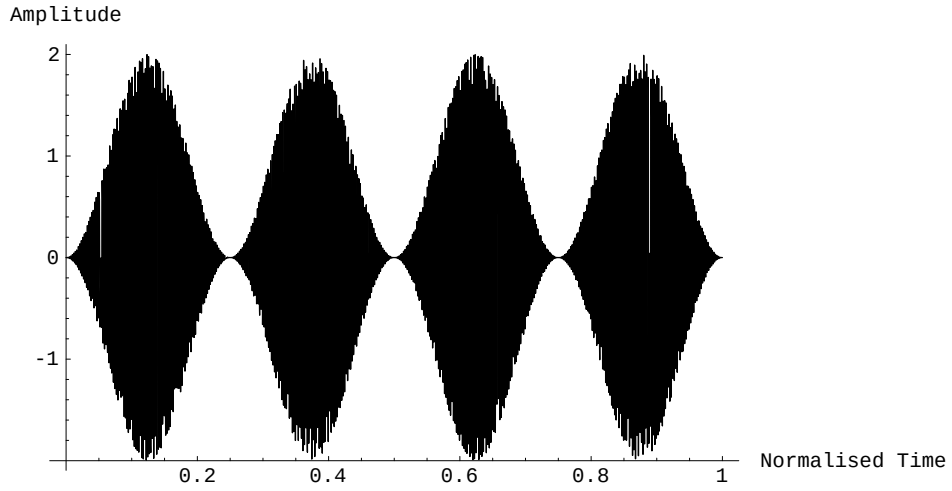

Figure 8: A time domain representation of the rhythmic amplitude function in Figure 5 multiplied (in the time domain, convolved in the frequency domain) with the acoustic 440Hz pitch function of Figure 6.

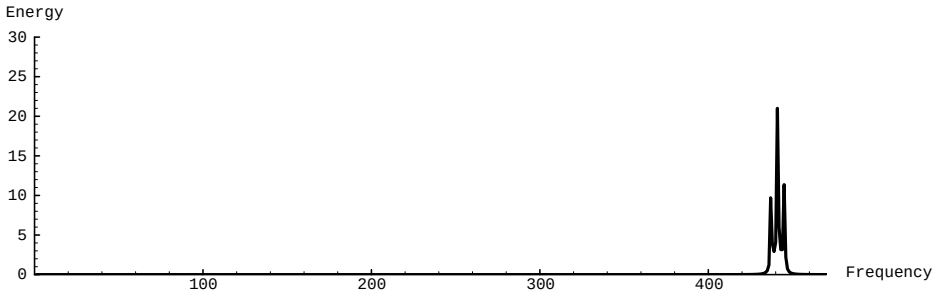

Figure 9: A Fourier domain representation of the rhythmic amplitude function in Figure 7 multiplied (in the time domain) with the acoustic 440Hz pitch function of Figure 6. The two rhythmic sidebands surround the central acoustic carrier frequency.

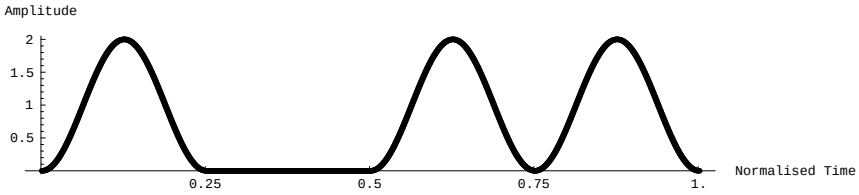

Figure 10: An anapestic rhythm.

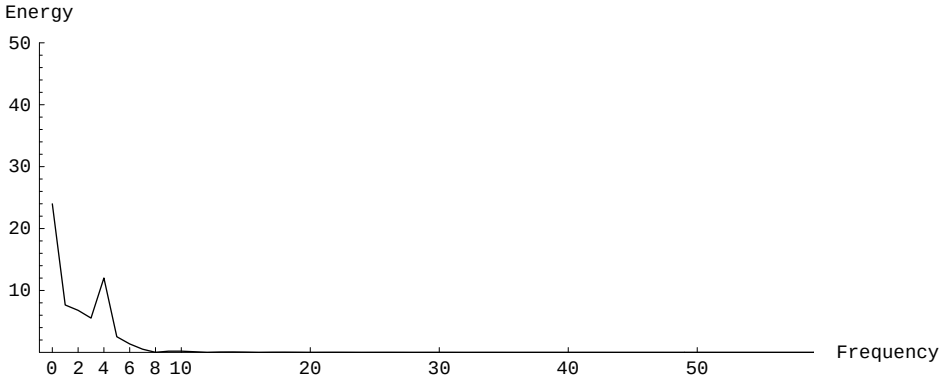

Figure 11: A Fourier domain representation of the energy of the rhythmic amplitude function in Figure 10.

order harmonics with energy inverse to harmonic order, and a spectral component corresponding to the frequency (within the window) of the beat (Figure 11).

Real instrument tones will comprise complex, time varying, clusters of spectral components rather than a pure sinusoid, and unless the rhythm is isochronous, its Fourier representation will also form a cluster of sidebands on either side of the carrier frequency. In addition, the pitch content of a rhythm will vary with chorusing and vibrato when performed on many instruments such as string, wind, percussion, or singing. The rhythmic frequency will therefore not be discriminable sidebands around an acoustic carrier.

This poses a problem in attempting to discriminate and analyse a rhythm from an acoustic signal (by first determining its acoustic carrier) in the spectral domain, even if the STFT is used. In order to extract such a rhythmic signal from an acoustic signal, a deconvolution would require a prior knowledge of the time-varying acoustic component of the signal. Deconvolution approaches such as homomorphic deconvolution [138] requires windowing between two separate frequency domains using the cepstral coefficients (logarithmic domain) of the two convolved signals. Another approach, adopted by Todd [103] is to take the signal energy of the acoustic signal of Equation 4, $|y|^2$, effectively a rectification, which will re-establish the rhythm modulation function independent of the acoustic carrier.

3.2.1 Capturing Musical Intention

An alternative pragmatic approach to this rhythmic convolution problem 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 [63] can transduce the time of a drum strike[4] and a measure of intensity of the strike.[5] This MIDI data is available for analysis before the synthesis hardware generates the audible sounds (sampled drum or synthetic sounds).

[4]The temporal resolution of MIDI is at best 1 msec [90, 118], however most MIDI synthesis devices are not able to respond within those time constraints [121] and a time resolution of *reception* is approximately 16 msec. Currently, the temporal resolution of the Roland PAD-8 drum controller *generating* MIDI data hasn't been tested but is confidently assumed to be within that measure. The processing required to transduce the drum strikes from a threshold measure of a piezo sensor to MIDI note values is low.

[5]The MIDI velocity value, a measure in arbitrary units of 1-127 [63].

An acoustic drum provides a wide variety of control over the timbre and pitch of the played sound, depending on the intensity of the strike and location on the drumskin of the strike. Currently available MIDI drum pads do not indicate such timbral information.[6] As reviewed in Chapter 2, timbre and non-tuned pitch will influence accent perception, but this will typically be accompanied by dynamic (intensity) based accents. Therefore the MIDI velocity value is assumed to be a meaure of the total intended accent by the performer. Other MIDI controllers (keyboards, wind instruments, guitars) are also available to capture rhythms from performances other than percussion music. For an interactive performance situation, there is a significant reduction in data processing by capturing the rhythm directly via MIDI, or another performance transduction process, without the preprocess of recovering the rhythm from the acoustic signal.

3.2.2 Representing Rhythm for Analysis

A rhythm can be perceived, memorised and reproduced independently of the music's original pitch and timbral content. As noted by Longuet-Higgins [86], rhythmic figures (such as a dotted rhythm), can be distinguished regardless of the accompanying tonal development. Compositions purely for percussion of indefinite pitch and percussion solos are interpretable. In addition, the articulation and even intensity components can be removed while still preserving the sense of rhythm [148, pp. 64]. Even using very short impulse-like clicks, a familiar rhythm can be recognised, or a new rhythm comprehended and tapped along with. The rhythm is induced from the IOI's between events alone [82, 148].

Pitch, tonality, harmony and timbre obviously play a role in rhythmic interpretation. In reciprocal manner, tonal interpretation is strongly influenced by rhythmic structure. This has been well characterised and modelled by Rowe's *Cypher* system and his other performance interpretation models as two parallel behaviours which provide mutual evidence to aid the other process [150, 149]. The rhythmic interpretation will inform the tonal interpretation and then that result will then revise the rhythmic interpretation, in a controlled feedback loop. The tonal interpretation can therefore be seen to be weighting the salience of each beat.

[6] Nor is the MIDI specification designed to adequately represent such timbral information [118]. For instance, Korg Corporation's "Wavedrum", which uses a standard drum skin to trigger physical models of drums uses the location of the drum strike as a control dimension. However it communicates an impoverished representation over its MIDI channel [151].

Therefore the approach adopted in representing a rhythmic function for analysis has been to take the short duration tap in the limit and represent the time of the onset of each beat as an unit impulse function [138]:

$$\delta(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where t is the sample index at the onset of the note. The rhythm function for a piece of music is therefore an uneven train of pulses with intervals of zero valued samples matching the IOI between beats. Such a representation has been used by other researchers [80, 12]. The uneven pulse train can also be viewed as a sampling of the amplitude envelope signal. By the sampling theorem [138], a continuous function $x(t)$ composed of component frequencies less than the Nyquist rate ($F_s/2$), may be discretely approximated by the summation of weighted, time-shifted impulse functions:

$$x(t) = \sum_{k=-\infty}^{\infty} x(k)\delta(t-k).$$

Recalling the amplitude function representing the rhythm of Equation 3, a pulse-train function can be seen to be a *minimal* or *critical* sampling of the amplitude envelope at the lowest sampling frequency which still accurately represents the rhythm function — one sample at the stationary point of each "lobe" of the envelope (see Figure 12). The sampling rate can be a reasonably low rate as the audible frequencies are not present. A rate of 200Hz is appropriate and was also used by Brown in her analysis of rhythm [12].

Weighting the impulse height by a normalised measure of the intensity of the beat

$$c(v) = \frac{v}{127}, \tag{6}$$

where v is the MIDI velocity value, incorporates the effect of dynamic accent, by

$$\iota(t) = c(v) \cdot \delta(t), \tag{7}$$

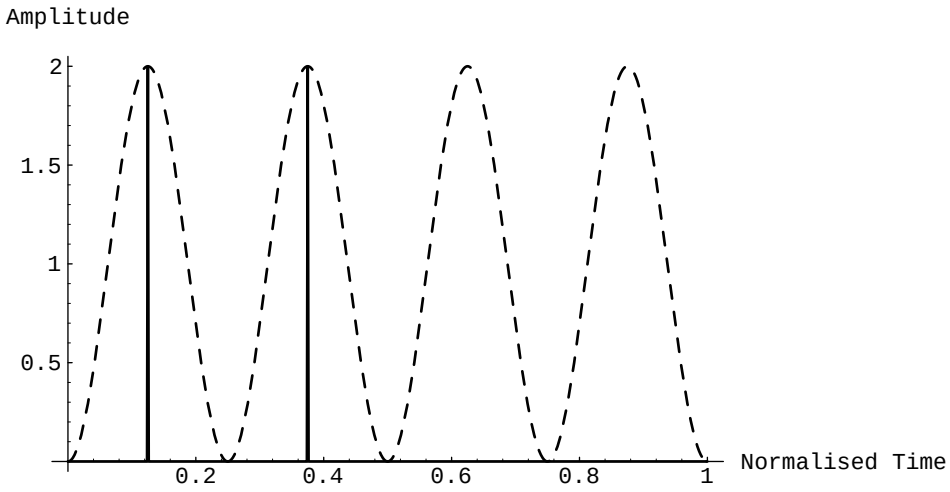

Figure 12: Points comprising a critical sampling of a rhythmic amplitude function.

where $\iota(t)$ is the rhythm function composed of sparse impulse values (0.0–1.0). Here we are assuming there is a linear relationship between the perceptual salience of an individual dynamic accent and the intensity of a beat.

This is ignoring the effect of masking of beats by temporal proximity and other non-linearities between intensity and its final perceptual salience. Masking, auditory streaming effects [10] and expectation (for example, from tonal structure) could be modelled by a hypothetical non-linear transfer function version of $c(v)$ in place of Equation 6, which would summarise the total effect of context on the perceptual impact of the beat. Alternatively, if a frequency representation is used which preserves energy (Parsevals relation [138]), such perceptual effects could be modelled in the frequency domain. As section 2.1.4 reported, masking and streaming occurs with IOIs shorter than 200ms, smaller than intervals found in performed rhythms analysed here, so the impact of local proximity is assumed here to be negligible. The opportunity to capture the intensity of the strike of the performer with a drum pad posits the linear weighting of the impulse as an acceptable approximation of the total intended accent to be communicated to the listener. As I have argued, expressive timing, agogic, dynamic and other objective accents will produce complex, frequency and amplitude varying rhythmic signals which will require a non-stationary signal analysis technique. Analytical wavelets are well suited to this task.

3.3 The Continous Wavelet Transform

Wavelet theory is a recent convergence between independent research in image processing, coding theory, applied mathematics and seismology [144, 20, 186, 48]. It 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.

In contrast to the STFT, the continuous wavelet transform (CWT) [52, 20], decomposes the signal onto scaled and translated versions of a *mother-wavelet* or *reproducing kernel* $g(t)$,

$$W_s(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(\tau) \cdot \bar{g}(\frac{\tau - b}{a})\, d\tau,\ a > 0, \tag{8}$$

where a is the scale parameter, controlling the dilation of the wavelet function, effectively stretching the wavelet geometrically over time. The translation parameter b centres the wavelet in the time domain. Each of the $W_s(b, a)$ coefficients weight the contribution of each basis to compose $s(t)$. The geometric scale gives the wavelet transform a "zooming" capability over a logarithmic frequency range, such that high frequencies (small a) are localised by the window over short time scales, and low frequencies (large a) are dilated over longer time scales [75].

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

$$s(t) = \frac{1}{c_g} \cdot \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_s(b, a) \cdot g(\frac{t - b}{a})\, \frac{dadb}{a^2}, \tag{9}$$

Equation 8 uses τ to indicate the time variable is used as an integrating parameter, whereas t in Equation 9 indicates the resulting signal at each time point. The constant c_g is set according to the mother wavelet chosen

$$c_g = \int_{-\infty}^{\infty} \frac{|\hat{g}(\omega)|^2}{|\omega|}\, d\omega < \infty. \tag{10}$$

The analysing mother wavelet must meet admissibility conditions of finite energy from absolute and square integratibility, given by,

$$\int_{-\infty}^{\infty} |g(t)| dt < \infty, \tag{11}$$

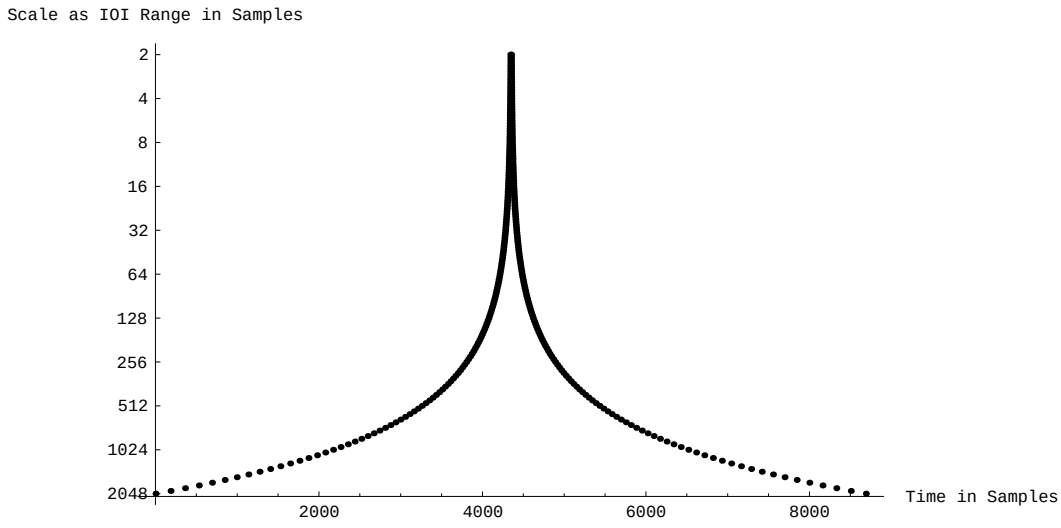

Figure 13: Time extent of scaled Morlet wavelets (Equation 8) over a range of 10 octaves.

and

$$\int_{-\infty}^{\infty} |g(t)|^2 dt < \infty. \tag{12}$$

The second admissibility condition (Equation 12) implies, in practice, a zero mean (no DC bias),

$$\hat{g}(0) = 0, \tag{13}$$

or,

$$\int_{-\infty}^{\infty} g(t)\, dt = 0,$$

which produces a finite time supported short wave — hence the term wavelets. The support of a function s is the closure of a set of points t where $s(t) \neq 0$. The time support of scaled wavelets of Equation 8 are shown in Figure 13.

The CWT indicated in Equation 8 is a scaled and translated filter from a constant relative bandwidth[7] (Equation 10) filter bank, comprised of an infinite number of filters or "voices". For implementation, the scale parameter a must be discretised with a sufficient density of voices per octave.

3.3.1 Morlet's Analytical Wavelets

Grossmann and Morlet [52], have applied a complex-valued Gabor mother-wavelet for signal analysis,

$$g(t) = e^{-t^2/2} \cdot e^{i\omega_0 t}, \tag{14}$$

where ω_0 is the frequency of the mother-wavelet (before it is scaled). In essence, this is a Gaussian window over cosine and sine curves which are in the real and imaginary planes respectively (See Figure 14). In the frequency domain the wavelet has the form

$$\hat{g}(\omega) = e^{-(\omega-\omega_0)^2/2} - e^{-(\omega^2+\omega_0^2)/2}. \tag{15}$$

Subsequently, Kronland-Martinet, Morlet and Grossman applied such a wavelet to sound analysis [76, 75, 74]. The research reported here differs from this earlier work in that it is the rhythm signal (the function that modulates the auditory carrier) that is analysed using so-called Morlet wavelets — not the raw sound signal. Here the rhythm is analysed independently (effectively deconvolved) of the auditory, carrier, component.

Equation 14 is close to a "progressive support" or "analytic" wavelet, which is defined as

$$g(t) = u(t) + iv(t), \tag{16}$$

where $v(t)$ and $u(t)$ form a Hilbert Pair

$$v(t) = u(t) * \frac{1}{\pi t}, \tag{17}$$

[7] Known in engineering terms as "constant-Q" [185].

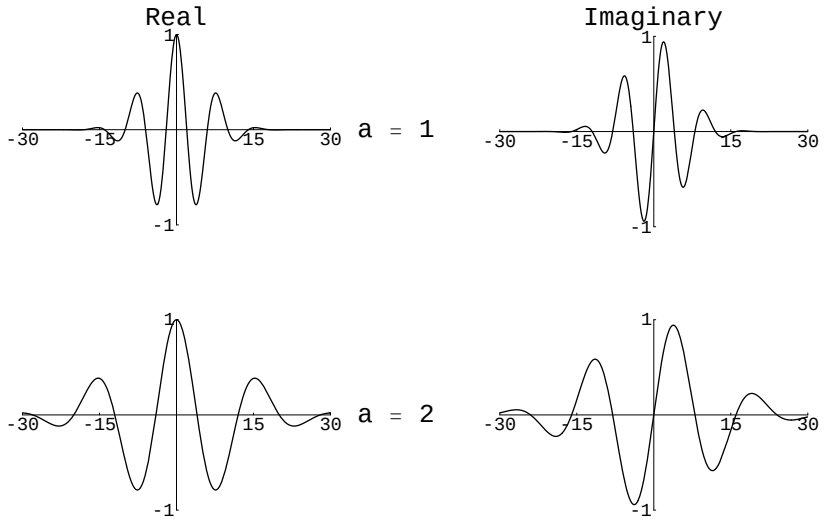

Figure 14: Time domain plots of Morlet wavelet kernels, showing real and imaginary components for the mother wavelet and a version dilated by $a = 2$.

where $*$ represents the convolution operator [138]. The real and imaginary components being the Hilbert transform of each other [54] produces

$$\forall \omega < 0 : \hat{g}(\omega) = 0. \tag{18}$$

That is, the wavelet analysis produces positive frequencies only, conversely, regressive support produces negative frequencies only [75, pp. 53]. However, due to the asymptotic tails of the Gaussian distribution, the Morlet wavelet kernel is not progressive, nor does it meet the admissibility condition of Equation 13 for reconstruction. With the exception of analysis-by-synthesis approaches, much analysis can be performed without requiring reconstruction, so this is not a problem in practice.

The internal frequency ω_0 determines the number and amplitude of oscillations under the Gaussian envelope in the time domain and effectively controls the frequency discrimination (bandwidth) of the wavelet. With a sufficient $\omega_0 > 0$, the negative frequency components can be made small enough to make Equation 14 nearly progessive, and they can be removed with a corrective term [57, pp. 31]. Daubechies [20, pp. 76] and Holschneider [57, pp. 32] have suggested ω_0 should be set so that the second oscillation of the real component of $e^{i\omega_0 t}$ meets the envelope $e^{-t^2/2}$ at half its maximum value. The smallest value is $\pi\sqrt{2/\ln 2} = 5.3364$, in this

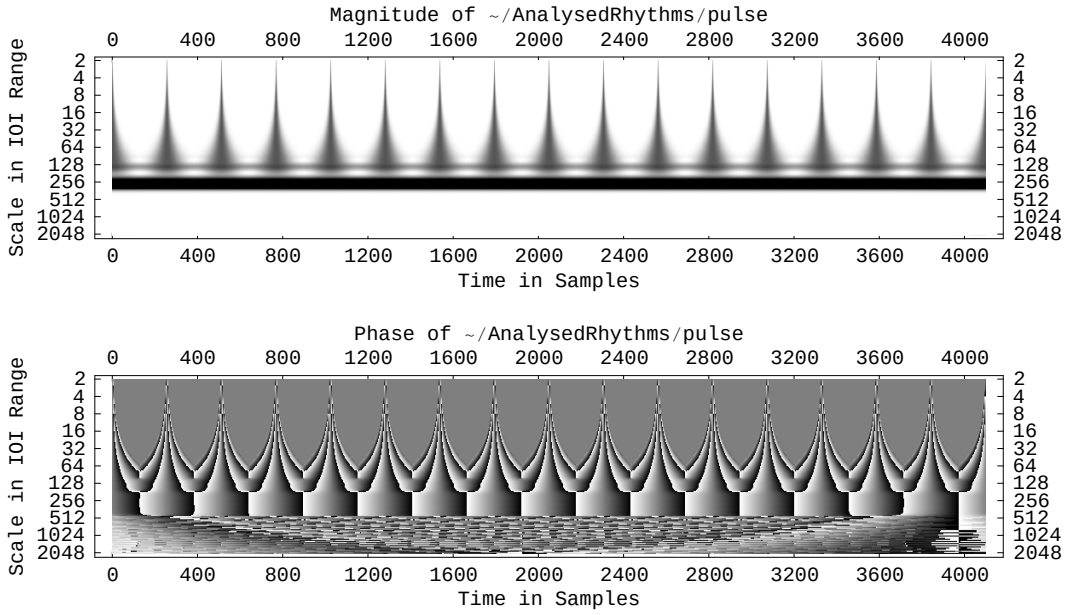

Figure 15: Scalogram and phaseogram plots of an impulse train spaced with an IOI of 256 samples. The most activated ridge is centered on the scale corresponding to an interval of 256 and a lower energy ridge is centered on the interval of 128. This secondary ridge occurs from interactions between secondary lobes of the wavelet.

research $\omega_0 = 6$ was determined experimentally to energise the correct scale with a reciprocal period matching the IOI between two impulses. This matches Guillemain and Kronland-Martinet's suggested value [53], while Daubechies suggested 5.

The conservation of energy 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

$$Z_s(t) = A_s(t)e^{i\phi_s(t)}. \tag{19}$$

The magnitude and phase of the wavelet coefficents $W_s(b, a)$ can then be plotted on a linear time axis and logarithmic scale axis in grey scales as a "scalogram" and "phasogram" (see Figure 15). The magnitude being

$$A_s(b, a) = |W_s(b, a)| = \Re[W_s(b, a)]^2 + \Im[W_s(b, a)]^2 \tag{20}$$

and phase

$$\phi_s(b,a) = \arg(\Im[W_s(b,a)], \Re[W_s(b,a)]). \tag{21}$$

$\Re[x]$ and $\Im[x]$ are the real and imaginary components of the CWT coefficients. These representations were first described by Grossman and co-workers [51]. Magnitude values are mapped from lowest energy levels to white, highest energy levels as black. 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. Alternatively, the phase can be mapped onto a colour wheel so that the white to black transition is less prominant, but a sense of phase regularity is still apparent. To improve clarity, phase values are clamped to 0 where they correspond to low magnitude values, otherwise

$$|W_s(b,a)| > \epsilon_m \tag{22}$$

where the magnitude threshold $\epsilon_m = 0.01$, registers the phase measure as valid.

3.3.2 Wavelet Properties

Non-orthogonality

Despite the admissibility conditions of Equation 11 and 12, there remains many choices for mother wavelets. Orthogonal basis functions produce a non-redundant transform with a compensated aliasing between filters that enables perfect reconstruction [20]. Such wavelets are used in coding and compression applications as they minimise the number of wavelets representing the signal through the use of *dyadic grids*, discarding redundant coefficients and enabling discretisation. The difference between such discrete wavelets and the continuous wavelet transform of section 3.3.1, is that discrete filters are not exact scaled versions of each other. Under certain conditions, the discrete transform will converge, after a number of iterations, to scaled versions of previous filters. In order for the filters to converge to a continuous limit, the filters must be *regular*, i.e. differentiable, requiring sufficient zeros [138] at the Nyquist rate to attenuate repeat spectra.

However, orthogonal wavelets are not suitable for this application because they are translation dependent as they do not preserve the phase of the signal [185]. That is, a signal shifted by one sample will produce a very different distribution over the wavelet coefficients compared to the unshifted decomposition.

Applicability of the Morlet wavelet to rhythm analysis

Without orthonormality of the basis, the wavelet expansions are not linearly independent. This implies there are many different superpositions of the basis functions which will sum to give the original signal. Therefore the choice of wavelet for signal analysis is concerned with making explicit the components of the signal that in general match the problem domain. Coefficients $W_s(b, a)$ of a decomposition of a signal represent how close the signal is to each scaled and translated basis function [144]. For signal analysis purposes then, the nature of both the wavelet and the signal strongly influences the interpretability of the decomposition.

As the time domain plots indicate (Figure 14), 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 [138]. Non-causality implies the wavelet transformation must be performed on a recorded copy of the entire signal, and so is physically unrealisable in real-time. Therefore the wavelet is best considered in terms of an ideal theoretical analysis kernel, rather than one existing in vivo as a listener's peripheral perceptual mechanism.

However, there are reasons to entertain the idea that the mechanisms used in the process of rhythm induction are not fully causal. Enculturation of rhythms from previous listening can be argued to construct a schema approach to perception. This has been argued for tonality by Leman [83], for rhythm by Jones [65], and in terms of pulse sensations, by Parncutt [130]. New rhythms are perceived with respect to previously heard rhythms and are organised and anticipated within the harness of a particular schematisation. In that sense, the perception of a beat in the present has an expectancy weighting, projecting from the future back towards the present.

Indeed, a performer will practice a phrase such that each beat is performed within the context of beats imagined and intended, but yet to be performed (see also Todd's similar argument [97]).[8] Given the cross-cultural nature of most musical development, most listeners will share and understand the cultural implications as the phrase develops. They will predict subsequent beats, and draw meaning and emotion from the confirmation of such predictions, as Leonard Meyer has argued (see section 2.4.4 and [112]). We need the corpus of learned rhythms as a reference point to subsume the new performance. A purely causal model will be limited in its

[8]Even in improvised music, with the notion of learned riffs, Indian paltras [169] and other improvisation training methods [3], each beat is performed within the context of intended future beats as well as those beats already performed.

success because it is not taking into account the (culturally relative) retrospection possible of the performance as it proceeds. This poses a serious, though perhaps not insurmountable, problem for computational approaches, namely providing the representation of *schematic* expectancies[9] that a musician will accumulate in the process of listening and performing.

The non-causal projection of the Morlet wavelet can therefore be viewed as an idealistic aggregation of such predictive memories. Backprojection of the filter is a sense of completion of an implied rhythm. It functions as retrospective assessment of the rhythm, as argued by Desain [23], Jones and Boltz [67] and Scheirer [154]. Its use purposefully does not seek to apportion rhythm perception behaviour between biological and cultural processes. It is in the non-causality that this work differs most dramatically from the philosophy and results of Todd's rhythmograms [103].

Clearly the Morlet wavelet is an oversimplification of the rhythm perception process. Construction of a wavelet which is a closer model of the auditory system, in a similar manner to the approach of Todd in using filters derived from gammatone and haircell models [103, 105, 106, 107, 108], may be possible. Despite the Morlet wavelet being a theoretic formalism, and being a basis for smooth functions, it has several positive attributes as a wavelet for rhythm analysis.

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 time and frequency domains of the Gabor transform (described in section 3.1) also holds for the Morlet wavelet. This wavelet therefore has the best simultaneous localisation of change in time and frequency. Other wavelet 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 quantifying the representative abilities of other multiresolution approaches to rhythm models, particularly Todd's rhythmograms. Towards such uses, the wavelet behaviour in analysing an impulse train is now considered.

[9] "Abstract structural regularities of the music of one's culture", in contrast to *veridical* expectancies, which arise from the particular musical events attended to in a performance [5, pp. 498].

3.3.3 Wavelet Analysis of an Impulse

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

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 (small a) and spreads the effect across longer finite time periods at lower scales. Due to the non causality of the Morlet wavelet, at each scale and translation of Equation 8, the impulse will be projected simultaneously forward and backward in time in the time-frequency plane, matching the support of the wavelet. This forms an *influence cone* [57] which has a time interval, for each scale and translation, between $[at_l + b; at_r + b]$ for a mother-wavelet with support over the interval $[t_l, t_r]$.

As detailed by Grossmann and co-workers [50, 51], [76, pp. 279], and Solbach et. al [168], a *singularity* such as an impulse will be marked by a localised increase in the modulus at high frequency scales and a constant phase across frequency scales, independent of the mother-wavelet used.

An analysis of an isochronous train of impulses (Figure 16, that is, of a constant beat frequency) with a bank of Morlet wavelets has been shown in Figure 15. The abscissa axis represents time in samples, the ordinate axis is logarithmic, represented here by the time extent of each wavelet voice, again in number of samples. The scale with the highest modulus, indicating energy, corresponds to the frequency of the beat — the reciprocal of the IOI, as indicated in Figure 17. The absolute timing of the onset intervals between beats will be reflected by the absolute scale number. The phase of a periodic component is indicated by a regular shade transition at the scale corresponding to 256 samples. For scales lower in frequency than this the modulus falls to zero and the calculation of phase becomes ill-defined.

From Figure 17, the relative energy levels of each scale is indicated. In addition to the most highly activated scale corresponding to 256 sample IOI, there is a secondary lobe of half amplitude energy at the first harmonic of the beat rate (128 samples). This is caused by coincidence of the half-amplitude second oscillations of the kernels in the time domain (Figure 18, by $\omega_0 = 6$). The forward time projection of the nth 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

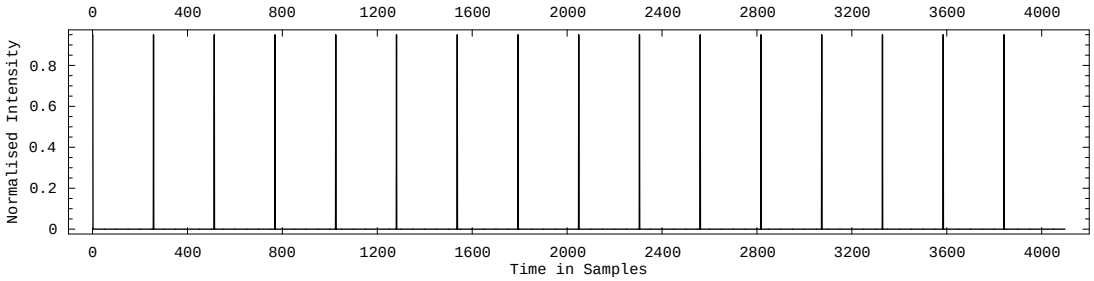

Figure 16: Plot of the time/amplitude signal of a simple isochronous pulse. The sample value at the time of each beat is non-zero, values at all other times are zero.

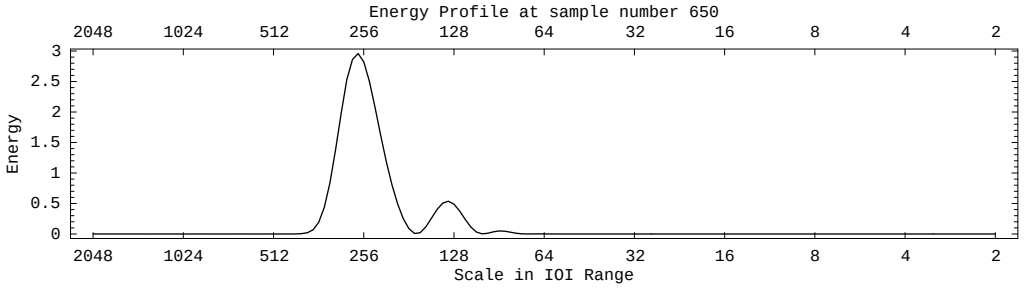

Figure 17: Modulus displaying the signal energy distribution over all wavelet voices at the 650th sample time point.

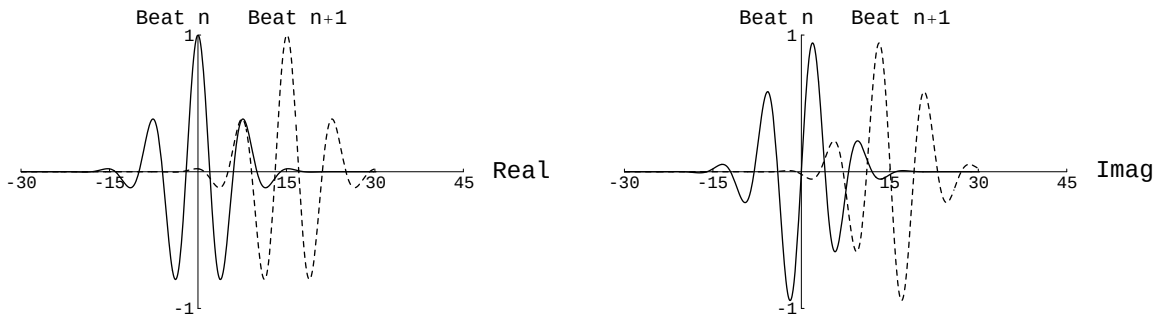

Figure 18: Time domain plots of the overlap of the real and imaginary components of Morlet wavelet kernels. These demonstrate the cause of the reduced energy second harmonic in the scalogram.

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 signal energy at the third harmonic can be discerned in Figure 17.

While artifactual in nature, these harmonics can be considered as representing a listeners lower propensity to perceive an isochronous rhythm as actually 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 was modelled explicitly by Desain's decomposable rhythm model [23], by forward projecting in time expectancy curves with reduced amplitude at second and third harmonics. This effect occurs in the Morlet wavelet as a by-product of the nature of the Gaussian envelope 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 rhythm perception,[10] rather than learned. Simplification of rhythmic ratios towards 2:1 in reproduction tasks [55, 135, 41] does indeed show that these ratios are privledged. Ubiquity in music notation, musical performance practice and and other activities involving doubling a motor behaviour are easily accomplished by humans. It is quite possible that this motor production optimisation is matched by an inherent perceptual process favouring simple subdivisions of

[10]Insofar as the assumption holds that rhythm perception is accurately represented in some skeletal sense by a decomposition into short term oscillations.

time.

Parncutt's findings in tapping experiments [130] suggest there should be near equal propensity of the listener to also consider an isochronous rhythm at half the beat rate (i.e 512 sample IOI). The relative energy levels of any rhythmic harmonics from Morlet wavelets would need to be derived by matching against perceptual measures. Parncutt's tapping experiments demonstrated listeners would tap occasionally at duple and triple meters to the stimulus meter, so second and third harmonics and subharmonics would need to be introduced. It is possible that tempo constraints applying to intervals longer than the subjective present are suppressing the fundamental rhythmic frequency, such that only sufficently fast harmonics of this fundamental are perceptible. However this conjecture has not been tested in this research and would not, by itself, explain Parncutt's near equal propensity for half-beat finding. Such an effect could be explained by top-down expectation from subjective rhythmisation.

3.4 Phase Congruency and Local Energy

Phase indicates the progression of a periodic wave though its cycle. Therefore an oscillating phase at a scale indicates that frequency is present in the signal being analysed. As suggested by Grossmann [51], inspection of a phasogram for regularly spaced smooth progressions through the grey scale leading to dark to white transitions indicate the presence of a frequency at a scale 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 partials of the image signal are most in phase synchronisation over a range of frequency scales [120, 73]. This behaviour has been termed *phase congruency* and peaks in the local energy function can be used to indicate points of maximum phase congruency. The model is capable of predicting the effects of Mach banding on trapezoidal intensity profiles. It is proposed here that phase congruency can be adapted to be a new measure of the structural significance of beats within a rhythmic context.

In the single dimension case, phase congruency is indicated by a constant shade or colour across scales at a given time point in the phase plot. The local energy function $E(t)$ of the signal $s(t)$ at time t is defined as

$$E(t) = \sum_{n}^{N} \sqrt{\Re[W_s(t,n)]^2 + \Im[W_s(t,n)]^2}, \tag{23}$$

where n is the scale index or "voice" of the wavelet, N is the number of voices in the discretisation and t, n substitute for b, a. Each $W_s(t,n)$ can be considered as a vector of a given magnitude and rotation (per Equations 20,21) around the scale axis at time t (see Figure 19). $E(t)$ is therefore the vector sum of those coefficents. To provide a normalised phase congruency measure $E(t)$ is weighted by the modulus over all voices at t

$$PC(t) = \frac{E(t)}{\sum_{n}^{N} A_s(t,n)}. \tag{24}$$

Phase congruency in the single dimension rhythm frequency case can be considered to be summing the contributions of all periodicities present at each time point, taking into account their phase and amplitude.

Phase congruency is therefore a measure of the synchronisation between oscillations of rhythmic components. These components constitute different rhythm time periods. Where these periods are highly synchronised indicates the end and beginning of phrases over more than one temporal level. As noted in Chapter 2 and summarised in Section 2.4.4, the interaction of temporal levels indicates rhythmic structure. Therefore high synchrony—high phase congruency—may be interpreted as points in time of high structural importance.

This is clearly coincident with the stratification approach of Yeston [192], phase congruency being a measure of Yeston's rhythmic consonance of phase between strata. Phase congruency however does not distinguish between strata dissonance from a phase shift and dissonance from difference in period. Phase congruency can also be seen as a computational approach to Lerdahl and Jackendoff's GTTM metrical decomposition [84]. Desain's summation, at each time point, of temporally forward projected expectancy measures over a limited range of time scales is another example of multiresolution rhythmic analysis which forms an analogue of phase congruency [23]. Parncutt's meter salience hypothesis also appears to be very close to phase congruency, if his pulse sensations are related to the responses of wavelet voices to analysed rhythms:

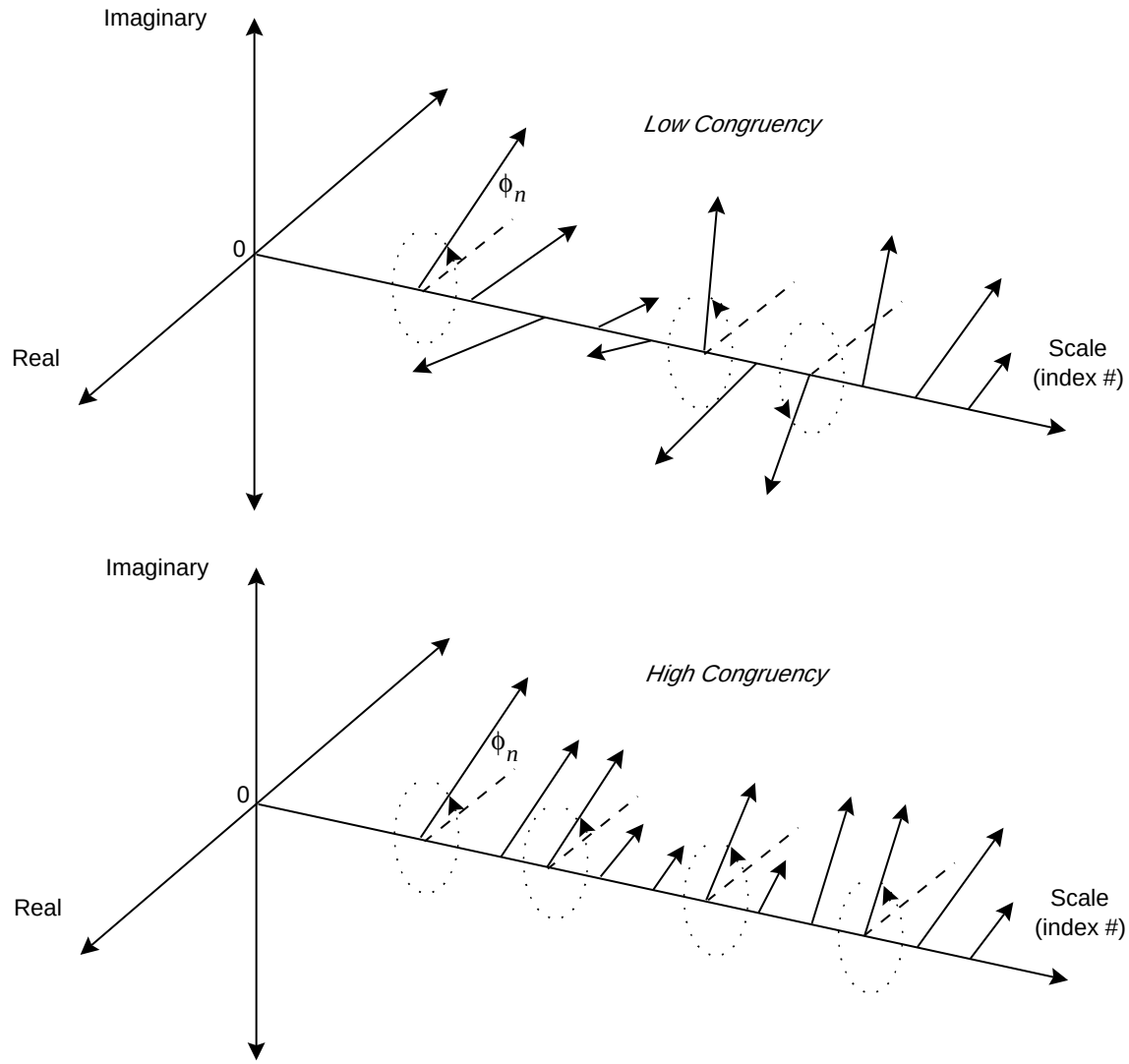

Figure 19: Phase congruency is the measure of angular alignment of all voices at each time point of the analysis. The diagram demonstrates the phase measures for all voices at a single time point.

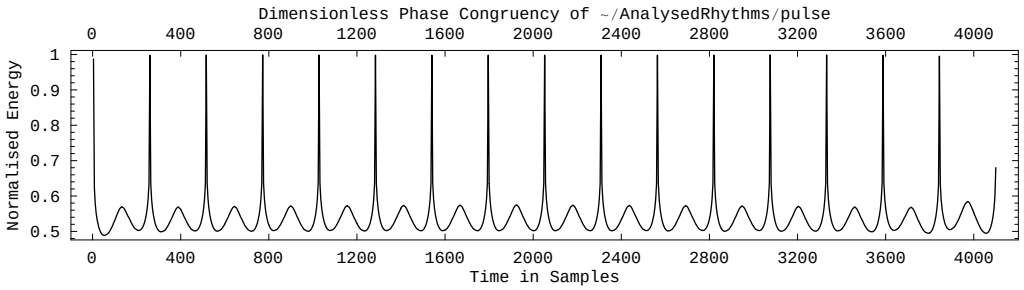

Figure 20: Phase congruency of the isochronous beat pulse train of Figure 15.

> "The salience of a perceived meter, or the probability of a given metrical interpretation, is proportional to the sum (or some other aggregate) of the salience of the pulse sensations that make it up. The most likely meter to be perceived is the one with the highest predicted salience." [130, pp. 443]

In an earlier paper, Parncutt describes qualitatively such an algorithm using direct summation of salience measures [128, pp. 135]. His pulse saliences are derived from a heuristic (his equation 4 [130, pp. 434]) based on phenomenal (i.e. objective) accent weighted by ordinal position within the pulse period. In contrast, phase congruency measures use the magnitude and phase of the frequency present in the entire rhythmic signal.

A similar conception to phase congruency has been taken by Todd [103] with respect to Marr's primal sketch theory of human vision [95]. As Todds filters are causal, a computational measure of his "temporal coincidence" across filters is not possible, instead resorting to qualitatively assessing "temporal contiguity" visually, using scatterplots of filter peak responses over the time-frequency plane.

The phase congruency measure of the isochronous beats (impulses) from Figure 15 is demonstrated in Figure 20. As noted in section 3.3.3, the local energy and phase congruency functions will indicate points where the impulses fall due to singularities having a high modulus at high frequency scales and a constant phase across frequency scales. Each impulse point achieves the same relative phase congruency measure; consistent with the isochronous (i.e. indistinguishable) nature of the beats. The slight hump between each impulse is from the second harmonic modulus.

3.5 Summary

The description of the traditional Fourier transform frequency analysis has been presented to indicate the extent of its analysis abilities. Unless the signal under analysis is periodic with respect to its analysis window, any change in the signal's frequency will be distributed across the harmonic components of the signal.

In order to analyse musical rhythm in terms of periodicities, a perspective of the rhythm independent of the acoustic component of the signal has then been detailed. Here the rhythm of a musical recording has been argued to be able to be represented as the amplitude modulation of changing frequencies in the acoustic range. The frequency characteristics of this amplitude modulation is the signal that must be analysed (not the acoustic carrier) to reveal the time dependent nature of musical rhythm.

Rectification of the amplitude modulation can be used to separate the rhythm from the acoustic element. Alternatively in prepared performance situations, the rhythm can be transduced from sensors, measuring the intensity of the instant of the beat, which is generalised as a measure of intended accent. Intention has been defined to refer to the conceptual beat structure prior to making those beats audible (by using it to modulate the acoustic carrier). This leads to the confirmation of the view that a train of impulses can represent the rhythm in signal processing terms.

The theory of non-orthogonal wavelet transforms has then been reviewed, with regard to its applicability to rhythm analysis. The wavelet transform enables change in frequency to be represented in 2-D plots of dilation scale and time, separately indicating magnitude and phase components. This enables rhythm to be formally viewed in terms of the frequency domain. Additionally, the magnitude and phase representations enable computation of a measure of phase congruency at each time point.

This is a new application of phase congruency and a new approach to rhythm analysis, unifying signal analysis and several music psychology theories. The approach has been demonstrated by analysing the primary case of an isochronous impulse train. The next chapter assesses the validity of such a multiresolution representation by analysing and interpreting a variety of musical rhythms.

Chapter 4

Analysis of a Corpus of Musical Rhythm Examples

Morlet wavelets were described in Chapter 3. This chapter examines their application to the analysis of musical rhythm and shows their usefulness on a range of musically typical examples.

Firstly, examples of simple rhythms created from dynamic and durational accents exhibiting changing meters are analysed. Rhythms undergoing asymmetrical ritardando and accelerando and using agogic accentuation are then demonstrated. The ability of the CWT analysis to display grouping of rhythms is then demonstrated on an anapestic rhythm. I then assess the degree to which deviations from strict quantization of a complex rhythm are revealed. A well known complex rhythm is then analysed, first in terms of its quantized, notated, rhythmic values, then analysing a performance of the same rhythm, preserving the rubato. This chapter is an expansion of examples of analysis of rhythms reported previously [162, 163, 165, 161].

The examples analysed are monophonic rhythms. The term monophonic is used here to mean a single rhythmic line, i.e as performed on a single pitch drum, or a rhythm tapped out on a tabletop. Even on such an impoverished sound generator, accentuation is still possible, e.g from intensity or timbre variation. Therefore the rhythms analysed are considered as a functional whole, exhibiting structure, including the possibility of syncopation through accenting, while being irreducible to parts played by different drums, limbs or performers.

The perception of musically typical rhythms is achieved by segregation of the received sound complex into separate streams of common sources [10]. It is thereby hypothesised that listeners use timbral, spatial localisation, pitch, tempo and other

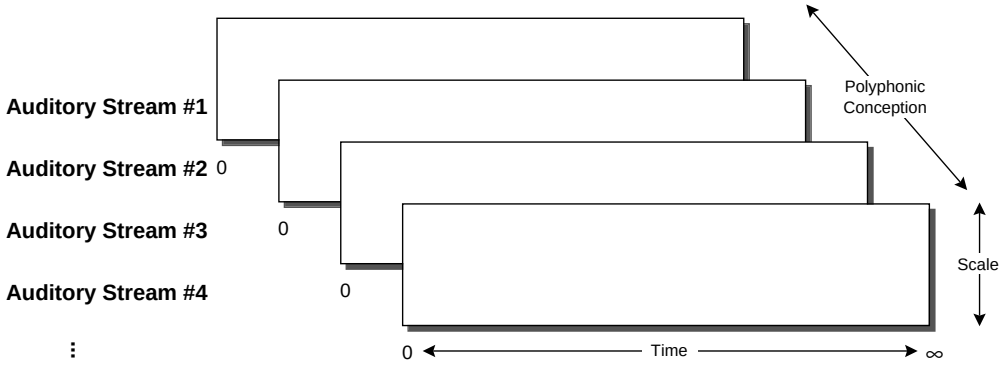

Figure 21: Polyphonic rhythms will segregate into parallel streams from objective differences between sources.

objective differences between sound sources to distinguish between independent rhythmic patterns. Thus, the process of perception of a polyphonic rhythm (for example a performance on a drumkit) would be represented by a number of wavelet analyses, one per line, running parallel in time. Where the listener can interpret a rhythm as comprising multiple rhythmic lines, rather than variations in accentuation of a single rhythm, this introduces two or more dimensions of independent rhythms. This concept is illustrated in Figure 21.

4.1 Implementation Details

The example rhythms presented in this chapter were either synthesised with the Common Music/Common Lisp Music software synthesiser [178, 179], or converted from MIDI files of performances on MIDI drumpads, into an impulse train. The impulse train is saved to a sound file at a sample rate of 200Hz, giving a minimum timing resolution of 5 msec. When sampling performance data (section 4.5.2), 400Hz was used, with 2.5 msec resolution being close to the limit possible with MIDI, regardless of sampling errors from machine load. The wavelet transformation typically extended over 10 octaves to a maximum wavelet wavelength (from its internal frequency ω_0 of Equation 14) of 2048 samples (10.24 seconds). The minimum wavelength was fixed at 4 samples (20 msec) as this still indicates impulses and represents most perceptible timing. Many of the test rhythms are drawn from other researcher's published examples, in order to allow comparison with other work, and

Rhythmic Value	IOI Ratio to ♪	Scale a	Dilation $2^{a/8}$
Quaver	1.0	0	1.0
		1	1.0905
Septolet Crochet (7 in time of 4)	1.1429	2	1.1892
Triplet/Sextolet Crochet (3 in 2/6 in 4)	1.3333	3	1.2968
		4	1.4142
Dotted Quaver	1.5	5	1.5422
Quintuplet Crochet (5 in 4)	1.6	6	1.6818
Alternative Septolet Crochet (7 in 6)	1.7143	7	1.8340
Crochet	2.0	8	2.0

Table 3: Musical rhythmic values, their relative ratio, and the degree of match to 8 voices per octave.

to begin to establish a standard rhythmic test corpus.[1]

Kronland-Martinet's analysis of musical sound [74] naturally suggested 12 voices per octave (one per equal tempered semitone). However 8 voices are sufficient to represent rhythmic variation possible within a doubling of beat frequency, within current computational capabilities. In a manner analogous to the approximation of equal temperament tuning systems to just intonation [189, 132], equal divisions of the rhythmic "octave" approximate low prime ratio time intervals typically considered within Western music theory [92]. As is apparent from their theoretical ubiquity, these "just" time ratios form the majority of the interval perception categories, so the 8 voice equal division approximations are appropriate.

Typical rhythmic values, their ratios and equal temperament approximations appear in Table 3. A special unequal wavelet dilation could be adopted, choosing scale parameters only to match expected ratios of rhythmic subdivisions. However, for this research, regularly spaced scales are used in order to indicate expressive timing which can deviate from the categorical rhythmic frequencies. Earlier plots of the figures in this chapter with 16 voices per octave did not significantly improve the interpretation of the scaleogram/phaseogram pairs.

The Morlet wavelet transforms were implemented initally in C and subsequently in Octave, a public domain Matlab[2] workalike. From a sound file of impulses, these

[1]DORYS — a Database of RhYthmic Stimuli is available from `http://www.cs.uwa.edu.au/~leigh/Research/Software/DORYS.tar.gz`.

[2]The mathematical language Octave can be found at `http://bevo.che.wisc.edu/octave/`. Matlab, Mathematica and Postscript are registered trademarks of their respective owners.

programs produce three output files for the magnitude, phase and phase congruency. Encapsulated Postscript plots of the two and a half dimensional scalograms and phasograms, and the one dimensional phase congruency plot were produced by Mathematica code. The C version computed the wavelets and the convolution in the time domain, while the Octave version constructed the wavelets in the Fourier domain and multiplied those with a precomputed FFT of the signal. This second version produces a run-time performance improvement due to the $n \log n$ time complexity of the multiplication and inverse FFT, but this requires the input data to be first padded at either end of the signal to a dyadic (2^n) length for the FFT. The padding consisted of reflected portions of the signal to create a repeated rhythm in order to preserve the low frequency strata from edge impulses. The displayed scalograms and phaseograms are then trimmed to their original length.

The abscissa of the scalogram plots time in samples, and the ordinate plots the frequency scale of the dilation of the wavelet in number of 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.

4.2 Generated Primitive Examples

Section 3.3.3 demonstrated the analysis of an isochronous pulse. In this section, other simple synthesised rhythms are analysed to further demonstrate the representational abilities of multiresolution analysis on typical rhythmic entities.

4.2.1 Changing Meters with Dynamic and Durational Accents

Intensity Accents and Changing Meters

Figure 22 illustrates a CWT of a rhythm composed of meters changing using accented downbeats. The IOI's remain equal across the pulse train (70 samples), only the beat that is accented is changed. Care was taken to ensure the rhythm does not change exactly in the middle of the analysis window, to avoid any artifacts arising from a signal symmetry within the window. The figure shows that a ridge

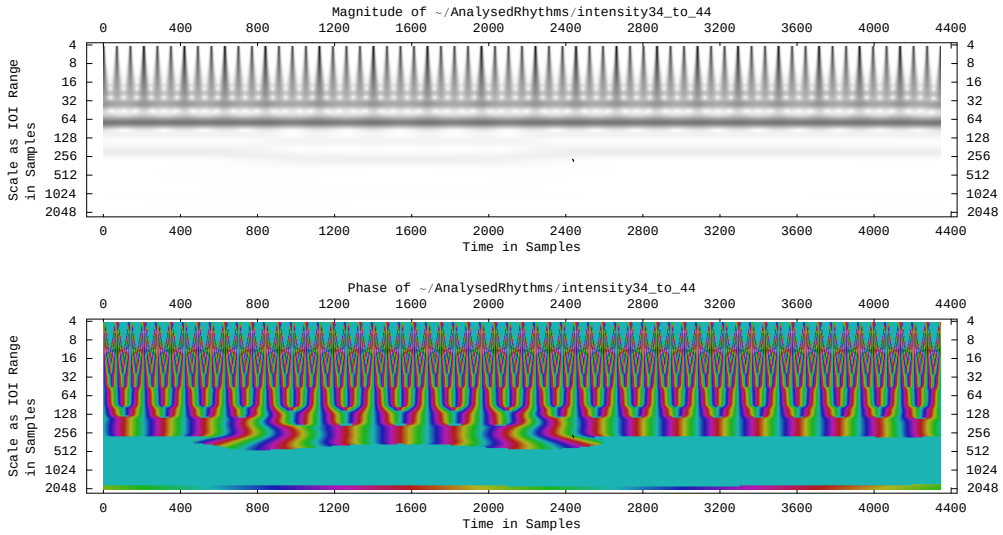

Figure 22: Scalogram and phasogram of a CWT of the rhythmic impulse function of a meter temporarily changing from $\frac{3}{4}$ to $\frac{4}{4}$. The time period of the meter transition is indicated by the change in the slight energy scale corresponding to the downbeat interval.

of frequency scales corresponding to the interval between the accented beats (210 samples) is established during the $\frac{3}{4}$ meter period, dips downwards for the $\frac{4}{4}$ (280 samples downbeat interval) and returns to the previous scale, demonstrating the zooming of the CWT and its ability to track a short term change in the frequency of the accented downbeats. The phasogram indicates congruence over ranges of scales corresponding to the rhythmic band.

The phase highlights the points in the signal, where a frequency (meter) change occurs. The phase oscillates at the lower scale during the $\frac{4}{4}$ region between beats 13 and 33. The non-causal nature of the convolution in the CWT pin-points the rhythmic alternation. A human listener can only *retrospectively* assign beats 13 and 33 as beginning points of a change in meter following contradiction of their expected downbeats. The non-causal CWT is imitating this behaviour.

The change in meter does not indicate a higher phase congruency measure at beats 13 and 33 as might be predicted. The phase congruency measure shown in Figure 23 is of a similar version of the rhythm, with significantly longer initial and following $\frac{3}{4}$ patterns, unequal in number to create analysis window asymmetry. The long $\frac{3}{4}$ patterns ensure the initial phase effects from the window edges do not

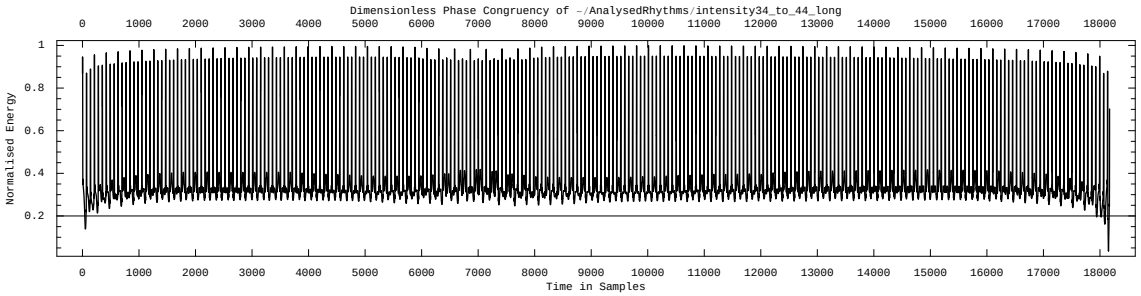

Figure 23: Phase congruency of the varying meter rhythm of Figure 22, extended with 30 bars of $\frac{3}{4}$ preceeding and 50 bars following. The $\frac{4}{4}$ region can be identified by the lower congruency values of the accented downbeats.

interfere with the congruency value. The meter change is identifiable by the *lower* congruency values for the accented downbeats of $\frac{4}{4}$ patterns. The congruency is being measured at each time point with respect to the entire signal.

Durational Accents and Changing Meters

To test the CWT behaviour on a durational accent, that is, a lengthening of the onset to offset time of each beat, an energy square wave per beat was used instead of a single impulse. Such an input signal will be the result following rectification of a sound file. A beat was formed with a 50% duty cycle, the accented beats extending the duty cycle to 71%, while intensity and IOI were held constant. The rhythm function input to the CWT is shown in Figure 24, again demonstrating a changing meter, from $\frac{4}{4}$ to $\frac{3}{4}$ and back. The transformed result shown in Figure 25 reflects the periodicity created by the accent at the downbeat of the measure. Here the rhythm frequency rises slightly at the point where the meter changes. Inspection reveals the frequency scale again corresponds to the IOI between the 71% duty cycle square waves. Substituting an amplitude modulating signal (as shown in Figure 5) for the square wave produces a similar scaleogram, and a phaseogram with much less high frequency phase variation, as is to be expected.

4.2.2 Ritardandi et Accelerandi

The ability of the wavelet transform to reveal a synthetically generated ritardando and a following accelerando is demonstrated in Figure 26. Every fourth beat is intensified (50% "louder" with respect to a normalised intensity scale) while the

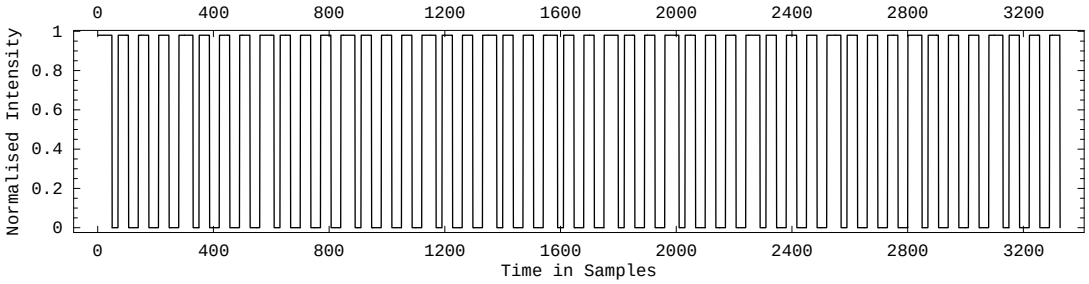

Figure 24: Plot of the rhythm energy square wave representation to be transformed with the CWT. Accents are created by lengthening the duration (the non-zero period) of the downbeat. The meter is temporarily changing from $\frac{4}{4}$ to $\frac{3}{4}$.

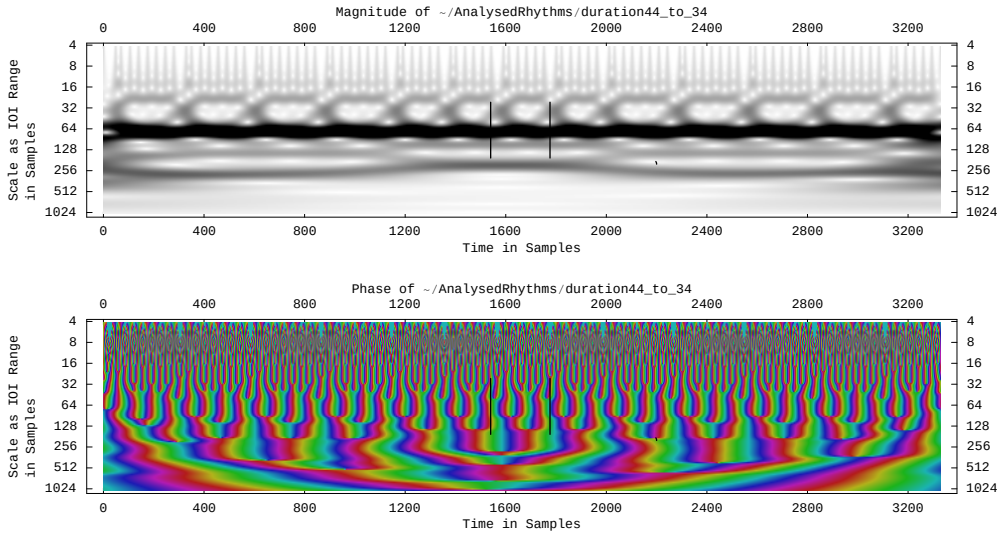

Figure 25: Scaleogram and phaseogram of the rhythmic energy square wave function shown in Figure 24. The rising frequency scale circled in the region of 1200–2000 samples (time axis) and an IOI range (y-axis) of 235 corresponds to the periodicity of the $\frac{3}{4}$ meter. This range is marked at samples 1541 and 1776 on the scaleogram.

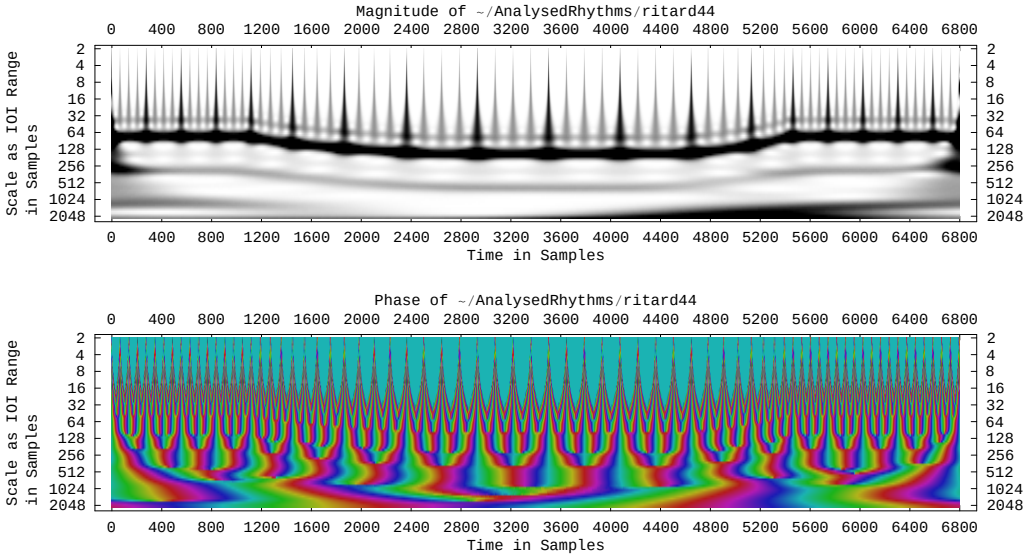

Figure 26: Time-Scale scalogram and phasogram display of a CWT of the rhythmic impulse function of a ritarding and then accelerating rhythm.

tempo begins at the 16th beat to slow linearly from 86 BPM to 42 BPM (at the 30th beat), holds at 42 BPM for 14 beats, and returns to 86 BPM by the 52nd beat. Frequency scales at the downbeat IOI (initially 280 samples interval) are mildly indicated in the scalogram and are more obvious in the phasogram at the same scales. This example demonstrates the tracking of a changing frequency due to the zooming nature of the scaled wavelet functions and the ability to discriminate frequencies created by regular, sparsely located, accented beats simultaneously changing over time.

The identical ritard/accelerate behaviour is demonstrated without intensity accents in Figure 27. The mild periodicity of the accented downbeat in Figure 26 does not occur, so only the highest magnitude corresponding to the IOI appears. The phaseogram pinpoints the events at which rate changes have occured, due to the non-causal nature of the wavelet.

The nature of a ritard has been the cause of some debate [32, 77]. A simple direct physical motion metaphor for a ritard is used here to demonstrate frequency tracking of the CWT, without assuming that a performed ritard will have such a frequency modulation. It remains a future research task to determine if the CWT will accurately reveal ritard patterns used by performers.

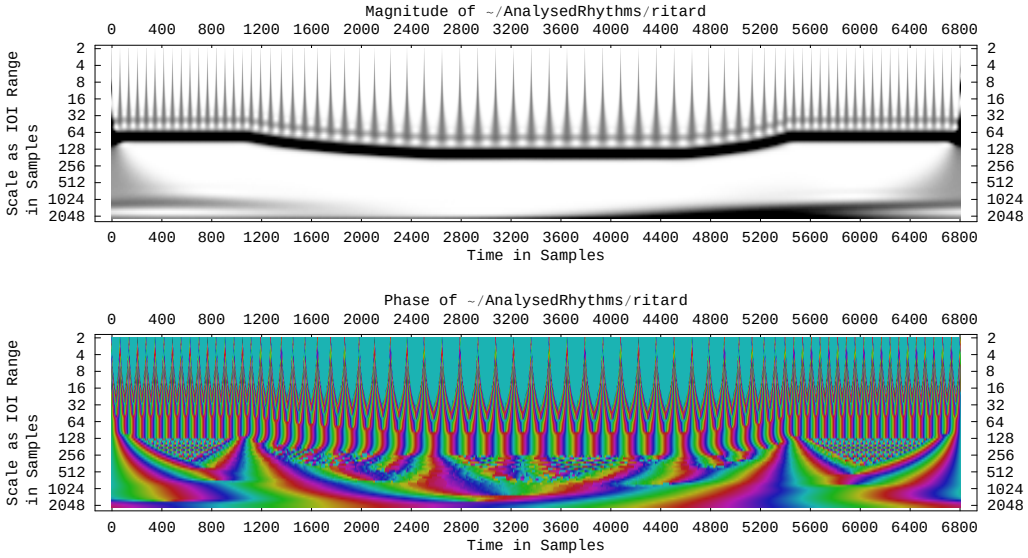

Figure 27: The same ritard-then-accelerate rhythm of Figure 26 without intensity accents, all impulses are the same weight.

4.2.3 Agogics

To test the wavelet analysis behaviour on a rhythm exhibiting agogic accent, the same ritard and then accelerate tempo curve in section 4.2.2 is applied to a quaver pulse, with a deviation away from isochrony every fourth beat (see Figure 28). This perturbation consists of an additional delay of 6% from the isochronous beat. Two different forms of agogic accentuation are demonstrated here. In the first, the degree of agogic deviation is scaled to be dependent on the tempo curve, i.e *relational invariance*, according with the proposal by Repp [142]. In the second, the event is shifted by a fixed amount independent of tempo in a similar manner to the work of Bilmes [6, 7, 8].

In either case, the example deviation used here is audibly quite pronounced, bordering on breaking the perceived periodic nature of the rhythm. Informal listening tests of the two rhythms by the author favour the dependent model. However, there is some debate against Repp's model from Desain and Honing as to whether there is a direct proportionate relationship between IOI and tempo [30]. The studies by Repp, and Desain and Honing, concerned complete pieces performed at different tempos, whereas the rubato presented here is far more local and varying over the phrase.

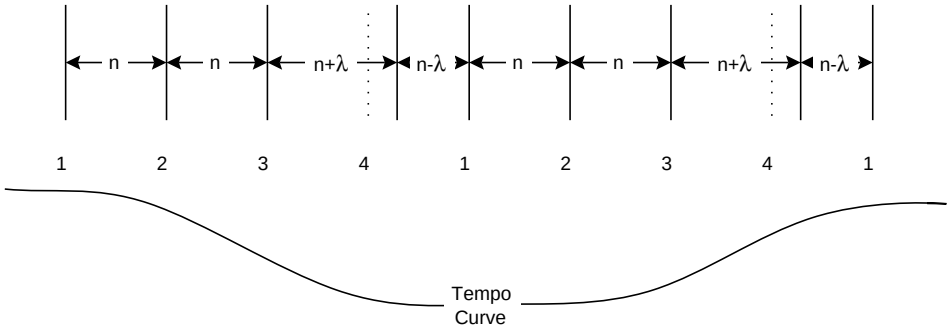

Figure 28: Implementation of agogic accent. The isochronous IOI is notated by $n = 350$ msec, the agogic deviation by $\lambda = 20$ msec, with a correction applied on the next IOI to avoid accumulation destroying longer term rhythmic structure. This canonical rhythm then has a tempo deviation applied to it.

In Figure 29, dependent agogic accenting is identifiable by the clear regular frequency modulation of the IOI ridge. However there does not appear an identifiable lower frequency activated scale corresponding to the rate of repetition of the agogic accent in a manner similar to intensity or duration accenting of section 4.2.1. Shifting by smaller ratios resulted in less ridge modulation, but still did not produce an accent period. In comparison, Figure 30 shows agogic accenting independent of the tempo curve, having noticeably less prominent modulations during the slower tempo region. While it is tempting to propose that this effect is caused by a multiresolution process mediating rhythm perception, a more logical explanation is that the perturbation is maintained at a similar ratio with respect to the underlying beat IOI. The scalogram's logarithmic dilation makes this effect clear.

4.3 Grouping of an Anapestic Rhythm

Wavelets not only give a reference to the periodicies from accent, but also the rhythm from grouping over longer time domains. A simple example of grouping is an anapestic rhythm.[3] Analysis of a similar rhythm has been demonstrated by Todd [103, (example A, figure 9, pp. 46)]. Examining the results in Figure 31 reveals two high energy scales, the highest corresponding to the beat IOI of 49 samples as expected, and the lower corresponding to the repeated period of the short-short-long group

[3]A repeated three beat rhythm, short-short-long ♪ ♪ ♩.

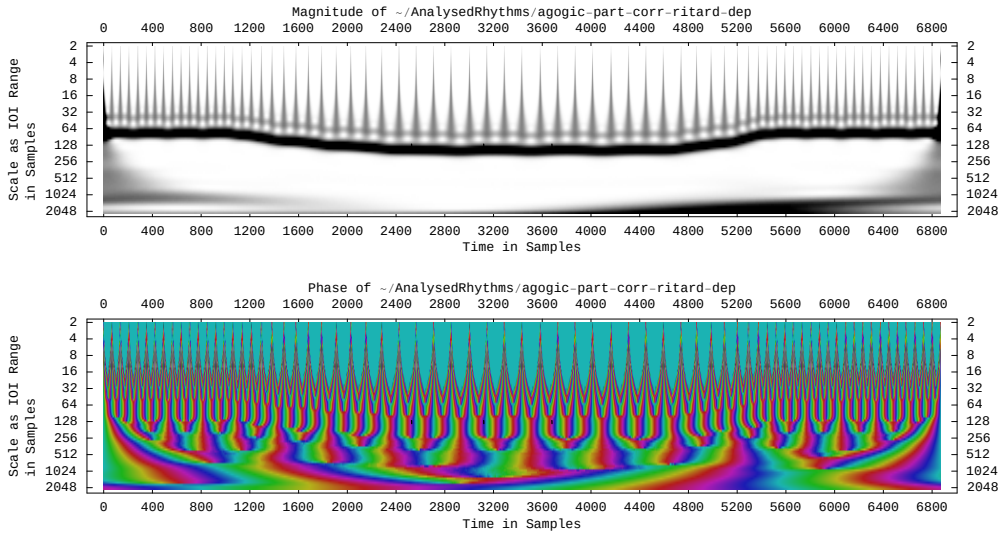

Figure 29: CWT of a rhythm composed of the same ritard-then-accelerate behaviour ("tempo curve") of Figure 27, with an agogic accent stretching every fourth beat by 20 msec, then applying the rubato, such that the final agogic deviation is rubato dependent and typically more than 20 msec. Three of the agogic accents are circled.

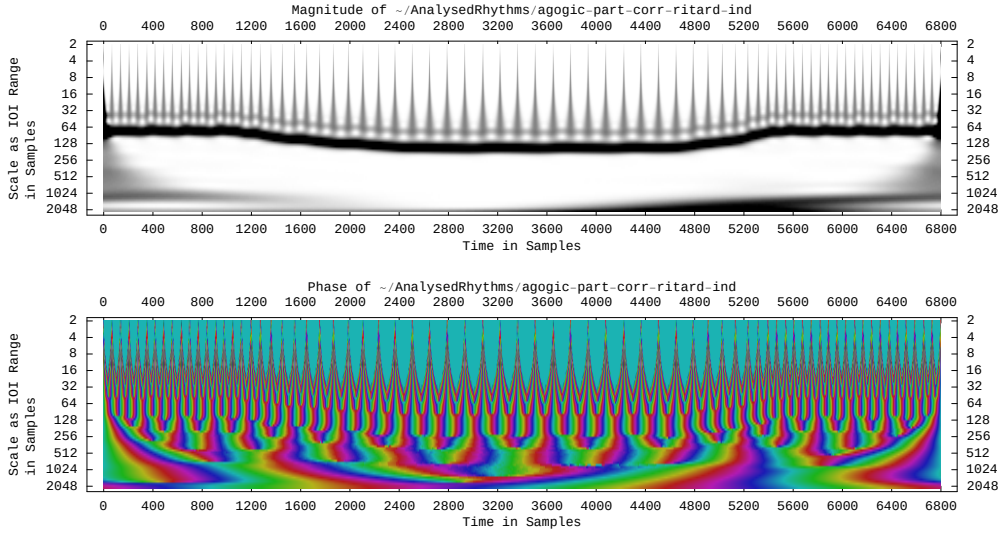

Figure 30: CWT of the same rubato rhythm as Figure 30 with an agogic accent stretching every fourth beat by absolute 20 msec, after rubato has been applied.

(198 samples as marked on the figure and expected). The most activated lower frequency scale extending over the analysis window corresponds to that anapest period. Due to estimation error in determining period from frequency scale, the computed period is marked on the scalogram at samples 1000 and 1215. Higher frequency sampling rates will reduce this reconstruction error. Other compound rhythms such as $\frac{7}{8}$ similarly are indicated by scales corresponding to the component periods (i.e 2 and 3) and the entire group period.

Due to the shift invariance of the CWT, if the rhythm is begun with the long beat leading the rhythm, forming a dactyl (long-short-short), the scalogram result will be identical, simply shifted in time. This is contrary to listener experience and music theory. For the short repetitions presented, the listener would remember the phase of the rhythm and group appropriately, with the caveat that Parncutt has reported wide deviation in listeners' beat phase preferences [130]. Clearly the CWT is not emulating the formation of higher order groups, simply indicating quasiperiodic repetitions of events. Construction of perceptually based grouping structures using a time-frequency representation is addressed in Chapter 5.

The distinctive phase measures at the beginning and end of Figure 31 are an artifact from the edges of the analysis window, effectively showing the periodicity of the entire analysis sample. The scalogram shows the energy measures of this artifact is low. Note that this differs from the signal window periodicity of the STFT, which is assuming the signal is entirely harmonic to its analysis window. The phaseogram indicates the periodicities of the components by the repeating progression through the colour spectrum.

4.4 Expressive Timing

4.4.1 Comparison of Performed and Quantized Versions of a Rhythm

To evaluate the performance of the CWT on a rhythm exhibiting expressive timing, an example previously used by Desain and Honing with their connectionist quantizer [25, pp. 154] was adopted (Figure 32). The timing data shown in Table 4 assigned to `desain-unquantized-rhythm` was normalised (by dividing by the first value, as per Smoliar's approach [167]), converted into a pulse train, and analysed as shown in Figure 33. The quantized version `desain-quantized-rhythm` is shown

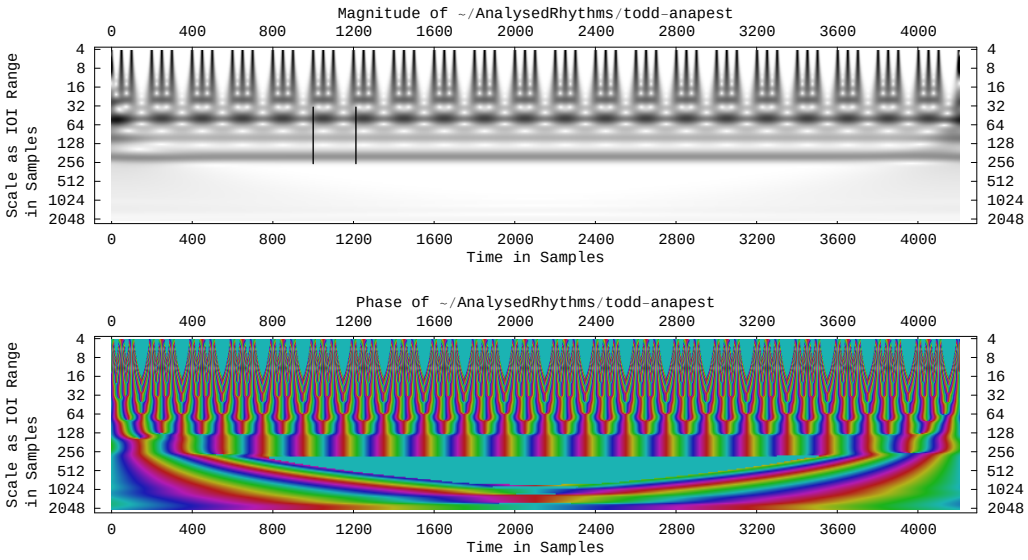

Figure 31: Analysis of an example of an anapestic rhythm. The anapest motive (short, short, long) period in the test rhythm was 200 samples. The computed period is marked on the scalogram at samples 1000 and 1215.

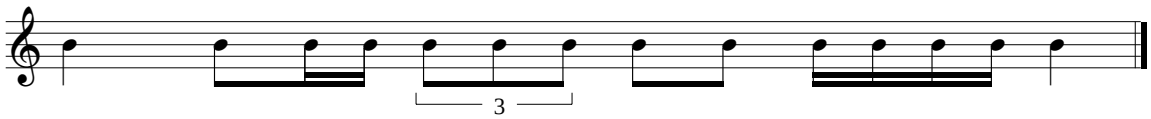

Figure 32: Desain and Honing's Connectionist Quantizer rhythm.

resulting from the standard 20 iterations of Desain and Honing's Micro Connectionist Quantizer as it has the tendency to not totally converge to integer ratios. This rhythm was also normalised and converted to an impulse train, resulting in the CWT magnitude and phase diagrams of Figure 34.

The magnitude display of the unquantized data demonstrates noticable bends in ridges between 300 and 600 samples, slowing to a local minima — half-way between the last triplet quaver and the first quaver — this can be seen as a graphic display of the "shaping" of a rubato. This beat pattern (three triplet quavers followed by two quavers) in the quantized data has a more dramatic and rapid transition between the two rates, without the bend towards a halfway point. Likewise, the effect of phrase final lengthening (Section 2.2.7) is apparent on the unquantized rhythm in the slight

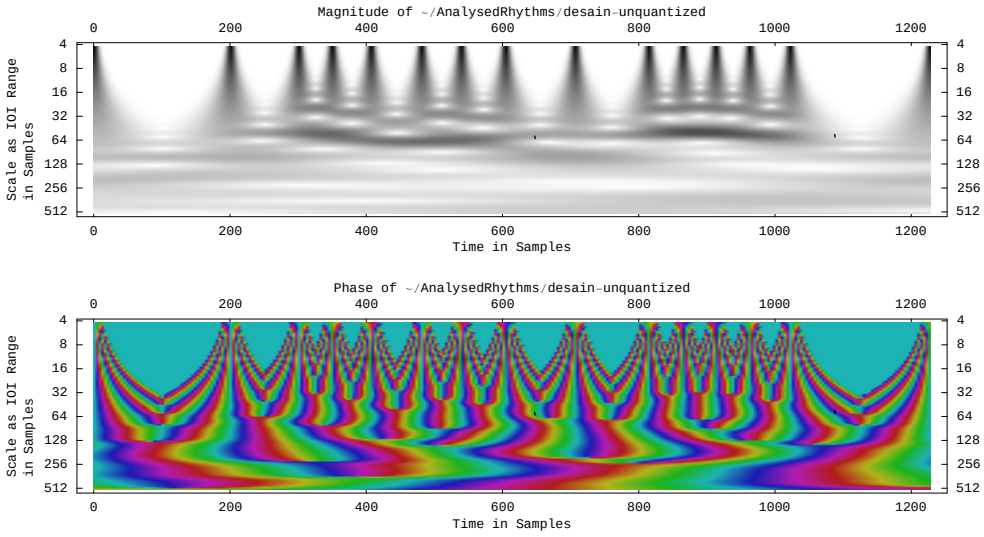

Figure 33: The scaleogram and phaseogram results of the unquantized data in Table 4 without intensity accents, all impulses are the same weight. The triplet to duplet quaver transition is circled, as well as the slowing of the final semiquaver.

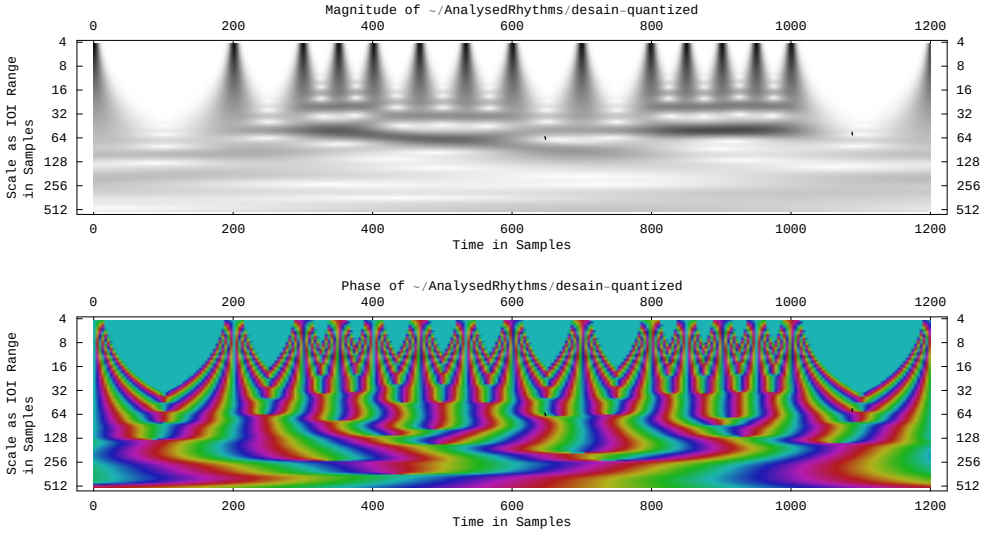

Figure 34: The scaleogram and phaseogram results of the quantized data in Table 4. The triplet to duplet transition is an immediate change from beat 5 to beat 6. The phrase final lengthening has been removed by the quantization.

```
(setf desain-unquantized-rhythm
  '(11.77 5.92 2.88 3.37 4.36 3.37 3.87 6.0 6.34 2.96 2.8 2.96 3.46 11.93))

(setf desain-quantized-rhythm
  '(12.02 6.00 3.03 3.02 3.98 3.97 3.98 5.97 5.96 3.07 3.08 2.95 2.98 11.99))

;;; normalising by the first interval indicates Smoliar's criticism
(defun normalise-rhythm (rhythm datum)
      (mapcar #'(lambda (n) (/ n datum)) rhythm))

;;; Make a Common-Music thread out of the IOI's
(thread desain-unquantized ()
        (dolist (beat (normalise-rhythm desain-unquantized-rhythm
                                        (first desain-unquantized-rhythm)))
          (object rhythm-onset note 'c4 rhythm beat duration beat
                  amplitude 0.95)))
```

Table 4: Common Music versions of the original input data used by Desain and Honing [25, p.167] for their quantizer and the quantized version following a run of their program.

dip of the scales corresponding to the IOI between the penultimate semiquaver and the final crochet.

The shaping of the triplet to duplet transition is also apparent comparing the phase congruency measures of the two versions in Figure 35. The equal intervals of the quantized version produce distinct ranges of equal congruency of phase over the 300–400 sample region and the 450–600 sample region. Since the wavelet is non-causal, the congruency measure is computed from phases derived from intervals forward and and backward in time from each translation sample. Therefore, the first triplet quaver (peak 5) is of the same phase congruency as the earlier adjacent semiquaver (peak 4), while the congruency is common at the second (peak 6) and third (peak 7) triplet quaver and the following quaver (peak 8). Whereas, the unquantized version's phase congruency measure is closely matched across nearly all of the beats. Although such general observations can be made regarding the phase congruency measures, in general it has been difficult to draw associations between phase congruency and the underlying musical structure that the rhythms embody.

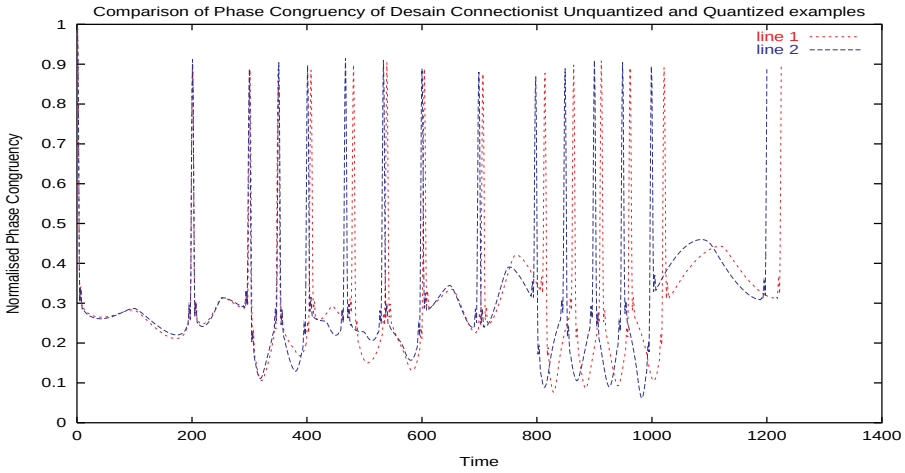

Figure 35: Comparison between the phase congruency measure of the unquantized (line 1, analysed in Figure 33) and quantized (line 2, analysed in Figure 34) versions of Desain and Honings rhythm.

4.4.2 Analysing Rubato Deformations of a Complex Rhythm

Figures 37, 38, 39 and 40 demonstrate the application of synthetic rubato to another of Desain and Honing's example rhythms, in this case, pre-quantized.[4] This rhythm, notated in Figure 36, establishes a high energy scale that corresponds to the dotted minim grouping from the duration of the motive. This is indicated in the relative energy profile taken at a representative time (one quarter of the window length) in Figure 38. The scale axis is marked with rhythmic values according to a tempo of 100 BPM, the original rate of the synthesised rhythm. It is significant to note that in Figure 37 the scale continues for the entire analysis period, since the grouping of the beats remains the same, even with variations in the rhythm.

The same rhythm is then perturbed with a complex tempo curve indicated in Table 5. The rhythm begins at 100 BPM as the quantized version, drops to 50 BPM on the fifth beat, accelerating to 80 BPM by the twelfth. It then jumps to 150 BPM and stays at that rate for the rest of the rhythm. The effect of this quite severe tempo curve on an isochronous crochet pulse is shown in Figure 39, and the effect on Desain and Honing's rhythm is demonstrated in Figure 40. This last figure reveals

[4]Taken from Desain and Honing's webpage `http://www.nici.kun.nl/mmm/SOUNDS/P31623.AU`, and also appearing in [31].

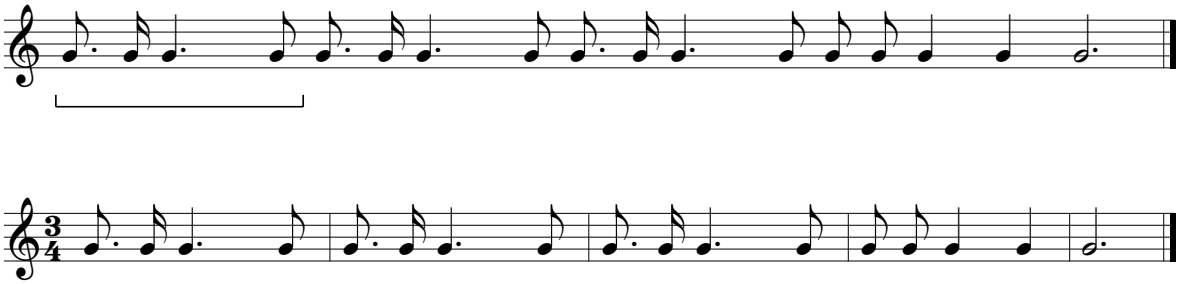

Figure 36: Desain and Honing's rhythm (top stave), producing a clear sense of grouping (lower stave). The functional group is underlined.

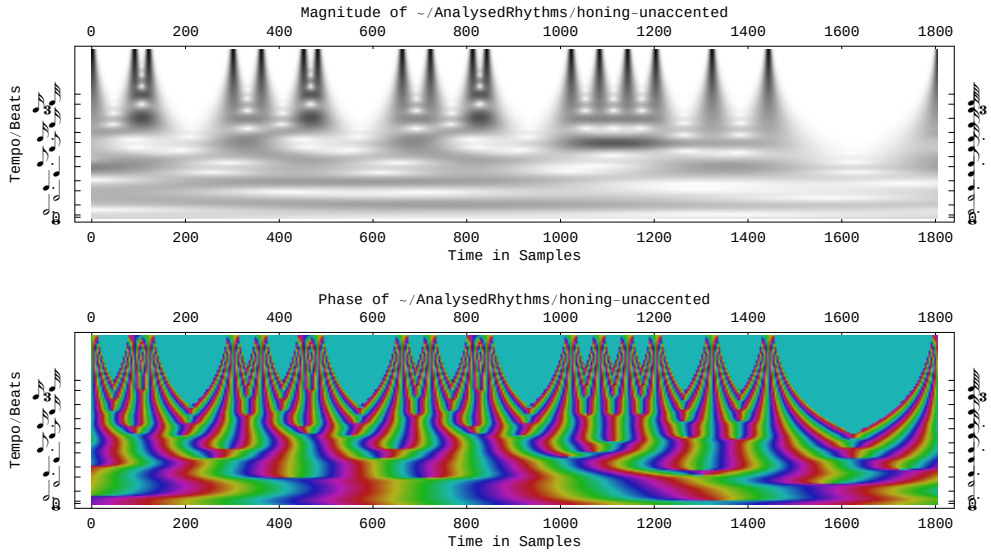

Figure 37: CWT analysis of the prequantized rhythm of Desain and Honing of Figure 36 without intensity accents.

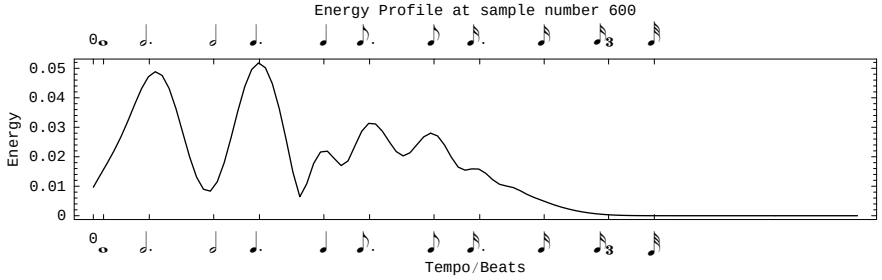

Figure 38: Demonstration of the activation energy distribution at the 600th sample time point and its concordance with grouping structures of the rhythm of Figure 36.

```
;; rubato perturbation to apply to Desain and Honings rhythm
(setf short-rubato-tempo-curve
      (tempo 0 100.0 3 100.0 5 50.0 12 80.0 13 150.0 17 150.0
             pulse 'q update after))
```

Table 5: Common Music version of the tempo curve applied to the rhythm of Figures 36 and 37.

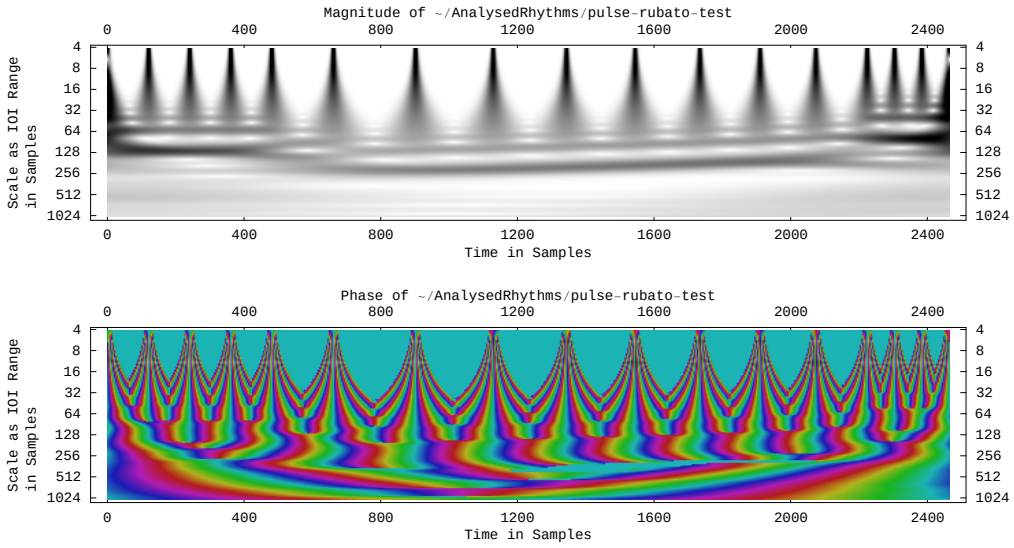

Figure 39: The tempo curve of Table 5 when applied to an isochronous crochet pulse.

the dotted minim rate deforms according to the tempo curve. A ridge of scales starts at the point of first rubato (approximately the 600th sample) corresponding to an interval of 535 samples and accelerates to an interval of 430 samples at the 2400th sample time.

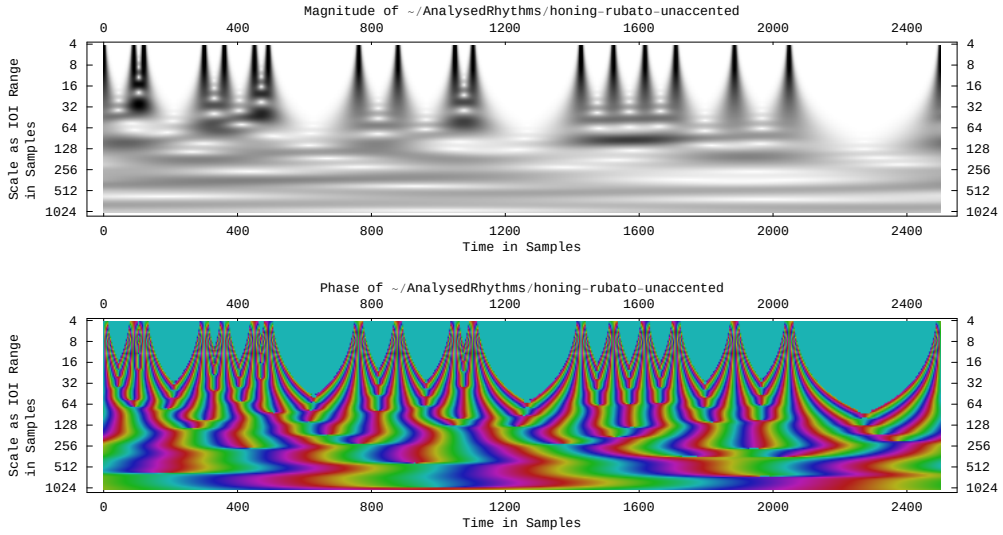

Figure 40: CWT analysis of the rhythm of Figure 37 after application of a synthetic rubato. The acceleration of the scale corresponding to the dotted minim is circled. In this figure, the scale axis is plotted in IOI spans as notation is only meaningful and can only be assigned according to a predetermined tempo, which in this case will be constantly changing.

4.5 Comparison of a Performed and Generated Musical Rhythm

4.5.1 Greensleeves

The previous rhythms have been short simple motives, chosen to independently demonstrate each effect. To demonstrate the scalability of the approach, the analysis is demonstrated on typical musical examples. In this section, a well-known rhythm example is analysed — "Greensleeves" (Figure 41, also used by Roberts [146, pp. 127]). It is composed of multiple IOIs grouped in musically typical proportions. Figure 42 demonstrates the rhythm synthesised with canonical IOIs directly from the notation. The scalogram indicates the hierarchy of frequencies implied at each time point due to the IOIs falling within each scaled kernel's support.

The scale corresponding to the period of the $\frac{\mathbf{6}}{\mathbf{8}}$ measure is most energised from the times of 0–2000 samples. Inspection of the activated scale corresponding to the dotted crochet in the time period of 2000 to 2800 samples shows this arises from a common interval of 300 samples which falls between the 17th and 19th beats,

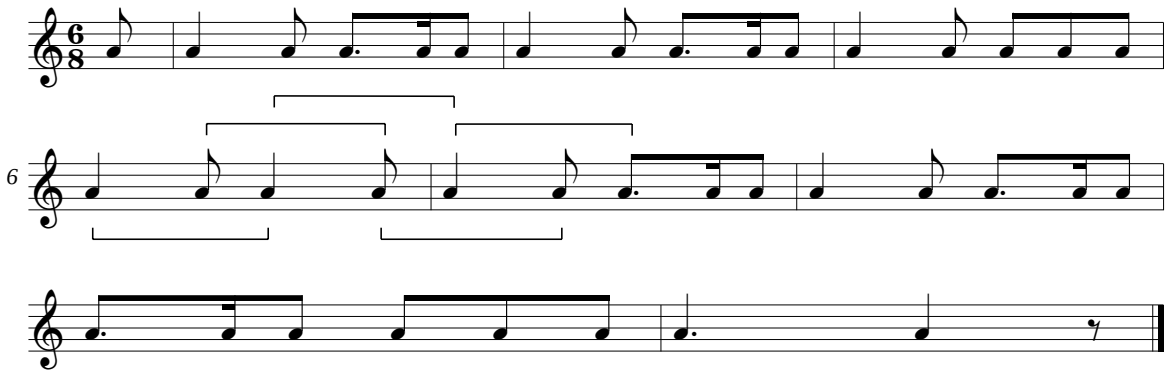

Figure 41: The rhythm of "Greensleeves". The interval of a dotted crochet appearing in the scaleogram of Figure 42 from samples 2000–2800 is shown in terms of the notation.

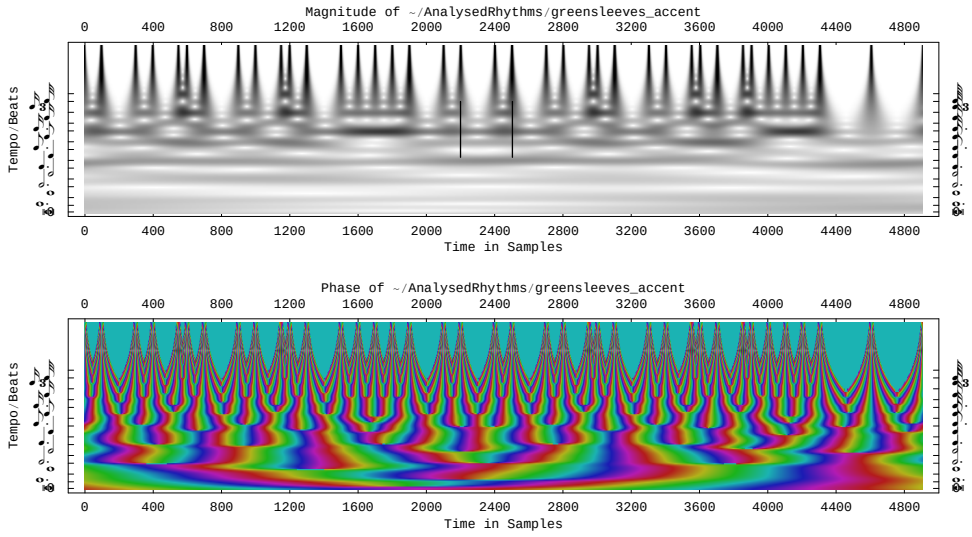

Figure 42: Magnitude and Phase of Greensleeves as notated with strictly rational IOIs. The period of the dotted crochet is shown aligned on the beat occuring on sample 2201, matching the next dotted crochet beat falling on sample 2501. The tempo of 60 BPM was chosen purposefully to create a dotted crochet IOI of 300 samples to aid finding the beats in the output sound file.

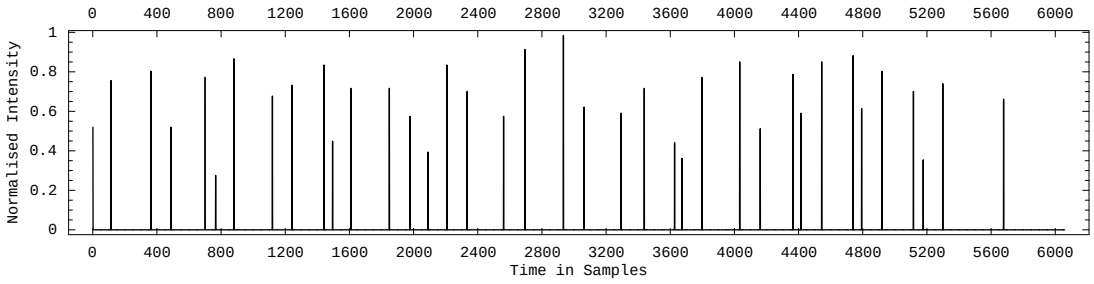

Figure 43: The impulse input produced from the velocity and timing parameters of a MIDI file recording of the rhythm being tapped on a drumpad without metronome. Note this rhythm uses 400Hz sampling rate, to reduce quantization of the performance, but is performed at a faster tempo than Figure 42.

18th and 20th beats, 19th and the 21st beats and so forth. With the return of the semiquaver/quaver figure (beats 24 and 25)), the scale is modulated in amplitude, but the recurring dotted crochet interval throughout the rest of the passage continues the scale activation. Of note is the wavelet's recognition of the overlapping dotted crochet intervals, independent of their metrical position, as notated in Figure 41.

The dotted crochet is, of course, half the duration of the $\frac{\mathbf{6}}{\mathbf{8}}$ meter the piece is notated in. Such a rhythmic figure would not normally be considered as a theoretical group. However, this interval is essential to establishing the expected theoretic grouping of the measure. The variety of periodic interpretations adopted by listeners such as in Parncutt's study [130] suggest listeners may simultaneous group at both rates, or as indicated here, change interpretation. The scales for IOI between consecutive beats also receive an energy distribution. Clearly the CWT is establishing close to the total feasible number of parallel strata capable of being evoked in the mind of the listener. In a human listening scenario, enculturated interpretative mechanisms are then bought into play in order to direct attention towards those strata that are most salient.

4.5.2 Greensleeves Performed

In comparison, in Figure 44, the CWT is demonstrated on an expressive performance of the rhythm of Greensleeves (Figure 43), tapped on a single drum-pad, without metronome at a medium tempo, roughly 96 BPM, converted from a standard MIDI file at 400Hz sampling rate as described in section 4.1.

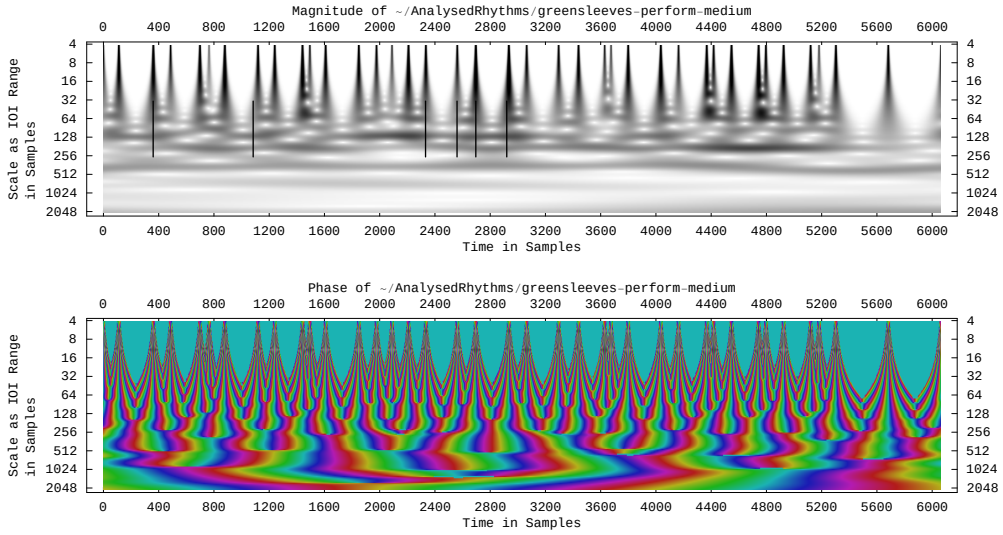

Figure 44: Resulting Scalogram and Phaseogram from Figure 43. Intervals corresponding to dotted crochet (362 samples) and dotted minim (724 samples) are marked.

Despite a different tempo, moderate rubato and significant variation in beat amplitude, features common to the canonical and performed versions can be discerned. The intervals of the dotted crochet and dotted minim appear again (362 and 724 samples respectively) and are marked on Figure 44 at samples 363, 1087, 2334, 2696, 2561, and 2923. There is a close but imperfect match on the dotted minim due to the rubato. The varying certainty of this period can be seen in the scale for 724 samples varying in energy across the analysis window. The half measure dotted crochet continues to be highly activated and extends the entire analysis window, with a slowing at the last two beats.

Inspection of the phase congruency (Figure 46), reveals the variation in the impulse weighting and the rubato produces less symmetrical events and therefore produces less overall phase congruency. Therefore the steadily increasing phase congruency measures for beats 13 to 17 seen in Figure 45 do not appear.

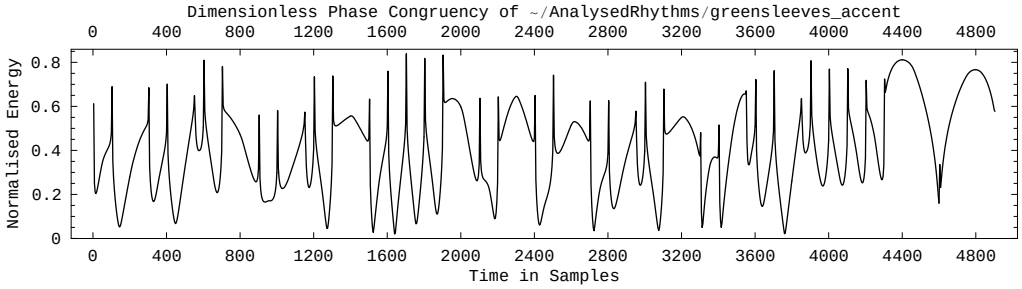

Figure 45: Phase Congruency plot of the rhythm analysed in Figure 42.

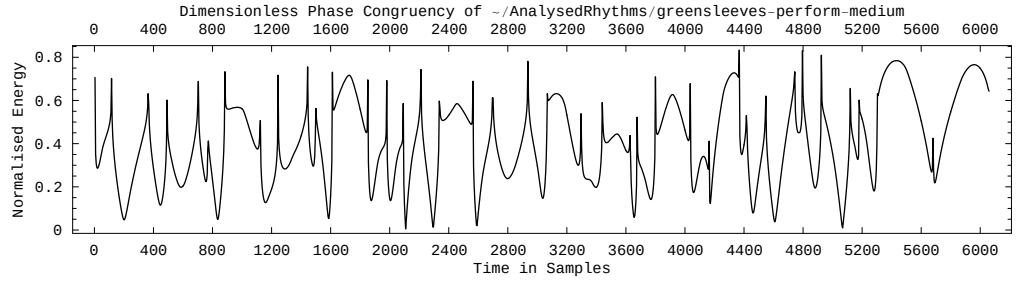

Figure 46: Phase Congruency plot of the rhythm analysed in Figure 44.

4.6 Summary of Results

Analysing the example rhythms in this chapter has demonstrated the illustrative abilities of a multiple resolution approach. The set of rhythms examined contains examples taken from previously published research and synthetic rhythms and is intended as a representative and reasonably complete sample. The synthetic rhythms have been chosen to independently illustrate rhythmic parameters described in Chapter 2. Analysis of this database of rhythms has indicated the degree to which musical knowledge and behaviour are made explicit by a time-frequency domain. Analysis has indicates the quantity of information inherent in the rhythm itself, prior to perceptual processing. This enables a systematic evaluation of a perceptual model of rhythm interpretation.

Multiple resolution analysis of rhythms represented as impulse trains undergoing fairly severe asymmetrical rubato has been shown to clearly track tempo deformations. This arises from the non-causal nature of the analysis and the zooming resolution of the CWT. Periodic accenting using duration or intensity — both increasing

signal energy at local time regions — is tracked by energy at scales matching the rate of accent, and correctly represents the effect under tempo modulation.

The local effect of agogic accent is apparent with the deviation of the scales corresponding to the IOI. However this deviation from metricality, when applied periodically, does not produce energy in scales corresponding to the rate of its repetition. The logarithmic scale representation makes agogic accenting appear more clearly when it is dependent on the (possibly changing) tempo of the rhythm. This tends to correspond to proportional scaling of timing with tempo as reported by other researchers [142, 30, 6].

Complex rhythms produce groupings that include those rates listeners would group on. However, there are other frequencies that correspond to component rhythms making up a compound rhythm, frequency doubled artifacts of the wavelet, or other rates corresponding to intervals between beats not attended to. This represents parallel multiple hypotheses that listeners construct in the process of rhythmic interpretation. The selection of salient strata from the multitude is clearly the next goal to be addressed.

The CWT demonstrates the effect of removing expressive timing, as shown by analysing a rhythm following quantization. The effect of such quantization is to "take the bends out of the tempo curve", effectively conforming the rhythm strata to a smaller set of scales without the degree of variation of the original.

Finally, applying the CWT to an accepted complete rhythm, has shown how multiple periodicities arise in a typical rhythm. Inspection of the results has shown that these strata do arise from intervals in the rhythm, taking into account overlapping intervals and representing simultaneous rhythmic strata. In that sense, the results represent listening with an almost infinite context, or memory, while human listeners will aim attention at the salient strata alone. Comparing a performed version has demonstrated the robustness of the analysis, allowing common temporal features to be identified and compared between the performed and synthetic versions of the rhythm.

Phase congruency has been shown to be a time-point measure of symmetry of the rhythm over multiple time scales. This can be used for assessing the relationship of a single beat, or group of beats, to the remainder of the signal. These congruency measures have been made over the full dilation scale range. While general features can be deduced from this phase congruency measure, it does not represent a direct interpretable measure of a listener's importance of the beat within its context.

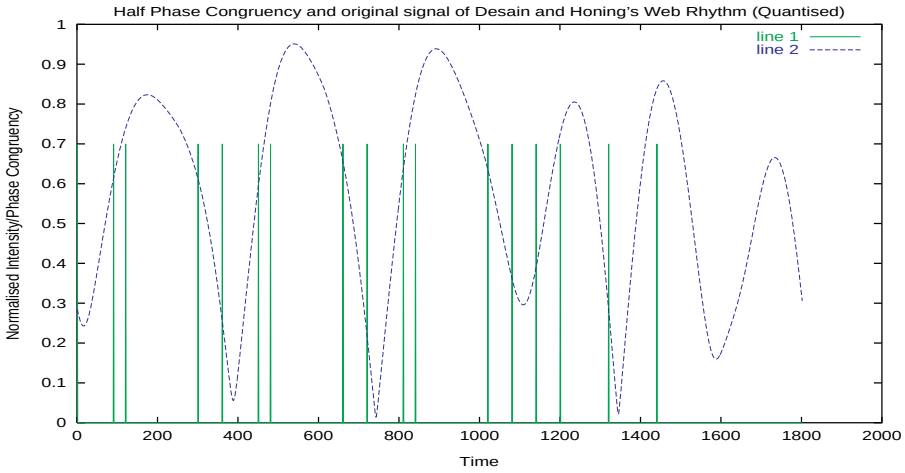

Figure 47: Phase congruency of Desain and Honing's rhythm in Figure 37 calculated over the reduced IOI range of 128 to 512 samples (two octaves), ignoring the IOI ranges which react to the impulse region.

While the idea of congruence of the phases of frequencies has appeal, the impulse points will always dominate the congruency measure, producing congruency peaks at each impulse. Even reducing the scales over which the phase congruency is calculated, only using scales with time intervals longer than rhythmic rates still produces a phase congruency measure heavily biased towards the impulses and the intervals following them. This is shown in Figure 47. A form of phase congruency measure based on those strata which are most salient to rhythm perception is required. One such form is *local phase congruency* which is detailed with other interpretative models selectively using time-frequency input in the next chapter.

Chapter 5

Interpretation of Time-Frequency Representations of Musical Rhythm

In the previous chapter, interpretations from the CWT transform were performed manually by the author. In this chapter several related approaches to using the time-frequency representation of rhythm for automatic interpretation are detailed and assessed. As a minimum demonstration of interpretation, the CWT is used to determine the tactus from several rhythms. This tactus is verified by using it to compute a foot-tap to accompany the original rhythm. This foot-tapping is demonstrated graphically. The intention in this chapter is to create a scenario to formalise and then test such interpretation hypotheses in accordance with the scientific method as close as possible. To that end, example rhythms are used where a clear tactus is already commonly agreed and can be compared against the computed version.

5.1 Tactus determination

In order to build and automatically interpret robust models of musical time (at least as understood within Western common practice conception), it is assumed that a fundamental requirement is the determination of the underlying tactus rate listeners will typically tap at. While notions of bell-lines as time-keeping references in non-western music may counter its universality (see section 2.2.5), the tactus can

be considered as an isochronous beat or pulse that is the conceptualised rhythmic backbone in western music.

The tactus is claimed here to function as the carrier in classical FM theory [119]. An isochronous beat is a "monochromatic" rhythmic signal, and the performer's rubato constitutes a frequency modulation of this idealised tactus frequency. Therefore, in performance the tactus of a rhythm is modulated but still elastically retains semi-periodic behaviour. A means of extracting the rubato frequency modulation of the tactus is required. This task is complicated by the fact that the rhythm, determined by the amplitude modulation function, only exists as onset time impulses. As detailed in Chapter 3, the rhythm of an acoustic signal is its amplitude modulation, typically extracted by rectification or by transducers sensing a performer's actions. In the ideal situation, the amplitude rectification will reveal the location in time of each single beat per energy burst. This is essentially a sharpening of the energy burst to a delta function (impulse) pin-pointing the event onset time.

Furthermore a number of interpretations of a particular rhythm are possible from the different beat rates that listeners have the option of tapping to. As shown in Chapter 4, wavelet analysis is capable of identifying many of the qualities of musical rhythm from an input signal consisting of time separated impulses. In particular, the multiple rhythmic strata are apparent. Within the context of signal processing, peaks in the scalogram/phasogram are termed ridges, representing the time behaviour of frequency components. A means of identifying and extracting the ridge that constitutes the tactus is required.

The approach adopted in this chapter is structured as follows: First, several forms of the multiple resolution technique of ridge extraction are described for determining the preferred rhythmic strata to tap to—the tactus—and its frequency modulation behaviour. These ridge extraction methods are stationary phase, modulus maxima and local phase congruency. Following explanation of these methods, principles applying to tactus are hypothesised and a simple algorithm implementing these principles is described. The ridge extraction methods and tactus algorithm are then tested on selected rhythms previously detailed in Chapter 4 and new examples. After assessing these, the tactus so extracted is demonstrated producing foot-tapping accompaniments and these are assessed. These steps are shown in the schematic diagram of Figure 48.

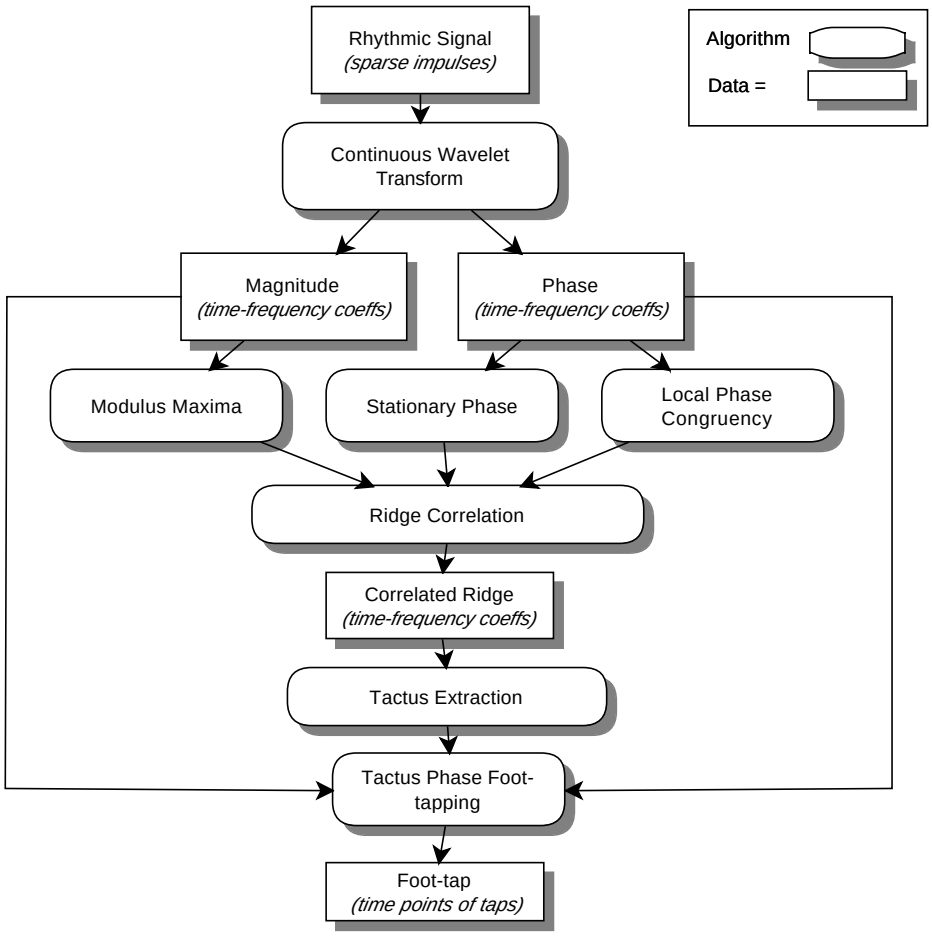

Figure 48: Schematic diagram of the multiresolution rhythm interpretation system.

5.2 Extraction of the Rubato Frequency Modulation Function

5.2.1 Review of Frequency Modulation Extraction from Ridges

A group of French researchers, Delprat [22, 21], Guillemain and Kronland-Martinet [53], Tchamitchian and Torresani [180], Escudié and others [37] (well summarized by Mallat [94, Chapter 4]) have used points of stationary phase to determine a *ridge* indicating the frequency modulation function of an acoustic signal. This determines the frequency variations over time of the fundamental and a finite number of partials.[1]

Analytic signals

Any real valued signal $s(t)$ can be represented as non-uniquely separated into amplitude and phase components [21, Equation 3.1]

$$s(t) = A_s(t)\cos(\phi_s(t)), A_s(t) \geq 0. \tag{25}$$

The analytic form [54] of $s(t)$,

$$Z_s(t) = u(t) + iv(t),$$

where $u(t)$ and $v(t)$ form a Hilbert Pair (Equation 17), produces a signal with only positive frequencies (Equation 18). Here of course $u(t) = s(t)$. However $Z_s(t)$ is completely characterised by the particular *canonical pair* [21] (A_s, ϕ_s), from Equation 19.

The real valued signal is recoverable from the analytic form since $s(t) = u(t) = \Re[Z_s(t)]$ produces Equation 25 from Equation 19.

Instantaneous frequency

The *analytic instantaneous frequency*, $\omega_s(t)$ of $s(t)$ is a theoretical notion[2] defined as a positive derivative of the signal's phase:

[1] Partials are considered the overtones or "harmonics" not neccessarily in strict harmonic ratio to the fundamental [119].

[2] See Hahn [54, pp. 516] for a discussion of the validity of a considering frequency at an instant in time.

$$\omega_s(t) = \frac{1}{2\pi}\phi'_s(t) \geq 0.$$

Instantaneous frequency will only be applicable to real-world signals that meet the condition of asymptotism. An asymptotic signal is defined as one whose phase changes significantly faster than amplitude with respect to time

$$|\phi'_s(t)| \gg \left|\frac{1}{A_s}A'_s(t)\right|. \tag{26}$$

Ridges

Two strategies to determine the ridge points are to either use a Gabor transform or its dilated version, the analytic wavelet transform. The wavelet has been defined by Morlet and Grossmann [52] and Kronland-Martinet [76] in the time domain previously by Equation 8 and Equation 14.

Points of stationary phase t_ς are defined such that

$$\phi'_s(t_\varsigma) = \frac{1}{a}\phi'_g(\frac{t_\varsigma - b}{a}). \tag{27}$$

In other words, the stationary phase points define where the rate of change of the phase of the signal (ϕ'_s) and the phase of the wavelet (ϕ'_g) (their instantaneous frequencies) are equal. The concept is illustrated in Figure 49. As the Morlet wavelet is analytic (Equation 16), the phase is independent and the wavelet itself can be represented as an asymptotic signal

$$g(t) = A_g(t)e^{i\phi_g(t)}. \tag{28}$$

The stationary phase points form a *ridge* where

$$t_\varsigma(b, a) = b, \tag{29}$$

that has the property of describing the frequency modulation function of the signal. The discretised version of the phase of the wavelet transform $\Psi(b, a) = \arg[W_s(b, a)]$ (the phasogram) can be recovered due to the nature of the near-analytic mother

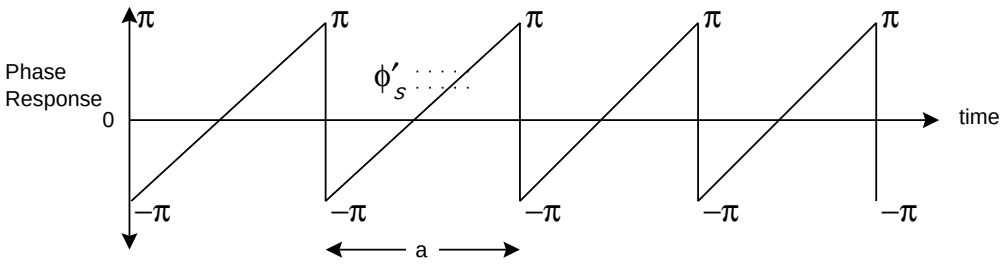

Figure 49: Representation of the stationary phase condition. The signal time interval between $-\pi$ to π transitions will match the wavelet time support on the ridge and indicate the signal frequency behaviour.

wavelet (Equation 14). The phase derivatives must be computed accounting for the $-\pi$ to π transitions, by adding back 2π at these time points. The phase allows forming a condition to determine the ridge:

$$\frac{\partial \Psi(b,a)}{\partial b} = \frac{\phi'_g(0)}{a}, \tag{30}$$

the right hand side being the frequency of the analysing wavelet at scale a. That is, when the signal instantaneous frequency equals the central frequency of the dilated wavelet. Delprat et. al [21][pp. 649] note the ridge of the wavelet transform can be extracted more precisely from the phase of the wavelet coefficients than the modulus maxima. They propose an iterative algorithm from Equation 30 for the ridges

$$a_{i+1}(t) = \frac{\omega_0}{D_b \cdot \Psi_{a_i}(t)}, \tag{31}$$

where D_b is a discrete differentiation operator, $\omega_0 = \phi'_g(0)$ from Equation 14 for a constant frequency mother-wavelet, a_i the scale at which the ridge occurs. The algorithm is deemed to have converged when the iteration change is insignificant [21][Eq 6.11, pp. 653].

A practical problem with this approach is comparing small value wavelet phase derivatives to small value signal phase derivatives, within some acceptable measure of equality with floating point hardware. For this research a more robust measure of the stationary phase ridge condition (Equation 30) has been adopted by comparing the reciprocal time support $2\pi/\phi'_s(t)$, corresponding to the instantaneous frequency

of the signal, and the dilated wavelet scale a respectively. An intuitive notion of equality is therefore when the time supports match within one sample difference

$$\left| \frac{2\pi}{D_b \cdot \Psi_{a_i}(t)} - a \right| < 1.0. \tag{32}$$

5.2.2 Application to Tactus Determination

The chief motivation of previous ridge research was to reduce the computation of the transform to only the ridges—the "skeleton", being the analytic version of the signal analysed—which can be achieved if the wavelet transform also meets the asymptotic condition [21, 53, 94]. In that application, the signal analysed was the sampled sound pressure profile. Additionally, signals composed of partials in close time-frequency proximity—within the Heisenberg bounds of the wavelet—create interaction between the partials ridges which prevents ridge separation [21, 94].

The motivation here is to extract the frequency modulation function for the purpose of determining a principal rhythmic partial that corresponds to the tactus. In this application, the rhythm signal is represented by sparse impulse points forming a critical sampling of the rectified amplitude envelope as described in Chapter 3.

Extraction of expressive timing $\gamma(t)$ from a performed rhythm $\kappa_p(t)$ is achieved by assuming the notated score rhythm (prior to performance) $\kappa_c(t)$ is a non-simple yet harmonically related carrier signal (possibly time-varying in its harmonic content) which undergoes a modulation by an inharmonically related expressive function

$$\kappa_p(t) = \gamma(t)\kappa_c(t).$$

Chapter 4 demonstrated the decomposition of a rhythmic signal into signal components corresponding to intervals between adjacent and non-adjacent beats. Due to categorical rhythm perception reviewed in Chapter 2, the typical intervals between beats will be close to integer subdivisions (2 or 3) of next shorter intervals. This creates a quasi-harmonic relationship between intervals. This harmonic relationship between onset intervals could be expected to ensure ridges do not interact significantly and that these ridges can be associated to the common modulation source $\gamma(t)$.

The condition of asymptotism described in section 5.2.1 will be compromised by the use of sparse impulses for the rhythm—effectively they are infinitely fast

amplitude envelopes. The degree to which this compromises the ability to extract worthwhile analysis data from musical rhythms has been a motivation of this research.

5.2.3 Modulus Maxima

In addition to extracting a frequency modulation function from the rhythm by restricting to points of stationary phase by the approach described above, ridges can also be determined from peaks in the modulus of the CWT with respect to the dilation scale axis

$$\frac{\partial |W_s(b,a)|}{\partial a} = 0, \tag{33}$$

when

$$\frac{\partial^2 |W_s(b,a)|}{\partial a^2} < 0. \tag{34}$$

Alternatively, modulus maxima with respect to the translation axis is defined by

$$\frac{\partial |W_s(b,a)|}{\partial b} = 0, \tag{35}$$

when

$$\frac{\partial^2 |W_s(b,a)|}{\partial b^2} < 0. \tag{36}$$

The dilation scale at each time instant will produce peaks due to the representation of the rhythm as impulse points, producing higher modulus points where wavelet impulse responses interact, as illustrated in the energy profile of Figure 17. Of the two modulus ridges methods, equation 33 more clearly reflects the nature of musical rhythm, finding the peak magnitude scales at each time point, demonstrating time-varying frequency components. In comparison, equation 35 finds the times of peak magnitudes for each scale.

5.2.4 Local Phase Congruency

The theory of phase congruency when applied to rhythmic signals was investigated in section 3.4. Phase congruency measures the match of phase angles between all scales for each time point. A new modified form—*local phase congruency*—is proposed here to have practical use in ridge determination. In this case, phase congruency over smaller consecutive scales indicates the presence of a frequency component in the rhythmic signal.

Due to the Gaussian shaped modulus profiles (see Figure 17), a wavelet transform of an isochronous pulse will activate several consecutive scales, with a common phase revolution (Figure 15). Therefore adjacent phase angles which are most in synchrony are indicative of a spectral component. Local phase congruency is therefore proposed as the "troughs" (local minima) of the absolute value of the first derivative of the phase with respect to scale. These are found by

$$\frac{\partial|\partial\Psi_s(b,a)|}{\partial a^2} = 0, \tag{37}$$

which finds the local extrema. The trough points along the scale a are found by

$$\frac{\partial^2|\partial\Psi_s(b,a)|}{\partial a^3} > 0, \tag{38}$$

given

$$\left|\frac{\partial\Psi_s(b,a)}{\partial a}\right| < \epsilon_p, \tag{39}$$

where $\epsilon_p = 0.05$ is an absolute threshold to ensure congruency measures are close to zero. The constant $\epsilon_m = 0.01$ for Equation 22 ensures congruency is only performed on scales and times where the phase is not ill-conditioned. This condition ensures the amplitude is large enough to allow phase to be calculated with some precision.

5.2.5 Combining Ridge Perspectives

As figures in Section 5.5 will demonstrate, no particular ridge method produces unbroken, unambiguous ridges across the analysis. Incomplete results from initial investigations with modulus maxima ridges formed the incentive to develop stationary phase and local phase congruency ridges as a means to disambiguate the ridges.

To that end, the combination of ridges derived from the three ridge methods aims to improve accuracy of the ridges in the time-frequency plane. This is essentially combining three perspectives originating from a single common source. It is notable that the non-causality of the Morlet wavelet enables the phase to add a degree of redundancy in ridge determination which is missing with real valued wavelets which lack an independent phase.

A number of different methods of correlating the ridge methods were attempted:

- Generating a ridge only where the three ridges methods coincided ("and" operation).
- Generating a ridge wherever proposed by a ridge method ("or" operation).
- Weighting contributions by each ridge method, such that the concurrence of both stationary phase and local phase congruency, or modulus maxima alone would suffice. Alternatively, any two of three would suffice.

As will be detailed in Section 5.5, no single combination method produced particularly better results. Finally ridge combinations were chosen manually for each rhythm in order to achieve best results. This therefore creates a future research agenda to devise a better method of combination. All results reported were achieved by correlating using an "or" operation between the modulus maxima and either local phase congruency or stationary phase or both. The "or" operation is justifiable in the sense it does not throw information away, but instead is overly inclusive of false ridges. Furthermore, all ridge methods are extracted from the phase and magnitude components of the same representation, so the "or" operation does not represent a divergence of perception between competing hypotheses.

5.3 Hypothesised Principles of Tactus

Because multiresolution analysis examines the whole signal over a contiguous range of dilation scales, relatively simple rules or principles can be formulated which have correspondance to music perception features described in Chapter 2. While axioms or principles may at first sight seem problematically reductionist, the principles proposed here extend over the entire rhythm, describing broad behaviours and matching intuition. This is not to propose the *necessity* of tactus, merely its sufficiency in

musical rhythms. Any rhythm can by decomposed into rhythmic strata, a subset of which, matching notions of metricality, can be analysed in terms of strata and a tactus interpreted from such.

First requirement is to formalise the concept of tactus continuity in order to construct an algorithm to identify which ridge of several candidates constitutes the tactus.

Principle 1 (Tactus Continuity)
The tactus perceived by the listener is a ridge that extends across the entire analysis window of the rhythm considered.

That tactus is proposed to be the most ubiquitious pulse over the entire analysis window. That is, the pulse rate which extends across the entire window length. This is created by the wavelet transform constructively summing time intervals between beats which reoccur in many overlapping relationships over the time period of the rhythm. As noted in Section 3.3.3, ridges will receive most contribution from the IOI, with minor contributions from intervals of a half or a third (double or triple the pulse rate).

Principle 2 ("Fundamental" Rhythmic Frequency of Tactus Ridge)
The tactus perceived is the lowest frequency ridge conforming to the principle of tactus continuity.

The requirement of the lowest frequency ridge avoids finding rhythms that are harmonics of a fundamental rate (half or third speed) tactus. Furthermore, only a (possibly modulating) tactus which extends beyond that of a beat interval is acceptable. Otherwise we are simply capturing a pulse, rather than the lower frequency rhythm. The lowest frequency continuous ridge allows for varying localised or additive rhythms, that combine to form more repetitive constructs over longer time periods, such as $\frac{5}{4}$ being formed from a group of two and a group of three.

Principle 3 (Tactus Modulation)
A performed tactus has a much slower modulation than the rate of the isochronous pulse (carrier). This reflects a preference towards constant repetition, with subtle variation generating longer term grouping.

This contrasts to classical definitions of tactus (such as Lerdahl and Jackendoff's GTTM [84]) which have been with respect to notated music. The tactus here is proposed as a simple pulse (longer in IOI than IOIs encountered in the rhythm, per

Principle 2) that is perceived by the listener during performance and is capable of tempo variation (modulation) according to the performance rubato.

That modulation will always extend over more than one period of the pulse. If the modulation was faster than the pulse this would be proposed to be the effect of a change of tactus, typically from a meter change. This would also be localised in time.

Principle 4 (Tactus Tempo Constraints)
A performed tactus has an upper and lower bound on its frequency.

As reviewed in Chapter 2 and itemised in the rhythmic periodic table (Table 2), the upper and lower bounds on perceivable intervals of time constrain the choice of tactus rate. The slowest rate can be proposed as an interval of 1800 msec, the fastest rate as 200 msec, per Section 2.4.2. These values will be modified by contextual effects, but should serve as typical boundaries. The influence of tempo could be considered to be a set of bounds on the range of dilation scales during analysis, or non-linearities in the dilated filter responses.

In using the CWT for rhythm analysis, the maximum wavelet period is the maximum retainable short term memory. The scale with longest time support is currently the theoretical limit of a quarter of the time period under analysis. This can result in a memory for rhythm that extends beyond human lower bounds. It is possible this may interfere with the ability to segment into groups due to an explosion of possible rhythmic parsings. It therefore seems necessary to match the longest wavelet to memory judgement limits from the psychological literature or experimental data.

Todd has proposed two peak receptive bandwidths in the rhythmic frequency spectrum corresponding to body sway (5000 msec interval), and foot-tapping (600 msec interval) that would strongly influence the listeners perception of tactus [104]. The foot-tapping rate is centered at favoured tempo rates, while the body sway rate is controversally derived from the vestibular system physiology [102].

Testing the Principles

These principles lead to a strategy for ridge extraction. It is hypothesised that the tactus can be found by determining the most complete ridge (apparent at all analysis times) through searching for maximal ridge continuity (i.e a tactus with least difference of scale numbers) using elimination of incomplete ridges.

This version did not apply tempo constraints to determining the tactus. While this limits the applicability to a wide range of tempos for rhythms, careful choice of examples which are close to typical tactus rates—Fraisses spontaneous tempo rate of 600 msec IOI—are assumed to negate effects of tempo preference. Principle 4 is therefore not scrutinised here.

Principle 1 does not allow for a tactus changing, for example, in the case of a meter change. It is however easy to imagine a relaxation of this principle with some perceptually derived tempo limitation to duration of tactus change. This would enable jumping from one ridge to another when it does not contradict the other principles and when that change preserves continuity from sample to sample in the sub-case.

5.4 A Greedy Algorithm for Tactus Extraction

Arguments have been presented in section 5.3 for the tactus being the ridge exhibiting the greatest degree of unbroken continuation across the analysed rhythm. An algorithm can now be formulated to determine a set of time-frequency points that lie on the most contiguous ridge across the analysed rhythm at the lowest frequency scales. This algorithm appears in Table 6. This uses a greedy-choice property [18, pp. 334] in its design and has similarities to a graphical flood-filling algorithm [39].

The parameter to the algorithm is an array containing the scales (the dilation indexes a) of all possible ridges (line 1). The initial most likely candidate for the tactus ridge is the highest scale ridge of the currently lowest frequency scales ($s(t)$) in the window (line 7). This solution is then expanded forward and backward from the time point of the highest scale *peakTime* (lines 8–15). This expansion extends along time points of scales lower or equal to the highest scale until there is a discontinuity ($d(n)$ in line 6). Three cases to be considered at discontinuity points are illustrated in Figure 50.

The first order difference of the scales are used to determine the continuity of the candidate. A minimum acceptable variation in scale between consecutive sample times (Δs) is used as a threshold condition. The ridge tracing algorithm produces prominent ridges, with $\Delta s = 2$ voices, on examples using 16 voices per octave. This is close to the theoretical single voice minimum $\Delta s = 2$, and all results reported here have used this value.

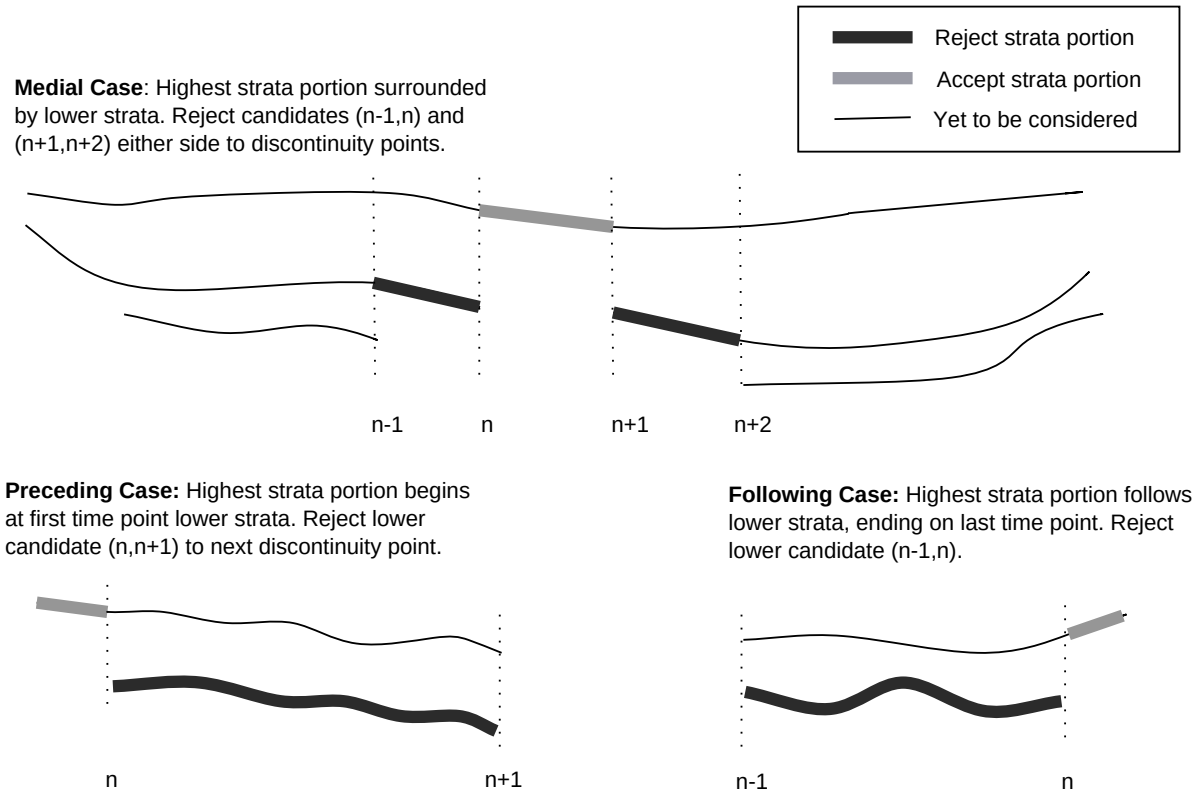

Figure 50: The three cases to consider within the greedy-choice algorithm (Table 6) when expanding from the established highest scale (i.e. most likely candidate ridge), by rejecting lower scale ridges. A set of subcases must also be considered when the reject portion beginning/ends are also the beginning/end of the analysis window.

greedy-tactus-extractor:

1: $S(T) \leftarrow$ array of the scales of all candidate ridges in the time-frequency plane.
2: **do**
3: $s(t) \leftarrow$ the lowest S for each time point $t \in T$
4: absFreqChange $\leftarrow |s'(t)|$
5: **if** max(absFreqChange) $> \Delta s$
6: $d \leftarrow \{t : t \in T,$ absFreqChange$(t) > \Delta s\}$
7: peakTime $\leftarrow t$, argmin $s(t)$
8: $n \leftarrow \exists n$ such that peakTime $\geq d(n)$ **and** peakTime $< d(n+1)$
9: **if** PRECEDING CASE: peakTime $\leq d(1)$
10: delete $[d(1) \rightarrow d(2)]$ from s
11: **else if** FOLLOWING CASE: peakTime $\geq d($length$(d))$
12: delete $[d(n-1) \rightarrow d(n)]$ from s
13: **else** MEDIAL CASE:
14: delete $[d(n-1) \rightarrow d(n)]$ from s
15: delete $[d(n) \rightarrow d(n+1)]$ from s
16: **while** max(absFreqChange) $> \Delta s$ **and** length$(s) > 1$
17: **if** length$(s) > 1$
18: **complete:** Tactus is the lowest scale in s
19: **else**
20: **incomplete:** No continuous tactus could be found.

Table 6: The greedy-choice algorithm for extracting the tactus from all candidate ridges.

5.5 Ridge Tracing Results on Selected Examples

5.5.1 Sinusoidal Signal

To demonstrate the correctness of the ridge tracing methods, a hyperbolically slowing constant amplitude sinusoidal signal

$$s(t) = \cos(2\pi t \omega_s + \alpha \ln(1 + \beta t))$$

is analysed in Figure 51 and ridges are plotted in Figure 52. Parameters are set as $\alpha = 100$, $\beta = 50$, $\omega_s = 40$. Spurious ridges are generated by slight magnitude variations at the edges of the analysis window by all three methods. However, all three identify the correct ridge without gaps, overlaying each other on the plot. Ridge tracing methods were also verified to correlate properly with the isochronous impulse signal shown in Figure 15.

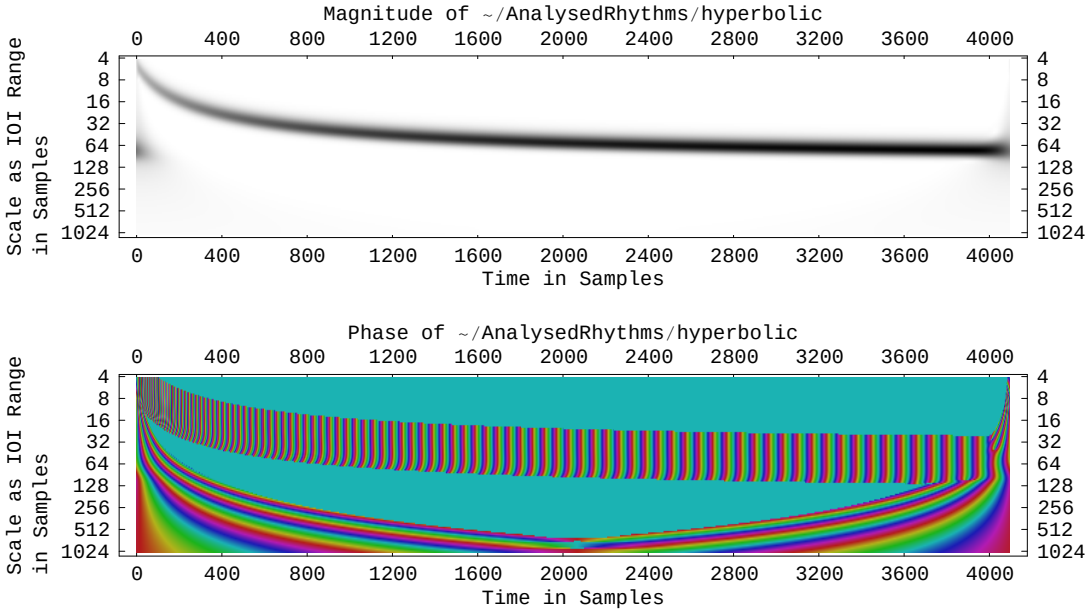

Figure 51: Scalogram and Phaseograms of a hyperbolically slowing constant amplitude sinusoidal signal.

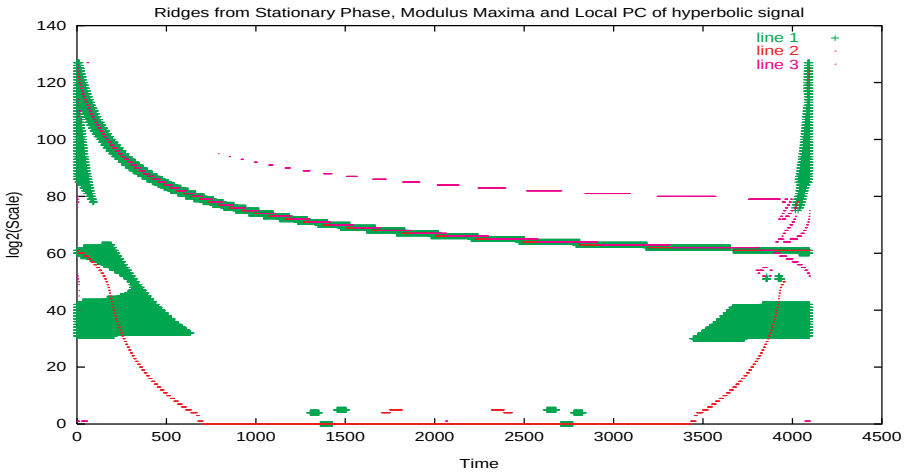

Figure 52: Ridges extracted from the signal analysed in Figure 51. The three ridges derived from stationary phase (Line 1), modulus maxima (scale derivative, Line 2) and local phase congruency (Line 3) all coincide.

5.5.2 Anapest with Rubato

The ridge tracing techniques are then demonstrated on an anapest rhythm undergoing an asymmetrical rubato, a ritard to half speed followed by a more rapid acceleration to the original rate (Figures 53 and 54). The interpretation of the axes is exactly the same as interpreting the scalogram and phasogram plots in Chapter 4. The expressive rubato appears clearly as a number of parallel partials reflecting the tempo curve.

While there is a more continuous ridge generated with modulus maxima, than with the stationary phase or local phase congruency ridges, there are time points where the dilation modulus is missing, between samples 5000 and 6000, at approximately dilation scale number 30, during the rapid acceleration back to the original beat rate. In order to create a continuous ridge, modulus maxima and local phase congruency were "or'd" to create the candidate ridges before the tactus extraction algorithm (Table 6) was run. For this rhythm, the stationary phase method tended to only produce spurious ridges around the main ridges which tended to defeat the continuity condition (Δs) in the extraction algorithm.

The ridge extracted as the tactus by the greedy algorithm is shown in Figure 55. The extracted ridge does not perfectly reconstruct the rapid acceleration between samples 5000–6000, but does produce a reasonable approximation. It correctly tracks the variation of the rhythm. The short acceleration from samples 0–500 is due to a spurious local phase congruency ridge which is within Δs continuity threshold. The ridge portion which should have been selected appears above. This ridge would have been selected if the congruency ridge had been rejected. This could have been achieved with a higher resolution of dilation discretisation, at the cost of a significant increase in processing burden.

5.5.3 Greensleeves

The ridge tracing approach is demonstrated on a more complex, musical example, the quantized rhythm of "Greensleeves". The quantized version was used in order to assess the accuracy of the ridge tracing which should, in this case, ideally produce an unvarying ridge across the window. The ridges are displayed in Figure 56, together with the expected tactus derived from its notation, and a tactus output by the algorithm described in section 5.4. The expected tactus corresponds to a dotted crochet interval (see Figure 41) at 60 BPM, 300 samples IOI at a 200Hz sample

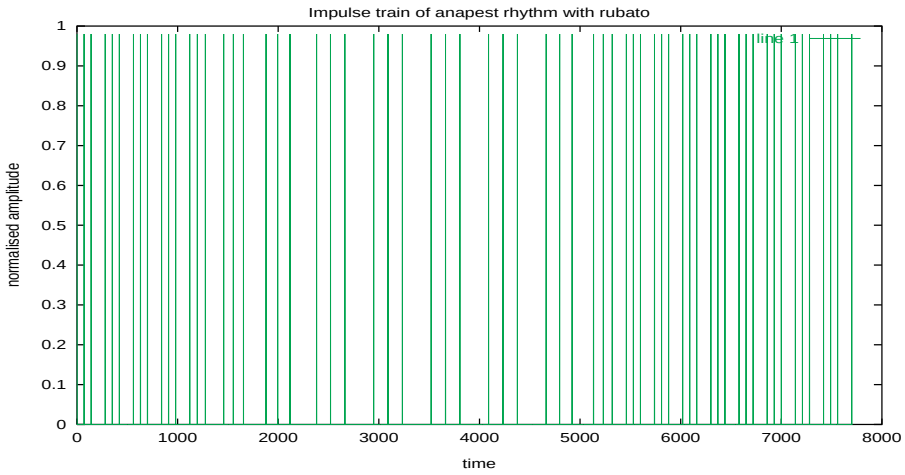

Figure 53: The impulse representation of the anapest rhythm as input to the ridge extraction algorithm.

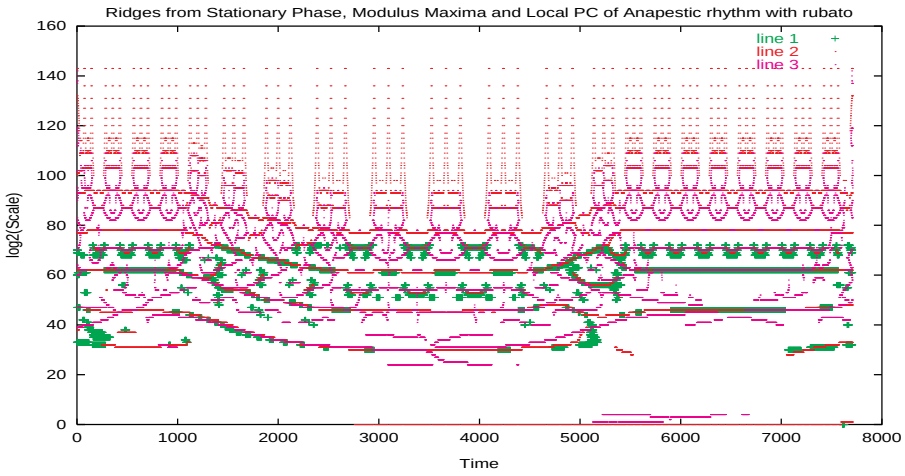

Figure 54: Ridges extracted from the anapest rhythm undergoing ritard then accelerate rubato of Figure 53. Line 1 is the stationary phase ridge, Line 2 is the modulus ridge with respect to dilation, Line 3 is the local phase congruency ridge. For clarity, the stationary phase ridges are only plotted below scale number 72.

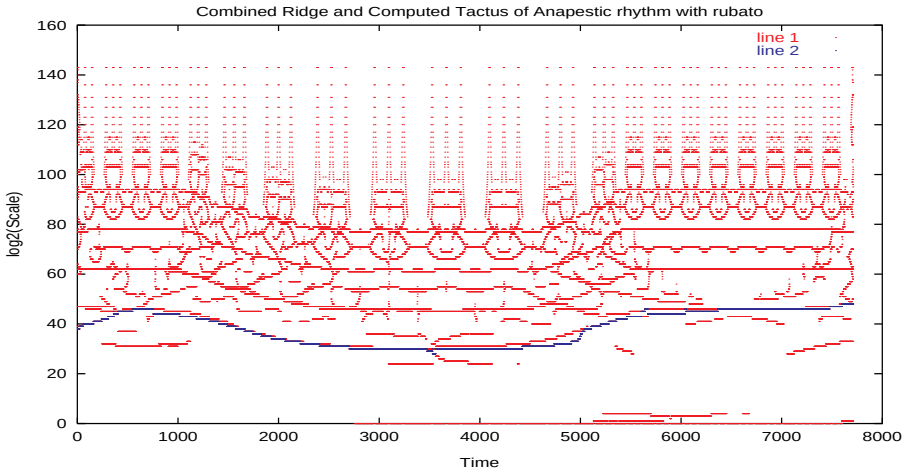

Figure 55: Tactus extracted from ridge candidates of Figure 54, formed from "or-ing" the modulus maxima and local phase congruency (Line 1). Line 2 is the extracted tactus ridge.

rate. This translates to a scale index of 44.

There are several points to note. Stationary phase activates particularly strongly on the frequency corresponding to the pulse rate,[3] but appears in only isolated patches for all other lower frequency ridges. These lower ridges are more strongly identified by the modulus maxima plots (Equation 33). Correlation with all three ridges failed to produce a ridge that could be extracted, due to the spurious ridges causing the continuity condition to fail to find a continuous ridge. Correlation of the modulus maxima and local phase congruency ridges produced enough spurious ridges to cause the tactus algorithm to skip the lowest ridge detected by the modulus maxima contribution alone. The second lowest continuous ridge selected by these two ridge methods corresponds to the expected tactus but is achieved "under false pretences", as the lowest ridge meeting the tactus principles is achieved with the modulus maxima alone. Therefore, the ridge extracted was obtained purely from the modulus maxima as shown in Figure 57. As will be demonstrated in Section 5.6.3, the tactus extracted is half the expected rate, corresponding to the downbeat. Results demonstrating the correct rate appear in that section.

[3] An IOI of 98.7 samples, corresponding to scale index 70 with 16 voices per octave, close to the quaver at 60 BPM, having an IOI of 100 samples at 200Hz sample rate.

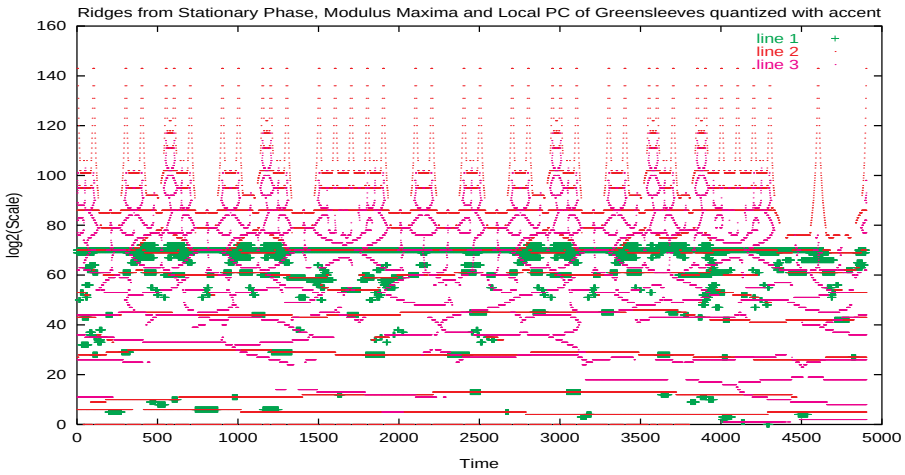

Figure 56: Ridges extracted from the dynamics accented quantized rhythm of "Greensleeves" (see Figures 41 and 42). The lines in numbered order are: stationary phase ridge, modulus maxima ridge, local phase congruency ridge.

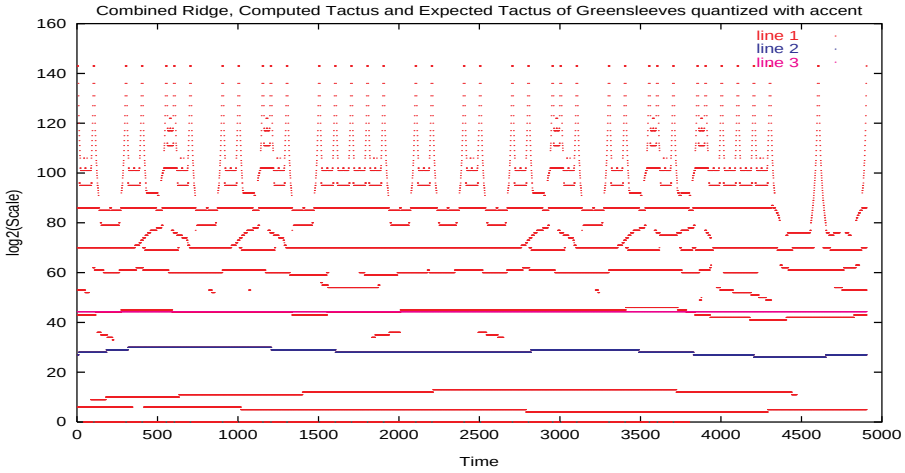

Figure 57: Tactus extracted from the ridge formed by the modulus maxima of the rhythm of "Greensleeves". The lines in numbered order are: modulus maxima ridge, tactus derived from the greedy algorithm in Table 6, tactus expected from notation.

5.6 Foot-tapping

5.6.1 Sampling the Tactus

Once the tactus has been extracted from the rhythm, it can be used to compute tap times. Preliminary investigations determined times of each foot-tap by sparse sampling the tactus [164]. Each foot-tap is computed from the period of the tactus frequency indicated by the dilation scale at the time of the previous foot-tap:

$$f_{i+1} = f_i + 2^{\frac{T(f_i)}{v}} \tag{40}$$

where v is the discretisation of voices per octave, and T is the vector of scale indexes produced by the tactus algorithm. This foot-tap is only synchronised to the original rhythm on the first beat, so there is considerable scope for cumulative error. Indeed, by the Heisenberg inequality, the accuracy of computing time from frequency can at best be the interval $(f_i \cdot 2^{\frac{-1}{v}}, f_i \cdot 2^{\frac{1}{v}})$. This error is dependent on the tactus values and v; with an IOI of 256 samples, the interval is $(245.15, 267.33)$, or an error of $\approx \pm 11$ samples, or 55.45 milliseconds for a 200Hz sample rate. This is too high an error and the asynchrony would be perceptible. Clearly a higher discretisation would reduce the error, at the expense of computation time.

Another problem with this algorithm is the requirement to compute all foot-taps relative to the first beat. This ignores the function of anacrusis beats, such as in the Greensleeves example. Unless human listeners have a memory of the rhythm, they will not begin clapping from the first beat, so a future research task is to identify appropriate first tap beat. Therefore the beat to begin tapping on (selecting the phase of the foot-tap) was manually chosen for each rhythm, so that the Greensleeves example started from the second beat, the other rhythms from the first.

It was initially hypothesised that phase congruency could be used to weight the intensity of the foot-tap, thereby reflecting the structural location of the beat with the intensity of the tap. However, as Figure 47 reveals, the original beats will dominate the phase congruency measures, rather than the congruency representing low frequency structure, it will simply indicate how close any foot-tap is to the original beats, not how close that tap lands near a structurally significant time point. It is a future research task to introduce a usable intensity weighting of each foot-tap. In addition, research will be needed to create some acceptable match of "beat importance" to auditory intensity, including the timbre of the foot-tap.

5.6.2 Reconstruction of the Tactus Amplitude Modulation

The simple foot-tapping algorithm of Equation 40 is problematic when sampling a tactus undergoing rubato due to the accumulation of error. This problem provoked a more appropriate approach of reconstructing a tactus in the time-domain and using it to compute the foot-tap times. From the time-frequency domain of the tactus, a FM sinusoidal foot-tap signal is reconstructed in the time-domain. Only the tactus ridge itself contributes to the foot-tap signal. The sinusoidal nature of the signal enables it as an amplitude envelope, that is, as a rhythm frequency that modulates over time. This signal was reconstructed from both the scalogram and phaseogram coefficients.

All scalogram coefficents other than those of the tactus ridge (identified by the greedy algorithm) were clamped to zero, while the original phase was retained. This altered magnitude matrix and the original phase matrix were converted back to complex valued coefficients and used as input to the reconstruction equation, originally introduced as Equation 9

$$s(t) = \frac{1}{c_g} \cdot \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_s(b,a) \cdot g(\frac{t-b}{a}) \, \frac{dadb}{a^2}. \tag{41}$$

The constant $c_g = 1.7$ was determined by calibrating the original time signal with its reconstruction $s(t)$ to achieve energy conservation. Due to the asymptotic tails of the Gaussian, the reconstruction cannot be perfect, but it was determined that the reconstruction would accurately resynthesize the frequency and phase of signals.

While the final version of the new algorithm clamped to a single dilation scale per sample, an initial version placed a Gaussian envelope across dilation scales at each time point, centered over the scale of the tactus. In practice, the reconstructed sinusoids were both the same frequency and phase. They only differed in their magnitude, which did not impact on determining foot-tap positions. The real component of the reconstruction

$$A_s(t) = \Re[s(t)] \tag{42}$$

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

$$\phi_s(t) = \arg(\Im[s(t)], \Re[s(t)]). \tag{43}$$

While the peaks of the amplitude modulation (Equation 42) were verified to produce the correct foot-tap points for an isochronous tactus from the pulse (Figure 15) that aligned with the original rhythm, problems would arise with rhythms which were phase shifted from the occurance of an anacrusis. Therefore the $\phi_s(t)$ value was noted for t at the first beat to begin tapping, and the remaining foot-taps were selected for each $\phi_s(t)$ that matched that initial phase value. These tap times were then used to generate a Common Music "thread" of note events [178] indicating times to synthesise a sampled hi-hat sound[4] that could be mixed with the original rhythm. This enabled the foot-tap accuracy to be audibly assessed.

5.6.3 Examples of Foot-tapping

Foot-tapping can now be assessed with selected rhythms for which the tactus is already in common agreement among listeners, forming an expected outcome. This can be potentially misleading, as an expected outcome that is understood and notated as the correct clap rate, and indeed is tapped to, may not be the underlying tactus—the pulse—the listener initially induces. Instead, a higher harmonic of the tactus could in fact be used to accompany (tap to) the presented rhythm. For example, a listener may sense the pulse falling at downbeats of $\frac{\mathbf{4}}{\mathbf{4}}$ measures and yet tap at the crochet, the fourth harmonic. Although this would strain the definition of a tactus, it is important to understand the tap rate as a performed accompaniment to a rhythm that can itself be "embellished" or "filled in" at harmonic rates. This variation in choice of tap rate and phase appeared in the experiments on subjects conducted by Parncutt [130].

Greensleeves Foot-tapping

Using the tactus extracted from the modulus maxima ridge of Greensleeves (Figures 56 and 57), produced the foot-tap accompaniment displayed in Figure 58. Typically the downbeat interval differed from the ideal by between 1 and 51 samples, with most under 12 samples. These errors resulted from the undulation of the extracted tactus ridge, but were not cumulative.

[4]The hi-hat cymbals on a drumkit typically have a time-keeping function in popular music, so

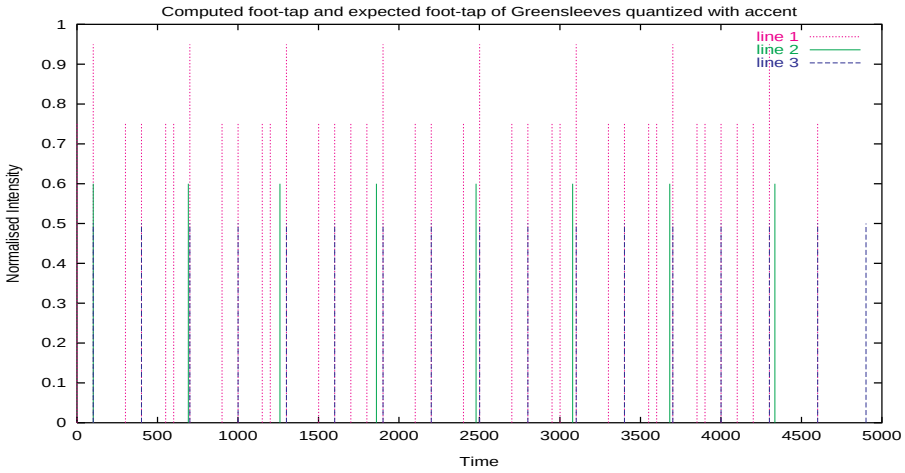

Figure 58: Foot-tap of Greensleeves from the modulus maxima derived ridge. Line 1 is the original rhythm, line 2 is foot-tap points computed from the tactus, line 3 is foot-tap points computed from the expected tactus, assuming it corresponds with the notated meter. The amplitude of the last two signals are arbitarily scaled for clarity, line 1 shows the amplitude of the original signal.

The Greensleeves tactus selected by the tactus algorithm using the modulus maxima ridge alone was half the expected tap rate. It is instructive to display the results when using both the modulus maxima and local phase congruency, which resulted in a tactus being extracted which is extremely close to the expected result (Figure 59). The phase of the tactus and the derived foot-taps appear in Figures 60 and 61, with the close match (maximum error +34/-5 samples) between the theoretical and computed foot-tap position both visually and audibly apparent.

Foot-tapping During Rubato

The extraction of ridges and tactus from the anapestic rhythm undergoing rubato was demonstrated in Figures 54 and 55. The foot-tapping beats are plotted overlaying the original rhythm in Figure 62 and 63. This example demonstrates the use of an extracted modulating tactus to foot-tap to a rhythm undergoing significant asymmetrical rubato. Regions of constant tempo, deceleration and faster acceleration are all correctly tapped on the first beat.

The artifactual acceleration of the ridge at the start of the rhythm (noted in

it was deemed an appropriate timbre to use to qualitatively assess the foot-tapping.

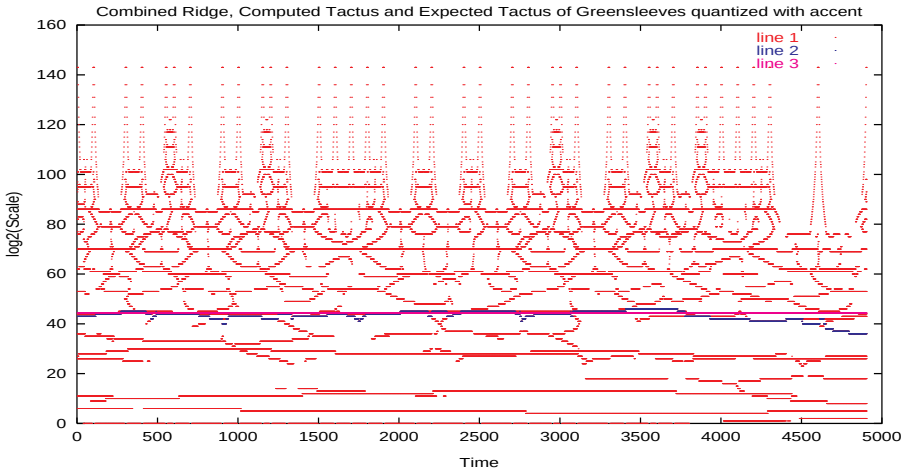

Figure 59: Alternative tactus of Greensleeves. The lines in numbered order are: ridge correlated from modulus maxima and local phase congruency, tactus derived from the greedy algorithm in Table 6, and tactus expected from notation.

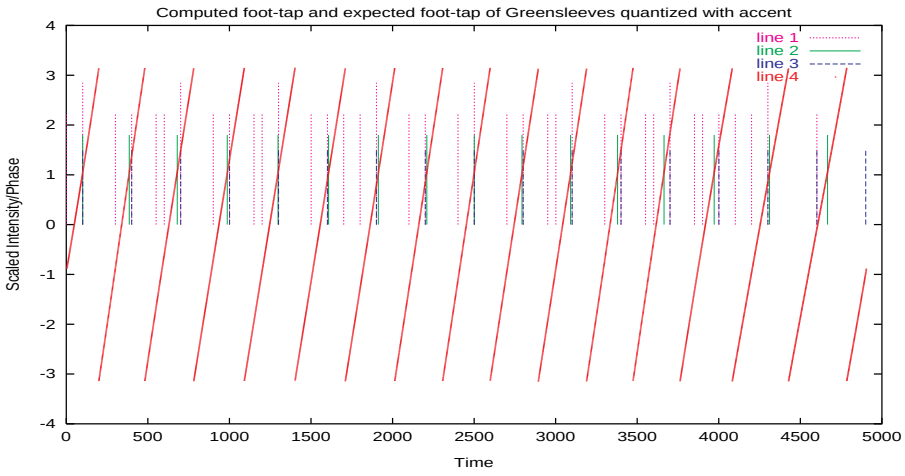

Figure 60: Alternative Foot-tap (line 2) of Greensleeves (line 1) derived from tactus phase (line 4), together with the expected foot-tap from the notation (line 3). Intensities of the original rhythm and the foot-taps have been scaled from 0–3 to clarify their relationships to the phase ($-\pi$ to π).

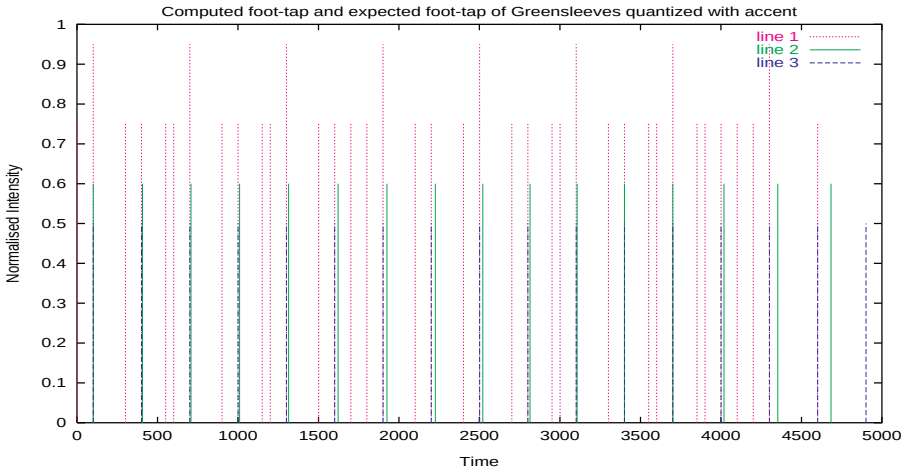

Figure 61: Alternative Foot-tap (line 2) of Greensleeves (line 1) derived from tactus phase, with the expected foot-tap from the notation (line 3).

Section 5.5.2) perturbs the phase used to compute foot-tap times, so the foot-tap algorithm of Section 5.6.2 was forced to begin tapping from the second anapest or third anapest group (the fourth or seventh beats respectively). If the tapping began on the first beat, the phase changed at a slower rate and the initial phase value would skew all subsequent beats. An improved tactus extraction algorithm extracting the correct initial rate would avoid this problem. If the tapping began on the fourth beat, which is still within the erroneous tactus acceleration region, the perturbed phase produces a maximum error of 20 samples (100 msec) from the ideal tap point. This gives the impression of a dragging feel, rather than a lack of rhythm. The failure is a relatively "graceful" degradation in performance. Starting the tapping from the seventh beat reduces this lag to nearly zero. An audible example of the foot-tap mixed with the original rhythm reveals an occasional slight asynchrony between the foot-tap and the first beat of each anapest group. While the initial tactus phase datum is identified from the seventh beat, the fourth beat is still identified as an appropriate time to foot-tap, though 100 msec early due to it's skewed phase.

In both cases, the foot-tapping algorithm does remarkably well on a rhythm that changes speed so rapidly. It should be noted that it can be quite challenging for a human listener to accurately tap to such a rubato rhythm. The deceleration will typically lead to human errors in judging the time of the next beat. While

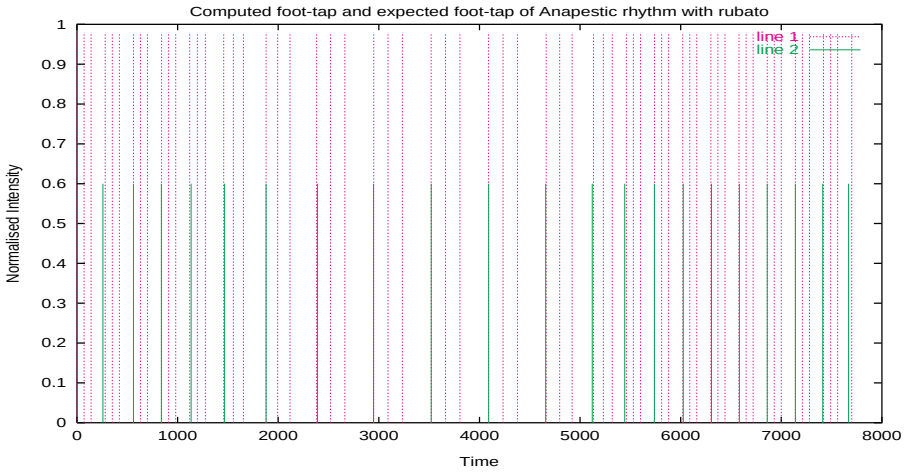

Figure 62: Foot-tap of the anapestic rhythm undergoing asymmetrical rubato. The first line is the original rhythm, the second the foot-tap accompaniment. Foot-tapping was determined from the phase of the seventh beat.

accurate in always clapping to the first beat of each anapest grouping, selection of the first beat ignores the effect of tempo on interval accents. Indeed, for this anapestic rhythm, listeners will typically accent the first beat at fast tempi and the last of the three at slow tempi. However this appears rather context sensitive. The tempo rates at which such accentual inclinations would induce listeners to continue to tap (at a syncopation) on the first beat, and when to switch to the third must be systematically investigated for human listeners, before a computational approach is proposed.

Investigation of an Example of Mismatch to Expected Tactus

Desain and Honing's quantized web rhythm (analysed in Figure 37) demonstrates the failure of the algorithm to match the expected tactus. This provides insights into the limitations of the greedy algorithm and the foot-tap generator. The ridges, tactus and foot-tap appear in Figures 64,65 and 66 respectively.

A ridge extracted from modulus maxima "or'd" with stationary phase enabled a tactus to be extracted. This tactus is of a lower frequency than the expected tactus from the notation (corresponding to a crochet at 100 BPM with an interval of 120 samples). As Figure 66 shows, the extracted tactus corresponds to an initial interval

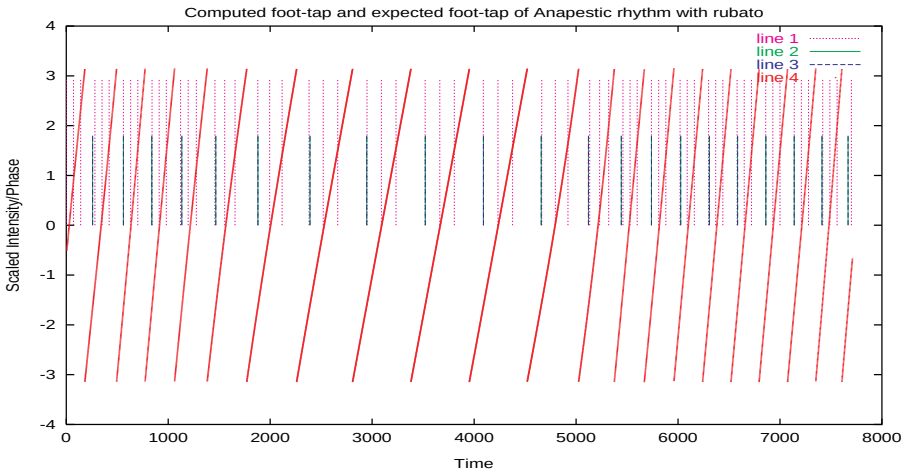

Figure 63: Phase of the foot-tap of the anapestic rhythm undergoing asymmetrical rubato. The first line is the original rhythm, the second and third the foot-tap accompaniment, the fourth the foot-tap phase.

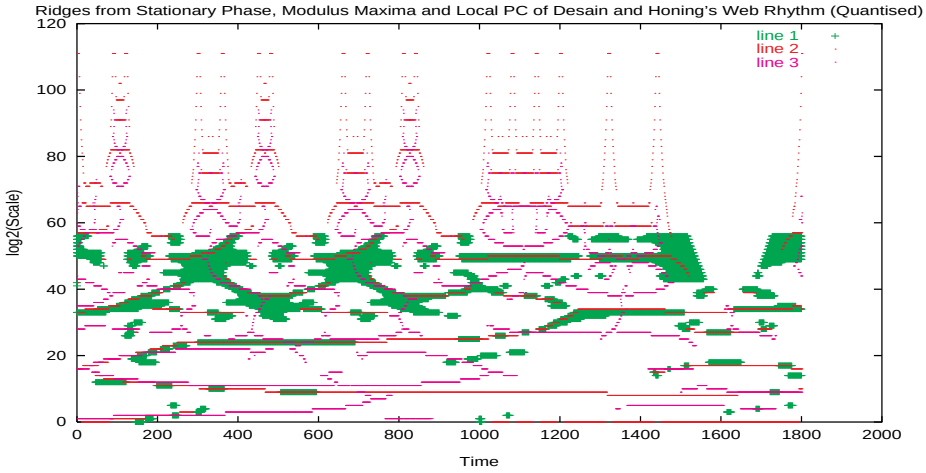

Figure 64: Ridges of Desain and Honing's rhythm analysed in Figure 37. The lines in numbered order are: stationary phase ridge, modulus maxima ridge, and local phase congruency ridge.

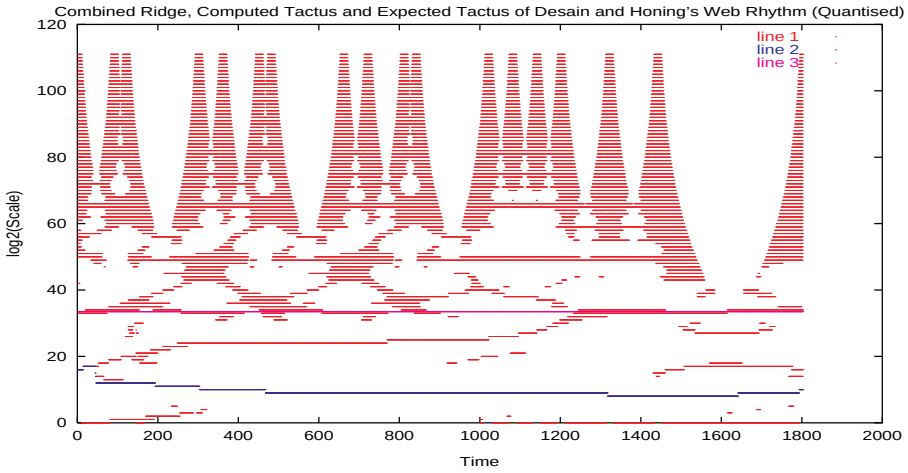

Figure 65: Tactus of Desain and Honing's rhythm extracted from the ridge formed by the modulus maxima or'd with stationary phase. The lines in numbered order are: combined modulus maxima and stationary phase ridge, tactus derived from the greedy algorithm in Table 6, and tactus expected from notation.

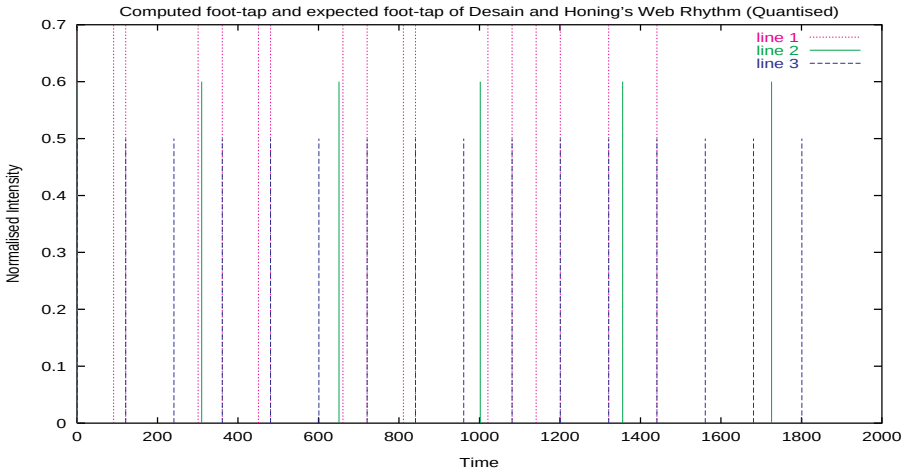

Figure 66: Foot-tap (line 2) of Desain and Honing's rhythm (line 1) derived from tactus phase, with the expected foot-tap from the notation (line 3).

of a minim tied with a quaver (2.5 crochets) which then descends in frequency to a dotted minim (3 crochets) for the remainder of the rhythm. As is apparent from Figure 65, the first interval is spurious and an artifact of the tactus extraction algorithm. However, if the tap times are used to synthesize audible beats accompanying the original rhythm, the claps create a syncopation, but still definitely create the impression of an unlikely but plausible accompaniment to the original rhythm.

As Figure 65 indicates, the expected tactus ridge is not continuous across the analysis time. The tactus pulse does not receive enough energy during the period of samples 900–1200 to retain its ridge. Re-examining the scalogram of Figure 37 reveals this is an artifact of the ridge extraction methods, as there is visually an energy at that period. However the energy does not always form a magnitude peak and therefore a ridge everywhere, as it is disipated by nearby ridges. In addition, this ridge does not correspond to the highest modulus peak. Therefore it would appear neccessary to introduce a tempo weight to scale this ridge to prominance. It can be concluded that the notion of a peak, producing a ridge, is too narrow a definition for faultless interpretation, but that within limited contexts, it can function well for interpretation purposes.

Fabricating the expected tactus and using it to compute the foot-tap using the phase indicates how successfully an isochronous pulse can be synthesised. As described in Section 5.6.2, this is achieved by constraining the magnitude to the single voice corresponding to the expected foot-tap rate, yet retaining the original phase. Figure 67 plots the match between the expected foot-tap resynthesised from the single voice and the known expected foot-tap interval. The deviation of beats 9 and 17 reveals that the original phase seems to skew the pulse from the ideal, but this is slight (maximum of +6/-4 samples error). While it is possible to synthesize the phase in the simplest case of an isochronous tactus, the general case of a modulated tactus seems difficult to correctly predict. More importantly it seems theoretically incorrect to ignore the signal's original phase when the intention is to resynthesize a particular partial of the rhythmic signal.

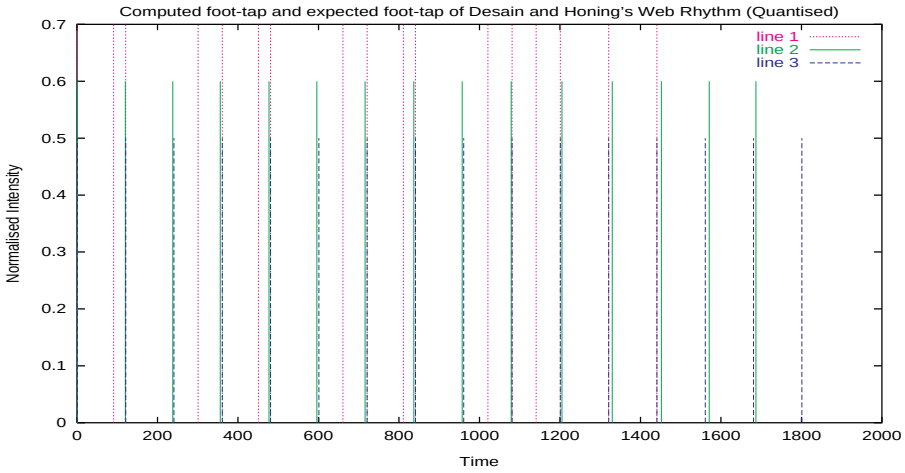

Figure 67: Foot-tap (line 2) computed from the expected tactus of Desain and Honing's rhythm (line 1). Line 3 is the expected foot-tap using the expected tap IOI.

5.7 Assessment of Results

5.7.1 Ridge Generation and Correlation

No single method of correlation of different ridge forms was found to reliably achieve a robust single ridge (Section 5.2.5). In fact the inclusive correlation (or'd ridges) could have the opposite effect, creating several ridges near the modulus maxima ridge which could result in valid ridge portions being rejected due to their proximity within Δs. This was illustrated in the Greensleeves tactus identification which skipped the lowest ridge when both local phase congruency and modulus maxima were correlated. The extra ridges also increase the running time of the tactus extraction algorithm, increasing the number of ridges to reject as well as the time spent calculating the stationary phase and local phase congruency ridges.

As mentioned in Section 5.2.5, the modulus maxima alone was often unable to provide a valid ridge which held to the continuity hypothesis in the case of rapid frequency variation such as the acceleration of the anapestic rhythm analysed in Section 5.5. The stationary phase algorithm detects ridges effectively on signals where the oscillation of the signal is clear, as with the hyperbolic example and an isochronous pulse. It has been much less successful with rhythmic signals, where

the stationary phase was only partially helpful in combination with the modulus maxima. Likewise, the local phase congruency was particularly effective in recovering rapid accelerations, but could also introduce spurious ridges which made it only effective in combination with the modulus maxima.

It seems undesirable to propose a heuristic as to what correlation mechanism should be used for each rhythm. It seems more fruitful to devise a tactus algorithm which does not rely on a single correlated set of ridges, but instead weights the contribution of each ridge approach during the tactus determination.

5.7.2 Asymptoticism and Undulating Ridges

All ridge extraction methods tended to produce slight variations in the ridges, despite in some cases the original signal being quantized to integer ratio IOIs. This undulation appears to be due to the discretised number of voices per octave in the case of modulus maxima (with respect to scale) and local phase congruency. Computational burden prevents increasing the resolution above that used here (16 voices per octave), but would result in less variation and therefore less inaccuracy when foot-tapping.

However, examining scalograms of Greensleeves (Figure 42) suggests that the ridges also undulate due to the interactions between ridges. This seems most likely from the contradiction of the asymptotic condition on the signal (Equation 26 described in section 5.2.1) composed of sparse infinitely fast impulses (see section 5.2.2). The infinitely fast amplitude "envelope" of the impulse contradicts the requirement that the phase of the signal changes faster than its amplitude. In the case of simpler signals such as the isochronous pulse (Figure 15), the ridge is perfectly straight.

While it seems that this contradiction compromises the analysis capabilities of the CWT, demonstrably there is still considerable information made apparent by the approach. The effect of these undulations is to produce unnecessary variations in the tactus and therefore the tap points. There are two possible solutions to this, to use a low pass filter to smooth the tactus before using it to compute foot-tap times, or to devise a tactus algorithm which does not operate only on the ridges, but on the entire scalogram, local phase congruency, and stationary phase fields, such that broader regions of data are assessed. The latter approach holds promise to compensate for interaction between ridges.

5.7.3 The Tactus Algorithm

Although it is important to emphasise that the tactus algorithm is simply identifying one of the ridges as prominent, it clearly is not biologically plausible. As a first means to extract out a modulating ridge that was visually apparent on the scalogram to allow testing that ridges are applicable to the resynthesis of foot-tapping, it has served its purpose.

The tactus algorithm is limited in a number of ways. It is susceptible to gaps in the ridges, and is therefore reliant on the ridge extraction and correlation approaches finding a continuous ridge across the window. It is sensitive to Δs (see Table 6). It can reject valid ridges if Δs is set too high to not detect a discontinuity properly. There is no accounting for likelihood weight of the ridge, all ridges are assumed equal. This is clearly too simple. However, resolving the relative contributions of ridge weightings becomes more complex when correlating between multiple ridge methods. The algorithm's lack of tempo sensitivity limits its use to rhythms that fall within narrow ranges of intervals. This clearly needs to be addressed in the future.

Statistical models, and (with less successful outcomes), neural or evolutionary approaches that do not rely on rules, have shown promise in overcoming contextual limitations. Such a limitation seems to occur with the continuity condition on the ridge extraction algorithm. The use of curve fitting statistical methods seems a fruitful avenue to pursue. Such a method must allow for distinguishing between close parallel curves, identifying several simultaneous candidates or weighting contributions from the ridges to identify the most likely candidate curve. Clearly with such an approach there is scope for biological weightings, such as absolute tempo constraints, to be incorporated.

In practical terms, and for the results reported here, better ridge extraction would be only an incremental improvement. The main tenant of the hypothesis—that time-frequency approaches to analysis of rhythm can be used to provide interpretable information for accompaniment tasks—has been satisfied. While a "cleaner" tactus would be required for applications such as event editing respecting rubato, the foot-tapping task is only sampling the tactus at isolated points, so the results for this task can be expected to only marginally improve. It remains to be investigated whether simply low pass filtering the currently extracted tactus will compare to a curve fitting approach.

5.7.4 Foot-tapping

The tactus phase based foot-tapping algorithm has proven effective in computing tap times of an accuracy that is acceptable in listening tests. The most apparent limitation is the requirement to manually determine when to start tapping. A future task is to investigate if it is possible to compute some measure of confidance of rhythm ambiguity from ridge acceleration behaviour. When the confidance rises above a threshold, foot-tapping can commence.

As indicated in Section 5.7.3, the accuracy of the foot-tapping is dependent on the tactus extracted. It was neccessary to closely monitor the tactus so extracted to determine that the foot-tapper was indeed accurate. As verified by using a fabricated (albeit constant rate) tactus, where the extracted tactus is incorrect, its fabricated substitute produced good tapping results.

Foot-taps are currently of fixed intensity and currently there is no determination of accentuation, in particular from tempo sensitivity. This is a future research task.

5.8 Summary

The multiple hypothesis aspects of rhythm perception are made explicit with wavelet analysed rhythms. The use of ridges derived from modulus maxima, stationary phase and local phase congruency have been demonstrated to successfully extract the tactus of sparse impulse rhythm signals. These different ridges are independent interpretation mechanisms acting on the modulus and phase representations of the time-frequency plane. These can be considered as perspectives of the time-frequency representation of the signal. These have been the focus of research to understand their use in rhythm interpretation applications due to such mechanisms being unique to the nature of multiresolution analytic (phase-preserving) wavelets.

The correlation between these methods as a means to produce well recovered ridges is new. Modulus ridges have been previously used by Todd [103], in the application of a biologically influenced, real-valued wavelet to rhythm perception. Whereas the research reported here are the first examples of ridges obtained using analytical wavelets. This approach enables magnitude ridges to be quantified and used in synthesis of accompanying rhythms. This is the first use of stationary phase to compute ridges from impulse characterised signals, in particular, rhythmic signals. The new concept of local phase congruency has been proposed and demonstrated as

a means to identify harmonics within an analysed signal.

With the construction of robust ridges, an algorithm for the extraction of one of several candidates as the tactus ridge has been documented. This is deliberately simple and does not seek to model rhythm by using currently reported biological features thought to contribute towards rhythm perception. Rather it adopts a behavioural model distilled into general principles, investigating the information inherent in the signal. The intention has been to determine clearly the information available and its use, and to quantify the limitations of such an approach. In particular, there is no explicit perceptual model proposed to distinguish interpretation of time spans within the subjective present (Section 2.1.2), compared to time spans longer than the subjective present. Again, this has been delayed to characterise exactly the degree to which an explicit model is necessary. In the future, a biologically motivated model can be constructed in the time-frequency domain, with a clear reference to information which any such multiresolution model must encounter. Several criteria for an improved tactus algorithm have been proposed for future investigation.

The accuracy of the tactus so extracted has been tested by applying this to the production task of foot-tapping to the rhythm. The results demonstrate that the tactus (when correctly extracted) can be successfully used to synthesize accurate foot-tap beats that respect structured tempo variation in the analysed rhythm. This also demonstrates the power of the CWT reconstructive capabilities in analysis-by-synthesis applications. This is the first application of reconstructive multiresolution approaches to musical rhythm, and the first demonstrations of the feasibility of the approach. The theoretical framework of wavelet transforms has enabled the problem of foot-tapping to be quantitatively cast in clear terms of selective attention to rhythmic strata.

Chapter 6

Conclusions and Future Directions

6.1 Concluding Assessments

In this thesis, phase preserving Gabor wavelets have been proposed, implemented and tested 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 existing music theories of the inducement of temporal structure, meter and expressive timing. The notion of multiple hypothesis during rhythm perception is made explicit with the time-frequency representation of rhythmic signals. This approach has allowed a rhythmic frequency to be determined that can be considered the foot-tapping rate. This modulating frequency has been used to resynthesize accompaniment rhythms that respect the tempo behaviour of the original rhythms. This task has provided a complete perspective on interpretation of a performance: from conception, to production, to listening, to accompaniment.

6.1.1 Frequency Analysis

This research has explored the idea of viewing a canonical pulse as an underlying sinusoidal oscillation at a frequency given by a wavelength equal to the inter-onset-interval. This is an approach that unifies many aspects of rhythm into a single conception.

As has been shown, a large amount of information can be obtained from the rhythm signal, independent of modelling perceptual processes. This has enabled

investigation of the value of the multiple resolution approach without inherent assumptions of processing mechanisms used in the cognition of rhythm. Indeed, many of the features of rhythm; the use of accentuation and time intervals, retrospective evaluation, and forward time projection of expectancy are all managed by the decomposition onto non-causal basis functions. Potentially, a model of tempo preferences and absolute performance limits could be constructed in the time-frequency domain, in terms of selective band-pass filters.

Of course, any model of music cognition will be limited by the degree to which it represents psychological behaviour. In that sense, the value of a theoretical model of time-frequency representation for rhythm perception modelling is currently limited despite the benefits of such an approach detailed in Section 3.3.2. Clearly a final version of a rhythm *perception* model must incorporate a model of human temporal constraints noted in section 2.4.2.

The multiresolution approach has shown good representation of expressive timing. Acceleration and ritards are made apparent, providing appropriate visualisation of rubato. It is possible to track tempo changes by following the apparent strata. This is a reproducible analysis that can be interpreted both manually and computationally. This has not required a specific model of the behaviour, it simply appears as part of the decomposition function of the wavelet representation. The formal concept of expressive timing as a frequency modulation of underlying canonical pulse hierarchies has intuitive and descriptive benefits. For example, phrase final lengthening seems to be the rhythmic equivalent of octave-stretching, that is, slight departures away from integral ratios. In the context of a total phrase there will not be a canonical meter, but instead a warping function that will deviate the beat frequency in a characteristic fashion, thereby communicating the information inherent in the expression. The use of time-frequency representations makes this activity apparent.

There are similarities between this multiresolution wavelet analysis of rhythm and Desain and Honings decomposable theory of rhythm projecting expectancy measures [23]. Both project a form of localised function over time. Their localised function is forward and backward projected in time, however the wavelet approach translates the localised function every sample, while Desain and Honings approach projects at intervals corresponding to harmonic ratios. For both models there is the opportunity to assess the value of the time-point based on convergence of time scales. The wavelet approach generalises this applied cognitive musicological theory

by unifying it with signal processing research. This has borne fruit in terms of performance (use of the FFT to perform the convolutions of the signal with the scaled and translated wavelet), verification through applications in other domains (examining sound signals), and generality (use of phase, phase congruency and ridge extraction).

6.1.2 Multiple Resolution and Ridges

In it simplest form, multiresolution analysis is akin to dividing down the most prevalent IOI's in the rhythm. This could be achieved by assessing the rhythm at time spans which are powers of 2 and 3 (duple and triple time) of the prevalent IOI's. However multiresolution analysis also captures many aspects of rhythm interpretation, in particular, modulation of pulse rates, in a coherent manner. While there are many demonstrated advantages to the multiresolution approach, a future task is to compare the results directly against such a divide-down approach to assess if the extra computation time of the wavelet approach is justified. It should be noted that no particular effort has been expended in optimisation, so there may be performance gains to be made.

In quite an opposite manner to the approach adopted by Tanguiane, the multiresolution analysis of rhythm does not reduce or compress data in order to determine encoding or least complexity [177]. Rather it creates an expanded (and invertable) representation that reveals structure that would otherwise be hidden. Widmer describes such a process as a more abstract representation [188, pp. 96]. As shown in Chapter 5, the redundancy of the decomposition has proved to be essential to achieve the extraction of the tactus.

Despite the demonstrated value of the approach, the use of ridges to determine the tactus seems flawed. While concepts like modulus maxima, local phase congruency and stationary phase are worthwhile, the binary selection of ridge/not-ridge can be seen to be throwing away quite a large amount of information before the tactus algorithm begins. This is counter to human visual interpretation of the scaleogram and phaseogram, in that the importance of each coefficient (by grey-scale value) is assessed as a unified body of data. Instead of isolating to single ridges, correlating then extracting, a future research endeavour suggested in Chapter 5 is to combine the entire continuously valued magnitude, local phase congruency and stationary phase fields and use a statistical curve fitting approach over the combined

fields for tactus determination.

6.1.3 Reconstruction

Achieving reconstruction as described in Section 5.6.2 provides a complete perspective on a performance from a performer's conception to another's accompaniment. Using a reconstructive transform (with at least a measure of error in reconstruction if not using orthonormal transforms) allows an analysis-by-synthesis approach, and the ability to systematically verify the results through resynthesis. In this thesis, the resynthesis has been used to address the foot-tapping problem [31]. The ability to measure the performance of the algorithms has produced highly encouraging results.

The reconstruction capabilities promotes the concept that time-frequency rhythm analysis is merely the translation to a new domain of time-frequency, without loss of information, so that it is equally valid to evaluate the rhythm in this domain as the time domain. When selectively reconstructing from certain coefficients in the time-frequency domain the resulting time domain signal has predictable results which are readily interpretable. This applies to non-orthogonal wavelets such as the ones used here rather than orthogonal wavelets which do not have an intuitive impulse response. Such an approach also enables modelling of perceptual constraints as a rhythm-band time varying filter in the time-frequency domain.

The reconstruction then allows the resulting analysis to be seen as purely a *descriptive* representation of the rhythm. This time-frequency representation is therefore declarative, in that the transformation is transparent, energy preserving and intuitive, and analysis is possible on the new representation. These match several of the attributes which are argued by Honing [61] to be important for research in musical time.

6.1.4 How Harmful is an Extracted Tactus?

There are important differences between the current notion of tempo curves (see Section 2.2.7) and the extracted tactus ridges investigated here. The extracted tactus is derived from a context of overlapping beat intervals, rather than computing deviations at each beat time from some supposed canonical tactus, original score or structural annotation. This latter approach of tempo curves does not allow for an accurate instantaneous frequency to be determined to compute new beat rates.

Desain and Honing emphatically proposed the link of structure to expressive timing. Indeed the approach of determining the modulation of the tactus from the original rhythm reflects this link. The extracted tactus is truely a low frequency component of the original signal, identified from the time-frequency domain (i.e the structure) of the rhythm.

When the time-domain tactus is resynthesised from the clamped magnitude and original phase of the signal (Section 5.6.2), the relationship to the original rhythm is preserved in the phase, and the time extents of the magnitude and phase (matching the time extent of the analysed rhythm). It is a future task to see if such tactus elements could be transferred and applied to other rhythms.

Desain and Honing have argued that the ability to transfer a tempo curve is a test of their claimed independence [28, 26, 27]. At first sight, transferring extracted tactus and phase does not seem achievable, as each rhythm will have its own length and characteristic scalogram and phasogram. A possible strategy may be a form of interpolation to match an extracted tactus to a new rhythm's time extent—stretching the tactus to fit, using the original phase. Even in the light of this proposal, the concerns of Desain and Honing remain, and it appears that the extracted tactus is intrinsically dependent on the original rhythm that produced it.

6.2 Contributions

The contributions made in this thesis can be summarised as follows:

1. Proposed a representation scheme for musical rhythm for time-frequency signal processing approaches.

2. Determined the theoretical applicability of *analytical* wavelets to analysing musical rhythm for the first time.

3. Applied the 2-D image processing phase congruency measure to 1-D rhythmic analysis for the first time, and evaluated the information provided by phase congruency.

4. Verified the suitability of analysing musical rhythm using analytical wavelets. This was done by devising and coding a rhythmic database and testing the wavelet analysis on this database. This has shown that there are considerable advantages in representing rhythm in a time-frequency domain.

5. Applied two existing ridge extraction methods of modulus maxima and stationary phase to rhythmic signals for the first time and evaluated the resulting extractions.

6. Devised a new ridge extraction method of *local phase congruency* and investigated the worth of correlation of all three ridge methods. Several alternative methods for correlation were also investigated.

7. Devised and implemented a tactus identification algorithm operating in the time-frequency domain. This is the first time such an algorithm operating in the time-frequency domain has been proposed. The performance of this algorithm was evaluated and further paths of development were suggested to improve its performance.

8. Implemented a foot-tapping system by reconstructing a new rhythm from the extracted tactus. This is the first such approach to foot-tapping, or any other form of structure interpretation, to demonstrate results from reconstruction. The results are extremely encouraging and are an improvement on existing approaches, especially on rhythms with expressive timing.

6.3 Practical Applications and Future Directions for Research

A powerful model of rhythm has a number of computer music applications—transcription, scorefile editing and computer accompaniment, such as score following or interactive performance systems [149, 160]. These applications suggest directions for further development of multiresolution rhythm analysis.

6.3.1 Structure Preserving Quantization

A significant problem in transcription (conversion of performance into visual notation) is the correct deduction of the rhythm back to its canonical conception before performance. The tactus determination method described here appears to be a powerful tool to identify the expressive timing and the canonical rhythm being performed. With the tactus determined, each beat's canonical duration is computable using the duration of the tactus rate current at the time of each beat.

Allied to the transcription task is the task of quantization—correcting for non-intentional performance errors when recording a musician, typically from MIDI input. Traditional approaches have attempted to align to a metrical grid, while a more powerful context sensitive connectionist approach has been demonstrated by Desain and Honing [25]. In both methods there is no distinguishing between an intentional rubato, in particular quantizing over a ritard, where no beat would have a small ratio IOI relative to its neighbours, and non-intentional performance error.

Non-intentionality of expressive deviations can now be viewed as producing overly complex modulation of the tactus. A low pass filtering of the frequency modulation function, before reconstruction to the foot-tap, is proposed here as a means of quantization while preserving intended expressive timing. Specification of the actual filters are a future research task. In particular, there remains the question of whether low-pass filtering localised deviations such as agogic accents, which produce short term modulations of the tactus (Section 4.2.3) will destroy this information.

As Western rhythm theory defines rhythmic units into small integer subdivisions of a slowest interval, those subdivisions can be considered harmonics of the rhythmic fundamental. If harmonic rhythms are what is currently considered a quantized rhythm, assessing the harmonicity of the ridges to the tactus can produce a measure of expression in a performance. If it proves possible to accurately factor a rhythm into harmonic and inharmonic components, where harmonicity is with respect to a predominant tactus, it may be possible to remove inharmonic components that comprise gross timing errors.

6.3.2 Structure Models

While the tactus is important, there is still much work to be done in inducing other concepts of rhythmic structure. The scalogram, phasogram and their ridges form a representation of underlying structure of a rhythm. As demonstrated in Chapter 4, grouping appears as a combination of parallel, harmonic ridges. It is possible an identification and labelling of several ridges could form a preprocessing step to determine musical structure. Such a structure representation allows editing of the times of recorded events while respecting their tempo. For such a task, the note's nominal value should determine where, on the ridge, the tempo should be retrieved. Changing note values (crochet to quaver etc) would effectively be

selecting different ridges.

The multiresolution technique is appropriate to apply to the analysis of final ritards (Section 2.2.7). This simply requires the production of suitable data from performance reduced to monophonic lines. While much literature exists reporting analysis of final ritards, the data is yet to be made available. Inclusion of such data into DORYS seems very worthwhile.

Other forms of time-keeping extraction appear possible. For polyrhythmic African music, it may be possible to extract the bell-line, by relaxing the continuity condition and instead looking for characteristic short term ridges (which should be discontinous due to the bell-line's asymmetry and not forming an additive meter). Given the cultural requirement to understand the bell as the time-keeping instrument, a possible scenario is to produce an analysis of the bell-line rhythm, then attempt to analyse the accompanying polyrhythms performed on other instruments by matching against the ridge behaviour of the bell.

6.3.3 Parallel Stream Segregation

The wavelet analyses have been over a single voice, drum or ensemble instrument, where timbre, pitch and spatialisation are conveniently assumed to provide an interrelated discriminator in the mind of the listener between instruments. Clearly the work of stream segregation (for example Brown and Cooke [11]), and specifically multiresolution approaches by McDonald [111], Schreirer [154] and Tait [175] needs to be tested as a preprocessor to multiresolution rhythm analysis. It is possible a full model would provide top down feedback from rhythmic structure to aid in stream segregation.

Assuming adequate stream segregation, further work needs to be done relating between multiresolution analysis of streams, as visualised in Figure 21. A 2-D wavelet transform typically used for vision seems initially appropriate to relate temporal coincidence between polyphonic events ("vertical timing", to use Desain and Honing's term [28]). This would allow dealing with temporal asynchrony between notes of a chord, voices in a choir, percussion instruments and so forth. However the dimension of polyphonic depth does not seem to have the relationship to the original signal that 2-D wavelets have when applied to the original image signal. This issue warrants considerable research.

6.3.4 Real-Time Operation

To achieve real-time accompaniment with the continuous wavelet model requires addressing the non-causality of the Morlet wavelet and the enculturated database of rhythmic examples (veridical expectancies) listeners use. It also requires a real-time unambiguous determination of the tactus.

The current Morlet transform is unable to run in real time due to the use of the convolution operator, which operates over the entire analysis window. Holschneider and others "algorithme á trous" implementation of the Morlet wavelet [58] as cascaded filter banks enables real-time operation. Obviously a period of time must pass before the lowest scales can be analysed, but this is a natural constraint on human listeners too. This requires establishing a lowest scale, that is, a maximum time interval that analysis will extend to. A first approximation seems to be a period of 2–5 seconds, conforming with the limits of the subjective present.

An assessment of the need for an enculturation database needs to be made, and what form it should take. This could take the form of changing the wavelet domain responses, or modifying the response following analysis.

The next issue is to develop a different approach to determining the tactus. This requires modelling of the constraints and behaviour of human perception and performance. It is likely there would need to be a limit to the memory of the tactus. Likewise, a measure of inertia to change from the current pulse rate would be needed (in a similar manner to the current continuity condition of the tactus algorithm). The weighted influence of absolute tempo constraints (frequencies) should be used. It may be possible to build from examples of performed rhythms to discover what transformations are possible.

6.3.5 Other Wavelets

As detailed in Chapter 3, Morlet wavelets have the best simultaneous time and frequency resolution with respect to the Heisenberg inequality. It would be instructive, however, to investigate the use of other wavelets for analytical purposes.

One candidate is Todd's one-dimensional version of Marr's Sombrero wavelet [103, 95]. However as noted in Chapters 3 and 5, the lack of an independent phase measure and reconstructive capabilities would limit the application of that wavelet. Todd's wavelet has similar characteristics to Solbach's Gammatone wavelet [168] which has a characterisation of the degree to which it deviates from the Gaussian

envelope, and has an independent phase.

Interval accents as demonstrated by Todd [103, 110] can be seen to be an artifact of the tail of the Marr wavelet, as such, accent effects on the last beat of an anapest group are not present with the non-causal Morlet wavelet. However, it is unclear whether this effect is tempo dependent in Todd's model. Tempo constraints are yet to be introduced in the multiresolution rhythm model as described in Sections 5.6.3 and 6.3.4.

Some design criteria of a constructed wavelet in order to be more applicable to rhythm analysis would be to minimise interaction between ridges and to quantify, and better control, the secondary and ternary oscillations of the wavelet. While the second criteria is trivial to achieve by changing the Gaussian envelope, reducing ridge interaction further seems a more challenging task.

As noted in Sections 5.6.2, 3.3.1 and 6.1.3, the reconstruction of Morlet wavelets is not perfect. While imperfect reconstruction has not hindered tactus determination, experimentation with bi-orthogonal wavelet transformations may be rewarding. These wavelets are capable of preserving phase, while reducing the representation redundancy and allow perfect reconstruction [185].

Bibliography

[1] Apple Computer Inc. *Audio Interchange File Format AIFF-C.* Apple Computer Inc., Cupertino, California, 1991.

[2] R. Ashley. Aspects of expressive timing in jazz ballad performance. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 485–90, Montreal, Quebec, 1996. Faculty of Music, McGill University.

[3] D. Bailey. *Improvisation: Its Nature And Practice In Music.* Da Capo Press, New York, second edition, 1992. 146p.

[4] I. Bengtsson and A. Gabrielsson. Analysis and synthesis of musical rhythm. In J. Sundberg, editor, *Studies of Music Performance*, volume 39, pages 27–60. Royal Swedish Academy of Music, Stockholm, 1983.

[5] J. J. Bharucha. MUSACT: A connectionist model of musical harmony. In S. M. Schwanauer and D. A. Levitt, editors, *Machine Models of Music*, pages 497–510. MIT Press, Cambridge, Mass, 1993.

[6] J. A. Bilmes. A model for musical rhythm. In *Proceedings of the International Computer Music Conference*, pages 207–10. International Computer Music Association, 1992.

[7] J. A. Bilmes. Techniques to foster drum machine expressivity. In *Proceedings of the International Computer Music Conference*, pages 276–83. International Computer Music Association, 1993.

[8] J. A. Bilmes. 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, September 1993.

[9] B. Boashash. Time-frequency signal analysis. In S. Haykin, editor, *Advances in Spectrum Analysis and Array Processing*, pages 418–517. Prentice-Hall, Englewood Cliffs, NJ, 1990.

[10] A. S. Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge, Mass, 1990. 773p.

[11] G. J. Brown and M. Cooke. Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research*, 23(2):107–32, 1994.

[12] J. C. Brown. Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America*, 94(4):1953–7, 1993.

[13] J. Cage. *Silence*. Marion Boyars, 1961.

[14] E. F. Clarke. Structure and expression in rhythmic performance. In *Musical Structure and Cognition*, chapter 9, pages 209–37. Academic Press, London, 1985.

[15] E. F. Clarke. Levels of structure in the organization of musical time. *Contemporary Music Review*, 2(1):211–38, 1987.

[16] M. Clynes. Secrets of life in music: Musicality realised by computer. In *Proceedings of the International Computer Music Conference*, pages 225–32. International Computer Music Association, 1984.

[17] G. L. Collier. The swing rhythm in jazz. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 477–80, Montreal, Quebec, 1996. Faculty of Music, McGill University.

[18] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

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

[20] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992. 357p.

[21] N. Delprat, B. Escudié, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, and B. Torrésani. Asymptotic wavelet and Gabor analysis: Extraction of instantaneous frequencies. *IEEE Transactions on Information Theory*, 38(2):644–64, 1992.

[22] N. Delprat, P. Guillemain, and R. Kronland-Martinet. Parameters estimation for non-linear resynthesis methods with the help of a time-frequency analysis of natural sounds. In *Proceedings of the International Computer Music Conference*, pages 88–90. International Computer Music Association, 1990.

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

[24] 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.

[25] P. Desain and H. Honing. The quantization of musical time: A connectionist approach. In P. M. Todd and D. G. Loy, editors, *Music and Connectionism*, pages 150–67. MIT Press, Cambridge, Mass, 1991.

[26] P. Desain and H. Honing. Tempo curves considered harmful. *Array: Journal of the ICMA*, 11(3), 1991. Spans 11(3),11(4),12(1).

[27] P. Desain and H. Honing. Tempo curves considered harmful. *Time in Contemporary Musical Thought, Contemporary Music Review*, 7(2):123–138, 1991.

[28] P. Desain and H. Honing. Tempo curves considered harmful: A critical review of the representation of timing in computer music. In *Proceedings of the International Computer Music Conference*, pages 143–9, 1991.

[29] P. Desain and H. Honing. Towards a calculus for expressive timing in music performance. *Computers in Music Research*, 3:43–120, 1991.

[30] P. Desain and H. Honing. Does expressive timing in music performance scale proportionally with tempo? *Psychological Research*, 56:285–292, 1994.

[31] P. Desain and H. Honing. Foot-tapping: A brief introduction to beat induction. In *Proceedings of the International Computer Music Conference*, pages 78–9. International Computer Music Association, 1994.

[32] P. Desain and H. Honing. Physical motion as a metaphor for timing in music: The final ritard. In *Proceedings of the International Computer Music Conference*, pages 458–60. International Computer Music Association, 1996.

[33] P. Desain and H. Honing. A reply to S. W. Smoliar's "modelling musical perception: A critical view". In N. Griffith and P. M. Todd, editors, *Musical Networks: Parallel Distributed Perception and Performance*, pages 111–4. MIT Press, Cambridge, Mass, 1999.

[34] C. Dodge and T. Jerse. *Computer Music: Synthesis, Composition, and Performance*. Schirmer Books, New York, 1985. 383p.

[35] W. J. Dowling and D. L. Harwood. *Music Cognition*. Academic Press, Orlando, Fl, 1986. 258p.

[36] W. Duckworth and R. Fleming, editors. *Sound and Light: La Monte Young, Marian Zazeela*, volume 40 of *Bucknell Review*. Bucknell University Press, Lewisburg, PA., 1996. 231p.

[37] B. Escudié, A. Grossmann, R. Kronland-Martinet, and B. Torrésani. Représentation en ondelettes de signaux asymptotiques: Emploi de la phase stationnaire. In *Proceedings Colloque GRETSI*, 1989. (In French).

[38] J. Feldman, D. Epstein, and W. Richards. Force dynamics of tempo change in music. *Music Perception*, 10(2):185–204, 1992.

[39] J. D. Foley and A. Van Dam. *Fundamentals of Interactive Computer Graphics*. Addison-Wesley, 1st edition, 1984. 664p.

[40] C. Folio. An analysis of polyrhythm in selected improvised jazz solos. In E. W. Marvin and R. Hermann, editors, *Concert Music, Rock, and Jazz since 1945: Essays and Analytical Studies*, pages 103–34. University of Rochester Press, New York, 1995.

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

[42] A. Friberg and J. Sundberg. Time discrimination in a monotonic isochronous sequence. *Journal of the Acoustical Society of America*, 98(5):2524–31, 1995.

[43] J. Frigyesi. Preliminary thoughts toward the study of music without clear beat: The example of "flowing rhythm" in Jewish Nusah. *Asian Music*, 24(2):59–88, 1993.

[44] D. Gabor. Theory of communication. *IEE Proceedings*, 93(3):429–57, Nov 1946.

[45] A. Gabrielsson. Once again: The theme from Mozart's piano sonata in A major (K.331): A comparison of five performances. In A. Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 81–104. Royal Swedish Academy of Music, Stockholm, 1987.

[46] K. Gann. *The Music of Conlon Nancarrow*. Cambridge University Press, 1995. 303p.

[47] M. Gasser and D. Eck. Representing rhythmic patterns in a network oscillators. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 361–6, Montreal, Quebec, August 1996. Faculty of Music, McGill University.

[48] A. Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2), 1995. available via http://www.best.com/~agraps/current/wavelet.html.

[49] P. Griffiths. *Modern Music: The Avant Garde Since 1945*. George Braziller Inc, London, 1981. 331p.

[50] 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.

[51] 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.

[52] A. Grossmann and J. Morlet. Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math.*, 15:723–36, 1984.

[53] P. Guillemain and R. Kronland-Martinet. Characterization of acoustic signals through continuous linear time-frequency representations. *Proceedings of the IEEE*, 84(4):561–85, 1996.

[54] S. L. Hahn. Hilbert transforms. In A. Poularikas, editor, *The Transforms And Applications Handbook*, chapter 7, pages 463–628. CRC Press, Boca Raton Fl., 1996.

[55] S. Handel. *Listening: An Introduction To The Perception Of Auditory Events*. MIT Press, Cambridge, Mass, 1989. 597p.

[56] S. Handel and G. Lawson. The contextual nature of rhythmic interpretation. *Perception and Psychophysics*, 34:103–20, 1983.

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

[58] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In J. Combes, A. Grossman, and P. Tchamitchian, editors, *Wavelets, Time-Frequency Methods and Phase Space*, pages 286–97, New York, 1989. Springer-Verlag.

[59] H. Honing. POCO: An environment for analysing, modifying, and generating expression in music. In *Proceedings of the International Computer Music Conference*, pages 364–8. International Computer Music Association, 1990.

[60] H. Honing. Expresso, a strong and small editor for expression. In *Proceedings of the International Computer Music Conference*, pages 215–18. International Computer Music Association, 1992.

[61] H. Honing. Issues on the representation of time and structure in music. In I. Cross and I. Deliège, editors, *Proceedings of the 1990 Music and Cognitive Sciences Conference, Contemporary Music Review*, volume 9, pages 221–38. Harwood Press, London, 1993.

[62] A. Houtsma, T. Rossing, and W. Wagenaars. *Auditory Demonstrations on Compact Disc*. Philips, Acoustical Society of America, 1987. 91p. (Compact Disc 1126-061).

[63] International MIDI Association. *MIDI Musical Instrument Digitial Interface Specification 1.0*. International MIDI Association, Los Angeles, 1983.

[64] M. R. Jones. Time, our lost dimension: Toward a new theory of perception, attention and memory. *Psychological Review*, 83(5):323–55, 1976.

[65] M. R. Jones. Attentional rhythmicity in human perception. In J. R. Evans and M. Clynes, editors, *Rhythm in Psychological, Linguistic, and Musical Processes*, chapter 2, pages 13–40. Charles Thomas Publishers, Springfield, Ill., 1986.

[66] M. R. Jones. Dynamic pattern structure in music: Recent theory and research. *Perception and Psychophysics*, 41(6):621–34, 1987.

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

[68] M. R. Jones, M. Boltz, and G. Kidd. Controlled attending as a function of melodic and temporal context. *Perception and Psychophysics*, 32(3):211–8, 1982.

[69] M. Kennedy. *The Concise Oxford Dictionary of Music*. Oxford University Press, Oxford, third edition, 1980.

[70] H. I. Khan. *The Mysticism of Sound and Music*. Shambhala, Boston, revised edition, 1996. 322p.

[71] J. Koetting. What do we know about African rhythm? *Ethnomusicology*, 30:58–63, 1986.

[72] R. Kostelanetz. *John Cage*. Da Capo Press, 1991. 239p.

[73] P. Kovesi. Image features from phase congruency. Technical Report 95/4, Computer Science Department, University of Western Australia, March 1995.

[74] R. Kronland-Martinet. The use of the wavelet transform for the analysis, synthesis and processing of speech and music sounds. *Computer Music Journal*, 12(4):11–20, 1988. (Sound examples on soundsheet with 13(1) 1989).

[75] 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.

[76] R. Kronland-Martinet, J. Morlet, and A. Grosmann. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 1(2):273–302, 1987.

[77] U. Kronman and J. Sundberg. Is the musical ritard an allusion to physical motion? In A. Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 57–68. Royal Swedish Academy of Music, Stockholm, 1987.

[78] C. L. Krumhansl. Music psychology: Tonal structures in perception and memory. *Annual Review of Psychology*, 42:277–303, 1991.

[79] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60:115–42, 1992.

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

[81] 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.

[82] C. S. Lee. The rhythmic interpretation of simple musical sequences: Towards a perceptual model. In P. Howell, I. Cross, and R. West, editors, *Musical Structure and Cognition*, chapter 3, pages 53–69. Academic Press, London, 1985.

[83] M. Leman. Schema-based tone center recognition of musical signals. *Journal of New Music Research*, 23(2):169–204, 1994.

[84] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, Mass, 1983. 368p.

[85] J. London. The binary bias of metric subdivision and the relative complexity of various meters, or why is 9/8 so rare? In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 357–60, Montreal, Quebec, 1996. Faculty of Music, McGill University.

[86] H. C. Longuet-Higgins. Perception of melodies. *Nature (London)*, 263:646–53, 1976. 544p.

[87] H. C. Longuet-Higgins. The perception of music. *Interdisciplinary Science Reviews*, 3:148–56, June 1976.

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

[89] H. C. Longuet-Higgins and E. R. Lisle. Modelling musical cognition. *Contemporary Music Review*, 3(1):15–27, 1989.

[90] D. G. Loy. Musicians make a standard: The MIDI phenomenon. *Computer Music Journal*, 9(4):8–26, 1985.

[91] T. Machover. Hyperinstruments: A progress report 1987–1991. MIT media laboratory internal memo, Massachusetts Institute of Technology, 1992.

[92] S. Macpherson. *Rudiments of Music*. Galliard Ltd, New York, second edition, 1970.

[93] J. M. Magill and J. L. Pressing. Asymmetric cognitive clock structures in West African rhythms. *Music Perception*, 15(2):189–222, 1997.

[94] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998. 577p.

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

[96] D. Massaro. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 79:124–45, 1972.

[97] N. P. McAngus Todd. A model of expressive timing in tonal music. *Music Perception*, 3(1):33–58, 1985.

[98] N. P. McAngus Todd. A computational model of rubato. *Contemporary Music Review*, 3:69–88, 1989.

[99] N. P. McAngus Todd. Towards a cognitive theory of expression: The performance and perception of rubato. *Contemporary Music Review*, 4(405–16), 1989.

[100] N. P. McAngus Todd. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, 91(6):3540–50, 1992.

[101] N. P. McAngus Todd. Multi-scale analysis of expressive signals: Recovery of structure and motion. In A. Friberg, J. Iwarsson, E. Jansson, and J. Sundberg, editors, *Proceedings of the Stockholm Music Acoustics Conference*, 79, pages 146–9. Royal Swedish Academy of Music, 1993.

[102] N. P. McAngus Todd. Vestibular feedback in musical performance: Response to: *Somatosensory Feedback in Musical Performance. Music Perception*, 10(3):379–82, 1993.

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

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

[105] N. P. McAngus Todd. Toward a theory of the central auditory system I: Architecture. In *Fourth International Conference on Music Perception and Cognition*, pages 173–8, Montreal, Canada, 1996. Faculty of Music, McGill University.

[106] N. P. McAngus Todd. Toward a theory of the central auditory system II: Pitch. In *Fourth International Conference on Music Perception and Cognition*, pages 179–84, Montreal, Canada, 1996. Faculty of Music, McGill University.

[107] N. P. McAngus Todd. Toward a theory of the central auditory system III: Time. In *Fourth International Conference on Music Perception and Cognition*, pages 185–90, Montreal, Canada, 1996. Faculty of Music, McGill University.

[108] N. P. McAngus Todd. Toward a theory of the central auditory system IV: Grouping. In *Fourth International Conference on Music Perception and Cognition*, pages 191–6, Montreal, Canada, 1996. Faculty of Music, McGill University.

[109] N. P. McAngus Todd and E. F. Clarke. The perception of rhythmic structure in expressive musical performance. In *Proceedings of the 15th International Congress of Acoustics*, volume III, pages 459–462, 1995.

[110] N. P. McAngus Todd and C. S. Lee. An auditory model account of interval produced accents: Experimental evidence. (manuscript to appear), 1996.

[111] S. McDonald. Biologicalesque transcription of percussion. In *Proceedings of the Australian Computer Music Conference*, pages 31–8, Canberra, 1998.

[112] L. B. Meyer. *Emotion and Meaning in Music.* University of Chicago Press, 1956. 307p.

[113] J. A. Michon. The complete time experiencer. In J. Michon and J. Jackson, editors, *Time, Mind, and Behaviour*, pages 21–52. Springer Verlag, Berlin, 1985.

[114] B. O. Miller, D. L. Scarborough, and J. A. Jones. On the perception of meter. In M. Balaban, K. Ebcioglu, and O. E. Laske, editors, *Understanding Music with AI*, pages 428–47. MIT Press, Cambridge, Mass, 1992.

[115] G. Miller. The magical number seven, plus or minus two: Some limits of our capacity for processing information. *Psychological Review*, 63:81–97, 1956.

[116] M. Minsky. Music, mind and meaning. *Computer Music Journal*, 5(3):28–44, 1981.

[117] P. Moisala. Cognitive study of music as culture – basic premises for "cognitive ethnomusicology". *Journal of New Music Research*, 24(1):8–20, 1995.

[118] F. R. Moore. The dysfunctions of MIDI. *Computer Music Journal*, 12(1):19–28, 1988.

[119] F. R. Moore. *Elements of Computer Music.* Prentice-Hall, New Jersey, 1990. 560p.

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

[121] C. Muir and K. McMillen. What's missing in MIDI? *Guitar Player*, pages 61–2, June 1986.

[122] K. Ohya. A rhythm perception model by neural rhythm generators. In *Proceedings of the International Computer Music Conference*, pages 129–30. International Computer Music Association, 1994.

[123] C. Palmer. Mapping musical thought to musical performance. *Journal of Experimental Psychology - Human Perception and Performance*, 15(12):331–46, 1989.

[124] C. Palmer. Structural representations of music performance. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, Hillsdale, NJ, 1989. Erlbaum Associates.

[125] C. Palmer and C. L. Krumhansl. Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology - Human Perception and Performance*, 13(1):116–26, 1987.

[126] C. Palmer and C. L. Krumhansl. Pitch and temporal contributions to musical phrase perception—effects of harmony, performance timing, and familiarity. *Perception and Psychophysics*, 41(6):505–18, 1987.

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

[128] R. Parncutt. The perception of pulse in musical rhythm. In A. Gabrielsson, editor, *Action and Perception in Rhythm and Music*, volume 55, pages 127–38. Royal Swedish Academy of Music, Stockholm, 1987.

[129] R. Parncutt. A model of beat induction accounting for perceptual ambiguity by continuously variable parameters. In *Proceedings of the International Computer Music Conference*, pages 83–4. International Computer Music Association, 1994.

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

[131] R. Parncutt. Template-matching models of musical pitch and rhythm perception. *Journal of New Music Research*, 23(2):145–67, 1994.

[132] H. Partch. *Genesis of a Music*. Da Capo Press, New York, second edition, 1974. 517p.

[133] S. T. Pope. A time-oriented taxonomy of computer music. *Computer Music Journal*, 19(1):4, 1995.

[134] E. Pöppel. Time perception. In R. Held, H. W. Leibowitz, and H.-L. Teuber, editors, *Handbook of Sensory Physiology*, volume VIII: Perception, chapter 23, pages 713–29. Springer, Berlin, 1978.

[135] D.-J. Povel. Internal representation of simple temporal patterns. *Journal of Experimental Psychology - Human Perception and Performance*, 7(1):3–18, 1981.

[136] D.-J. Povel and P. Essens. Perception of temporal patterns. *Music Perception*, 2(4):411–40, 1985.

[137] D.-J. Povel and H. Okkerman. Accents in equitone sequences. *Perception and Psychophysics*, 30(6):565–72, 1981.

[138] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications*. Macmillan Publishing Co., New York, second edition, 1992. 969p.

[139] J. A. Prögler. Searching for swing: Participatory discrepencies in the jazz rhythm section. *Ethnomusicology*, 39(1):21–54, 1995.

[140] B. H. Repp. Further perceptual evaluations of pulse microstructure in computer performances of classical piano music. *Music Perception*, 8(1):1–33, 1990.

[141] B. H. Repp. Music as motion: A synopis of Alexander Truslit's (1938) "Gestaltung und bewengung in der musik". *Psychology of Music*, 21(1):48–72, 1993.

[142] B. H. Repp. Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research*, 56(4):269–284, 1994.

[143] B. H. Repp. Quantitative effects of global tempo on expressive timing in music performance: Some perceptual evidence. *Music Perception*, 13(1):39–57, 1995.

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

[145] C. Roads. Research in music and artificial intelligence. *ACM Computing Surveys*, 17(2):163–90, 1985.

[146] S. Roberts and M. Greenhough. The detection of rhythmic repetition using a self-organising neural network. In *Proceedings of the International Computer Music Conference*, pages 125–8. International Computer Music Association, 1994.

[147] D. Rosenthal. Emulation of human rhythm perception. *Computer Music Journal*, 16(1):64–76, 1992.

[148] D. F. Rosenthal. *Machine Rhythm: Computer Emulation of Human Rhythm Perception*. PhD thesis, MIT Media Lab, Cambridge, Mass, August 1992.

[149] R. Rowe. *Interactive Music Systems*. MIT Press, Cambridge, Mass, 1992. 278p.

[150] R. Rowe. Machine listening and composing with Cypher. *Computer Music Journal*, 16(1):43–63, 1992.

[151] G. Rule. Keyboard report: Korg Wavedrum. *Keyboard*, 21(3):72–7, 1995.

[152] D. E. Rumelhart, J. L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press, Cambridge, Mass, 1986.

[153] C. Sachs. *Rhythm and Tempo: A Study in Music History*, chapter 5: The Near and Middle East, pages 83–97. J.M. Dent and Sons Ltd, London, 1953.

[154] E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, 103(1):588–601, 1998.

[155] U. Seifert. Time constraints and cognitive modelling. In J. Laaksamo and J. Louhivuori, editors, *Proceedings of the First International Conference on Cognitive Musicology*, pages 288–99, Finland, 1993. Department of Music, University of Jyväskylä.

[156] U. Seifert, F. Olk, and A. Schneider. On rhythm perception: Theoretical issues, empirical findings. *Journal of New Music Research*, 24(2):164–95, 1995.

[157] L. Shaffer. Performances of Chopin, Bach and Bartok: Studies of motor programming. *Cognitive Psychology*, 13:326–76, 1981.

[158] J. A. Sloboda. The communication of musical meter in piano performance. *Quarterly Journal of Experimental Psychology*, 35:377–96, 1983.

[159] J. A. Sloboda. *The Musical Mind: The Cognitive Psychology of Music*. Clarendon Press, Oxford, 1985. 291p.

[160] L. M. Smith. Surveys for design criteria of interactive computer music performance systems. Post graduate diploma in computing science thesis, School of Computing Science, Curtin University of Technology, Bentley, Western Australia, 1991.

[161] L. M. Smith. Listening to musical rhythms with progressive wavelets. In *Proceedings of Tencon '96: Digital Signal Processing Applications*, volume 2, pages 508–13. IEEE, 1996.

[162] L. M. Smith. Modelling rhythm perception by continuous time-frequency analysis. In *Yanchep '96: Proceedings of Department Research Conference*, pages 1–20. Department of Computer Science, University of Western Australia, 1996.

[163] L. M. Smith. Modelling rhythm perception by continuous time-frequency analysis. In *Proceedings of the International Computer Music Conference*, pages 392–5. International Computer Music Association, 1996. http://www.cs.uwa.edu.au/~leigh/Research/Papers/ICMC96.ps.Z.

[164] L. M. Smith. Application of ridge extraction to tactus determination. In C. MacNish, A. Czarn, and P. Taylor, editors, *Proceedings of the Ninth University of Western Australia Computer Science Research Conference*, pages

173–8. Department of Computer Science, University of Western Australia, 1998.

[165] 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*, pages 197–202, Montreal, Quebec, August 1996. Faculty of Music, McGill University. http://www.cs.uwa.edu.au/~leigh/Research/Papers/ICMPC96.ps.Z.

[166] S. Smoliar. Mental structures. *Array: Journal of the ICMA*, 12(3):8–9, 1992.

[167] S. W. Smoliar. Modelling musical perception: A critical view. *Connection Science*, 6(2+3):209–22, 1994.

[168] 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*, Montreal, Canada, August 1995. Preprint available via ftp://ftp.ti6.tu-harburg.de/pub/paper.

[169] N. Sorrell and R. Narayan. *Indian Music In Performance: A Practical Introduction*. Manchester University Press, 1980. 190p.

[170] E. S. Spelke. Exploring audible and visible events in infancy. In A. Pick, editor, *Perception and its Development*, pages 221–35. Erlbaum, Hillsdale, NJ, 1979.

[171] M. J. Steedman. The perception of musical rhythm and metre. *Perception*, 6:555–69, 1977.

[172] S. Sternberg and R. L. Knoll. Perception, production and imitation of time ratios by skilled musicians. In R. Aiello and J. A. Sloboda, editors, *Musical Perceptions*, chapter 10, pages 240–57. Oxford University Press, Oxford, 1994.

[173] J. Sundberg, A. Askenfelt, and L. Frydén. Musical performance: A synthesis-by-rule approach. *Computer Music Journal*, 7(1):37–43, 1983.

[174] J. Sundberg and V. Verrillo. On the anatomy of the retard: A study of timing in music. *Journal of the Acoustical Society of America*, 68:772–9, 1980.

[175] C. Tait. Audio analysis for rhythmic structure. In *Proceedings of the International Computer Music Conference*, pages 590–1. International Computer Music Association, 1995.

[176] E. Tamm. *Brian Eno: His Music and the Vertical Color of Sound*. Faber and Faber, 1989. 223p.

[177] A. S. Tanguiane. *Artificial Perception and Music Recognition*. Number 746 in Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, 1993.

[178] H. Taube. Common Music: A music composition language in Common LISP and CLOS. *Computer Music Journal*, 15(2):21–32, 1991.

[179] H. Taube. An object-oriented representation for musical pattern definition. *Journal of New Music Research*, 24(2):121–29, 1995.

[180] P. Tchamitchian and B. Torrésani. Ridge and skeleton extraction from the wavelet transform. In M. B. Ruskai, editor, *Wavelets and Their Applications*, pages 123–51. Jones and Bartlett Publishers, Boston, Mass., 1992.

[181] J. Tenney and L. Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–41, 1980.

[182] P. M. Todd and D. G. Loy, editors. *Music and Connectionism*. MIT Press, Cambridge, Mass, 1991. 268p.

[183] P. Toiviainen. An interactive MIDI accompanist. *Computer Music Journal*, 22(4):63–75, 1998.

[184] H. H. Touma. *The Music of the Arabs*. Amadeus Press, 1996. 238p.

[185] M. Vetterli and C. Herley. Wavelets and filter banks: Theory and design. *IEEE Transactions on Signal Processing*, 40(9):2207–32, 1992.

[186] B. Vidaković and P. Müller. Wavelets for kids: A tutorial introduction. Technical report, Duke University, 1991.

[187] P. G. Vos. Temporal duration factors in the perception of auditory rhythmic patterns. *Scientific Aesthetics*, 1:183–99, 1977.

[188] G. Widmer. Learning expressive performance: The structure-level approach. *Journal of New Music Research*, 25(2):179–205, 1996.

[189] S. R. Wilkinson. *Tuning In: Microtonality in Electronic Music*. Hal Leonard Books, Milwaukee, WI, 1988. 120p.

[190] H. Woodrow. Time perception. In S. S. Stevens, editor, *Handbook of Experimental Psychology*, chapter 32, pages 1224–36. Wiley and Sons, New York, 1951.

[191] M. Yako. The hierarchical structure of time and meter. *Computer Music Journal*, 21(1):47–57, 1997.

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

Colophon

Preparation

This thesis was prepared on the teTeX distribution of LaTeX2e substituted with Basil K. Malyshev's **BaKoMa** and Adobe Sonata, Calliope and ZapfDingbat Postscript fonts, running NeXT/Apple's OpenStep V4.2 on a Pentium 166MHz clone. The DVI file was converted with dvi2ps and then to PDF with Frank Siegert's **PStill**. The diagrams were generated with Diagram! V2.0 and Mathematica routines, and music notation with William Clocksin's **Calliope**. I couldn't resist itemize lists with Dingbat symbols.

The Sound File examples

This thesis is presented in both printed and softcopy forms, the latter as an Adobe Acrobat PDF file. This includes links to audio examples of the rhythms encoded in AIFF [1] format. These were generated with Rick Taube's Common Music system, substituting an enveloped sinusoid instrument for each impulse, and using 44.1KHz sample rate. The timing and intensities are identical to the analysed examples. Clicking on the link will play the example file.

Sources

Online versions of this thesis and software are available from:

- `http://www.cs.uwa.edu.au/~leigh/Research/Thesis.tar.gz`
- `http://www.cs.uwa.edu.au/~leigh/Research/Software/DORYS.tar.gz`
  The "Database Of RYthmic Stimuli" written in Common Music, described in Section 4.1.

- `http://www.cs.uwa.edu.au/~leigh/Research/Software/MultiresRhythm.tar.gz`
  The software performing analysis and plotting of musical rhythms and producing Common Music score files of foot-tapping rhythms described in Chapters 4–5.