A review of automatic rhythm description systems

Fabien Gouyon, Simon Dixon

18th October 2004

[draft version, to be published in Computer Music Journal 29:1, 2005]

Rhythm belongs with harmony, melody and timbre as one of the most fundamental aspects of music. Sound by its very nature is temporal, and in its most generic sense, the word *rhythm* is used to refer to all of the temporal aspects of a musical work, whether represented in a score, measured from a performance, or existing only in the perception of the listener. In order to build a computer system capable of intelligently processing music, it is essential to design representation formats and processing algorithms for the rhythmic content of the music.

Computer systems reported in the literature offer different interpretations of the words "automatic rhythm description" as they address diverse applications such as tempo induction, beat tracking, quantisation of performed rhythms, meter induction and characterisation of intentional timing deviations. Although some rhythmic concepts are consensual, no single representation of rhythm has been devised which would be suitable for all applications. In this paper, we propose a unifying framework for automatic rhythm description systems, and review existing systems with respect to the functional units of the proposed framework.

1 Representing musical rhythm

A naive approach to describe the rhythm of musical data (whether audio or symbolic) is to specify an exhaustive and accurate list of onset times, maybe together with some other musical features characterising those events (e.g., durations, pitches and intensities in the MIDI representation). However, such a representation lacks abstraction. There is more to rhythm than the absolute timings of successive musical events. There seems to be agreement on the fact that, in addition, one must also take into account the metrical structure, tempo and timing (Honing, 2001). However, there is no consensus regarding explicit representations of these three rhythmic concepts.

A first reason is that different rhythmic features are relevant at each step in the musical communication chain, at each step where rhythmic content is produced, transmitted and/or received. As we illustrate in the next sections, metrical structure, tempo and timing take slightly different meanings for composers, performers and listeners. Indeed, even if a goal in the field of music psychology is to seek representational elements, or processes, that would stand as "universal" or "innate" (i.e. functioning at birth, independent of environmental influence) (Drake and Bertrand, 2001), a more widespread objective is to determine differences in perception according to listeners' culture, musical background, age or sex (Drake, 1993; Lapidaki, 2000; Gabrielsson, 1973; Drake et al., 2000b).

A second reason is that the diverse media used for rhythm transmission suffer a trade-off between the level of abstraction and the comprehensiveness of the representation. Standard Western music notation provides an accepted method for

communicating a composition to a performer, but it has little value in representing the interpretation of a work as played in a concert. On the other hand, a MIDI file might be able to represent important aspects of a performance, but it does not provide the same level of abstraction as the score. At the extreme end, an acoustic signal implicitly contains all rhythmic aspects but provides no abstraction whatsoever. In an application context, the choice of a suitable representation is based on the levels of detail (respectively abstraction) of the various aspects of music which are provided by the representation.

1.1 Metrical structure

Western music notation provides an objective regular temporal structure underlying musical event occurrences and organising them into a hierarchical metrical structure. This is independent of the hierarchical phrase structure which may be explicit in the notation or implicit in the composer's, the performer's and/or the listener's conceptualisation of the music.

The Generative Theory of Tonal Music (GTTM, Lerdahl and Jackendoff, 1983) formalises this distinction by defining rules for a "musical grammar" which deals separately with grouping structure (phrasing) and metrical structure. While the grouping structure deals with time spans (durations), the metrical structure deals with durationless points in time, the *beats*, which obey the following rules. Beats must be equally spaced. A division according to a specific duration corresponds to a *metrical level*. Several levels coexist, from low levels (small time divisions) to high levels (longer time divisions). There must be a beat of the metrical structure for every note in a musical sequence. A beat at a high level must also be a beat at each lower level. At any metrical level, a beat which is also a beat at the next higher level is called a downbeat, and other beats are called upbeats. Beats obey a discrete time grid, with time intervals all being multiples of a common duration, the smallest metrical level.

Music psychology research asserts that humans perceive at least part of the objective temporal structure. Drake and Bertrand (2001) advocate a universal "predisposition toward simple duration ratio", and claim that "we tend to hear a time interval as twice as long or short as previous intervals." The Dynamic Attending Theory (Drake et al., 2000a; Jones and Boltz, 1989) proposes that humans spontaneously focus on a "referent level" of periodicity, and they can later switch to other levels to track events occurring at different time spans (for instance, longer-span harmony changes, or a particular shorter-span fast motive). However, metrical structure perception is strongly dependent on musical training (Drake et al., 2000a).

1.2 Tempo

Given a metrical structure, *tempo* is defined as the rate of the beats at a given metrical level, for example the quarter note level in the score. There is usually a *preferred* or *primary metrical level*, which corresponds to the rate at which most people would tap or clap in time with the music, and this is commonly used to define the tempo, expressed either as a number of beats per minute, or as the time interval between beats (the *inter-beat interval*). In many cases the primary metrical level corresponds to the denominator of the time signature, and the next one or two higher levels are specified by the numerator of the time signature.

However, the perception of tempo exhibits a degree of variability. It is not always correct to assume that the denominator of the time signature corresponds to the "foot-tapping" rate, nor to the actual "physical tempo" that would be an inherent property of audio flows (Drake et al., 1999). Differences in human perception of tempo depend on age, musical training, musical preferences and general listening

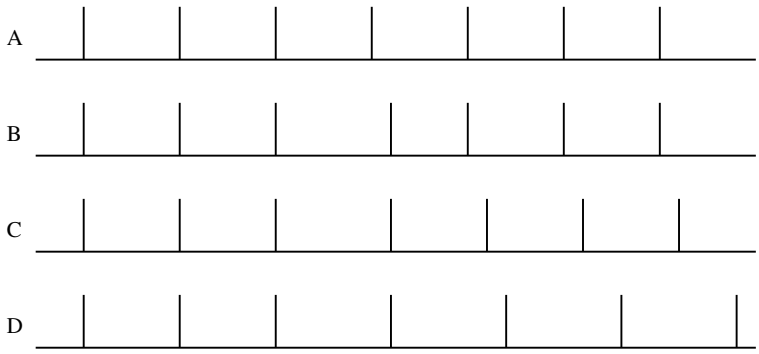

Figure 1: Four time-lines, marked with onsets, illustrating the difference between tempo and timing changes: (A) an isochronous pulse; (B) a local timing change; (C) a global timing change; and (D) a tempo change.

context (e.g. tempo of a previously heard sequence, listener's activity, instant of the day) (Drake, 1993; Drake et al., 2000a; Lapidaki, 2000; Drake et al., 2000b). Differences in tempo perception are nevertheless far from random; they most often correspond to a focus on a different metrical level, e.g. differences of half or twice the inter-beat interval (when hearing duple meter music) or one-third or three times the inter-beat interval (when hearing triple or compound meter music).

1.3 Timing

Although it is supposed to model the listener's intuitions, a major weakness of the GTTM is that it does not deal with the departures from strict metrical timing which are apparent in almost all styles of music. Thus it is only really suitable for representing the timing structures of musical scores, or as an abstract representation of a performance, where the expressive timing is not represented.

There are conceptually two types of non-metrical timing, which come under the headings *tempo* and *timing* respectively. These are illustrated in Figure 1, which shows a strictly metrical (isochronous) pulse (A), followed by three variations on this pulse. There are two types of timing changes: in the first case (B), just one beat in the pulse is displaced, whereas in the second case (C), all beats from a particular time onwards are displaced, as when a pause occurs in the music. In both of these cases, the change is in the timing; there is a discontinuity in the pulse, but the rate of the pulse on both sides of the discontinuity is the same. In this sense we can associate timing changes with short term changes in the pulse. On the other hand, a tempo change is a change in the rate of the pulse (D), which is a long term change in the pulse.

It is important to note that at the time of the first change (the 4th beat in the pulse), it is impossible to distinguish cases (B), (C) and (D). This makes causal analysis impossible (i.e. algorithms which do not use information about future events in analysing present events, as, for example, any real-time algorithm), since with no knowledge of the future, a single "out of time" beat could be due to either a tempo or timing change (Cambouropoulos et al., 2001).

One of the greatest difficulties in analysing performance data is that the two dimensions of tempo and timing are projected onto the single dimension of time. Mathematically, it is possible to represent any tempo change as a series of timing

changes and any timing change as a series of tempo changes, but these descriptions are somewhat counterintuitive (Honing, 2001). The parsimony of the representation is an important factor in its psychological plausibility (Tanguiane, 1993).

In order to represent changing tempi, various approaches can be used. If tempo is considered as an instantaneous value, it can be calculated as the inter-beat interval measured between each pair of successive beats. A more perceptually plausible approach is to take an average tempo measured over a longer period of time. A measure of central tendency of tempo over a complete musical excerpt is called the *basic tempo* (Repp, 1994), which is the implied tempo around which the expressive tempo varies. The end result of any of these approaches is a value of tempo as a function of time, which is called a tempo curve. Often, timing is also modelled by the tempo curve representation, an approach which is sharply criticised (Desain and Honing, 1991; Honing, 2001) for failing to separate the dimensions of tempo and timing. This criticism is well supported by examples where transformations applied to a tempo curve representation do not preserve musically important features.

Among others, Bilmes (1993) and Baggi (1991) propose to represent timing deviations as systematic event shifts occurring within the span of the fastest pulse, while keeping a constant execution speed. They found evidence of the suitability of such a representation in analysing respectively Latin percussion music and jazz music. Friberg and Sundström (2002,1999) propose to focus on the swing. The term originates in Jazz music, and is often characterised by the long-short pattern of performing consecutive eighth-notes. The *swing ratio* refers to a mathematical expression: the duration of the first eighth-note divided by that of the second.

Music psychology research presents evidence that listeners perceive performers' intentional timing deviations. Clarke (1987) shows that "categorical perception" differentiates expressive timing from rhythmic structure: a small number of categories are used to characterise the *continuously* variable temporal transformation of the *discrete* (integer ratio) structure. Further, timing and structure are tightly linked. Repp (1992) confirms listeners' sensitivity to timing deviations, but, most importantly, also shows that this sensitivity is a variable of the position in the metrical structure. Complementary to this finding, there is strong evidence that performers do not produce timing deviations at arbitrary points in time (Palmer, 1997). They rather deviate from pure mechanical performance in specific ways. The metrical structure provides "anchor points" for timing deviations, and "every aspect of musical structure contributes to the specification of an expressive profile for a piece" (Clarke, 1999, p.492). Expressive timing is also systematic; the timing in repeated performances can be very stable over a period of years (Clynes and Walker, 1982, pp.181-187).

2 A common analytical framework for rhythm description systems

The chief goal in automatic rhythm description is the parsing of acoustic events that occur in time into the more abstract notions of metrical structure, tempo and timing, as illustrated in Figure 2, where the goal is to derive a representation like (B) from (A) or (A'). A major difficulty is the inherent ambiguity of rhythm, as discussed in the previous section. This concern is also expressed by Parncutt (1994, p.423), Chung (1989, p.19) and Rosenthal (1992, p.12). This is a problem because computer implementations demand precise definitions, and any systematic comparison of program performances must be based on some "ground truth." For instance, it is difficult to compare programs that extract the tempo if their definitions of tempo do not explicitly refer to the same metrical level. Indeed, tempo induction

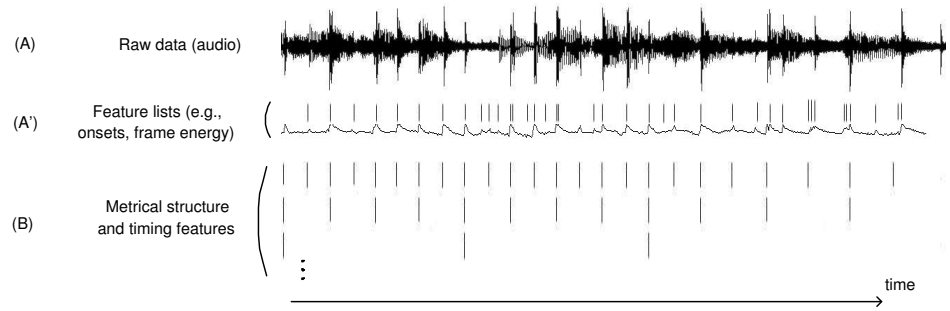

Figure 2: Example of an audio signal, examples of feature lists, corresponding metrical structure and timing features, showing a gradually decreasing tempo.

systems typically make errors of simple integer ratios, such as 60 BPM instead of 120 (Dixon, 2001; Goto and Muroaka, 1997). Also, even if existing scores can be taken as ground-truth references to the quantisation or time signature determination tasks, "correct" time signatures or quantised durations can always be the object of controversy. For this reason Cemgil et al. (2000) define music transcription as "the extraction of an *acceptable* ... music notation" (original emphasis).

The ambiguity of rhythm representations becomes apparent when we consider the following questions: Given a musical signal, how many metrical levels are relevant? Is there one most important level? Is there solely one *correct* perceptual tempo? Which metrical level defines the MM tempo of the music? Which metrical levels define the time signature? What are the relevant categories of timing deviations? In terms of Figure 2, what elements of (B) are relevant, and how can they be named and clearly defined? Are the answers to these questions common to all listeners?

These questions have no simple answers. There is no canonical form for representing rhythm, and lacking this ground truth, it is difficult, if not impossible, to provide a meaningful quantitative comparison of the various computer systems which each have different answers to these questions. Further, there is no common database on which the systems have been tested.

Some systems derive the beats and the tempo of just one metrical level, where this level is somewhat arbitrarily chosen. Others aim at deriving complete rhythmic transcriptions (i.e. scores) from musical performances. Still other programs aim at determining some timing features from musical performances, such as tempo changes, event shifts (timing changes) and swing factors.

These computer programs share some functional aspects. For instance, a prevalent aspect is the handling of symbolic data derived from (or instead of) raw audio data. These feature lists are usually made up of onset times (see (A') in Figure 2), which are sometimes used in conjunction with other features (temporal, timbral, harmonic or melodic). We define feature lists somewhat broadly, to include frame-based feature vectors as well as lists of symbolic events, since the algorithms subsequently used to process the lists are similar for both cases, even though the timescales differ by an order of magnitude. The distinction between high-level and low-level representations, although conceptually important, does not necessarily play a large role in determining the suitability of algorithms for the discovery of temporal patterns.

This review provides a qualitative comparison of systems with respect to the functional units of the general model illustrated in Figure 3, consisting of feature list creation (e.g., onset detection), pulse induction (including periodicity computation and handling of event shifts), pulse tracking, time signature and quantised duration

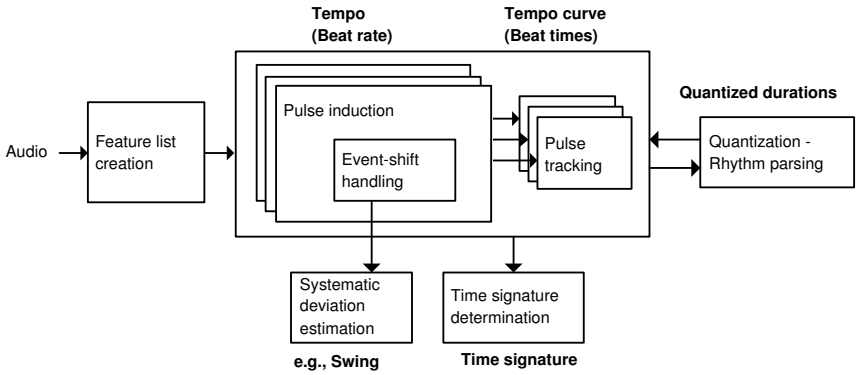

Figure 3: Functional units of rhythm description systems

determination and finally estimation of short term timing features. In the remainder of the paper, we discuss each of these functional units in turn.

3 Feature list creation

Some computer systems deal with symbolic data, such as MIDI or manually parsed scores containing solely onset times and durations (Brown, 1993; Longuet-Higgins and Lee, 1982). Recent systems tend to deal directly with acoustic signals or with compressed audio (Wang and Vilermo, 2001), although some early systems also used audio input (Schloss, 1985; Chowning et al., 1984). No matter what input data is used, the first analysis step is the creation of a feature list, i.e. the parsing, or "filtering", of the data at hand into a sequence of features. These features range from note onset features (as time, duration and amplitude) to frame-based signal features, and they are assumed to convey the predominant information relevant to rhythmic analysis.

In this step, monophonic excerpts are often parsed into sequences that resemble note features (e.g. onset time, duration and pitch). For polyphonic music, Allen and Dannenberg (1990) propose separating instrumental streams (a very challenging goal) and building a feature list for each monophonic stream, which is merged with other streams after some rhythmic analysis steps. Another possibility is to describe a polyphonic excerpt by a single feature list, giving a global (homophonic) view where features ("summary events" in Rosenthal, 1992, p.29) represent musical chunks such as chords or energy components.

Onset time The extraction of note onset times for rhythmic analysis is ubiquitous in the literature. Musical event occurrence instants are very important cues for rhythm perception. Onsets can be extracted (with more or less reliability) from virtually any musical format. For instance, Longuet-Higgins (1987) processes onsets manually parsed from scores. They can also be easily parsed from MIDI data (Cemgil et al., 2000; Raphael, 2002; Dixon and Cambouropoulos, 2000; Cemgil et al., 2001). More complex is their automatic extraction from audio signals, the details of which are out of the scope of this paper. An exhaustive overview of musical onset detection can be found in (Bello, 2003).

Duration In addition to onset times, some systems also handle durations, or alternatively inter-onset intervals (IOIs), which can be considered as roughly equivalent to durations (Brown, 1993), and are easier to compute for audio data. Like on-

sets, durations can be easily parsed from MIDI or scores, but can not be computed reliably from audio data (especially polyphonic music). Durations are extracted from scores by Brown (1993) and Longuet-Higgins and Lee (1982), whereas Mont-Reynaud and Goldstein (1985), Dannenberg and Mont-Reynaud (1987) and Allen and Dannenberg (1990) use durations derived from MIDI data to filter out "weak" onsets, being those onsets whose duration is either shorter than some fixed threshold (20 to 50 ms for Allen and Dannenberg, 1990) or much shorter than the preceding one. Chung's "note importance agencies" (1989, pp.61-62) and Temperley and Sleator's model (1999) also parse MIDI data into note onsets and durations. Parncutt (1994, p.426-432) weights onsets proportionally to their subsequent IOI, using a perceptually justified "saturation function." Perceptual experiments by Snyder and Krumhansl (2001) show that timing information alone (onsets and durations) is sufficient for the perception of a pulse in Ragtime music.

Relative Amplitude Relative amplitude is a factor which contributes to perceptual accentuation, and is easily computable from MIDI or audio data, but it is largely absent from score notation. Various systems use relative amplitude for weighting onsets derived from symbolic (Smith and Kovesi, 1996; Smith, 1996; Dixon and Cambouropoulos, 2000; Gasser et al., 1999) or audio data (Dixon, 2001; Dixon et al., 2003).

Pitch Pitch is easily obtained from scores and MIDI data, but can not be reliably extracted from polyphonic audio. Pitch is rarely used in rhythmic analysis; an exception is Dixon and Cambouropoulos (2000), who extract duration, amplitude and pitch from MIDI data in order to calculate the "salience" of musical events, which is shown to improve the performance of their beat tracking system.

Chords Chords are used in two ways in rhythmic analysis: by counting the number of simultaneous notes as a measure of accentuation (Dixon, 2001; Rosenthal, 1992), and by detecting harmonic change as evidence of a downbeat (Goto and Muraoka, 1999; Temperley and Sleator, 1999, p.25). Just like pitch, chords are easily readable in scores and MIDI data, but much harder to derive from audio data.

Percussive instrument classes Some authors use specific drum sounds (particularly snare and bass drum) as clues to distinguish between upbeats and downbeats (Goto, 2001; Gouyon et al., 2000; Zils et al., 2002; Goto and Muroaka, 1995).

Percussive events can easily be extracted from MIDI data. Dealing with audio signals, isolated percussion samples can be automatically classified with a high reliability (Herrera et al., 2003). This facilitates labelling tasks in transcription of percussive tracks ("drum loops") whose onsets have been previously detected (Schloss, 1985; FitzGerald et al., 2002; Bilmes, 1993). However, the recognition of percussive events in polyphonic audio mixtures is still ongoing research (Herrera et al., 2004; Gouyon et al., 2000; Zils et al., 2002).

Frame features Honing (1993) comments that "there seems to be a general consensus on the notion of discrete elements (e.g. notes, sound events or objects) as the primitives of music ... but a detailed discussion and argument for this assumption is missing from the literature." Further, Scheirer (2000) argues that solely well-trained musicians hear the music in terms of its conventional musicological structures, and he criticises the "transcriptive metaphor", maintaining that the modelling of the perceptual mechanism should not be based upon abstract symbols such as durations, pitches, and chords. For example, he showed in an informal experiment that

replacing the harmonic content of a musical signal with modulated noise did not change the sensation of tempo (Scheirer, 1998).

Based on this rationale, some systems do not focus on note onsets and their features, but refer to a data granularity of a lower level of abstraction: frames. A frame is a short chunk (typically 10 ms) of audio, from which both time and frequency domain features can be computed. Consecutive frames are usually considered with some overlap for smoother analyses. The analysis step, the *hop size*, equals the frame size minus the overlap.

The simplest feature is energy, which can be calculated for the whole frame or for frequency subbands of the frame. Assuming that low-frequency instruments communicate much of the rhythmic information, Alghoniemy and Tewfik (1999) and Blum et al. (1999) focus on the energy in low-frequency components, as a simple alternative to the percussion detection methods mentioned previously. Others decompose the signal into several subbands, compute energy in each subband, then optionally postprocess them (e.g. assign them different weights) and sum them back (Vercoe, 1997; Tzanetakis et al., 2002). Finally, another procedure is to compute one feature list per frequency subband, yielding e.g. 6 feature lists for Herre et al. (2002) and Wang and Vilermo (2001), 20 for Pampalk et al. (2002) and 23 for Sethares and Staley (2001).

Rather than focusing on frame energy values, some systems measure the variation of the energy between consecutive frames. For instance, Foote and Uchihashi (2001) use the cosine distance between the magnitude spectra of consecutive frames. In (Laroche, 2003) the magnitude spectrum is transformed by a compression function (e.g. a hyperbolic sinus) to give higher weights to high frequencies than low frequencies, and then a first-order difference is computed. Scheirer (1998) also computes the first-order difference of frame energy values in 6 frequency bands. In (Klapuri, 2003), the computation of "registral accent" entails the aggregation of the energy values computed in 36 frequency bands in a smaller set of feature lists (e.g. 6). Here also, a first-order difference replaces the frame energy value.

Low-level features other than energy (e.g. spectral flatness, temporal centroid) have also been recently advocated (Gouyon and Herrera, 2003b,a).

One might note that these procedures resemble the first stages of an onset detector. The main difference is that there is no discretisation of frame energy values, nor any explicit thresholding and peak-picking. Further rhythm description stages deal with a data granularity defined by the hop size.

Features computed on lower metrical levels Several authors propose to compute low-level features over the time span of a given metrical level. For instance, Seppänen (2001) and Gouyon and Herrera (2003a) compute beat indexes from low level features computed on segments of audio defined by the smallest metrical level. Also, Goto and Muraoka (1999), Meudic (2002) and Gouyon and Herrera (2003b) derive downbeat indexes from descriptors of beat segments. The latter points out the relevance of a specific feature for downbeat computation: the temporal centroid of the beat.

4 Pulse induction

A metrical level (a pulse) is defined as the periodic recurrence of a feature in time. Therefore, computer programs generally seek periodic behaviours in feature lists in order to select pulse periods and possibly also their phases. The process of *pulse induction* aims at highlighting intrinsic periodicities of feature lists, and thus it is central to any form of rhythm understanding (see Figure 3).

The resulting pulses often serve as input to a *pulse tracker*. This division in the processing is motivated in Desain's "(de)composable theory of rhythm perception" (1992) that highlights the need to consider events with respect to the rhythmic context. This context can be defined mathematically as an expectancy curve, a function of past IOIs. Further, Desain and Honing (1999) argue that human perception of pulse exhibits two dichotomic processes: a bottom-up process that forms a pulse percept very rapidly from scratch, and a top-down process (a persistent mental framework) that lets this induced percept guide the organisation of incoming events.

In pulse induction, a fundamental assumption is made: the pulse period (and phase) is *stable* over the data used for its computation. That is, there is no significant speed variation during the excerpt used for inducing a pulse. In that part of the data, remaining timing deviations (if any) are assumed to be short term (considered as either errors or expressivity features). They are either "smoothed out" or cautiously handled within the pulse induction process so as to derive patterns of short term timing deviations as e.g. the swing.

For pulse induction, computer programs either proceed by *pulse selection*, evaluating the importance, or salience (Parncutt, 1994), of a *restricted number* of possible periodicities, or by a *periodicity function computation*, generating a *continuous* function plotting pulse salience versus pulse period (or frequency). The former procedure is simpler, and is typically used for processing symbolic data, where pulse selection is usually considered jointly with subsequent tracking. Systems handling finer-grained data (e.g. frame features) often implement a periodicity function computation. We now detail these two approaches in turn.

4.1 Pulse selection

The first approach to pulse selection is an instance-based approach, where each IOI defines a possible pulse period, and the corresponding events define the phase. For example, Longuet-Higgins and Lee (1982) simply consider the first two events as the first two beats, whereas Dannenberg and Mont-Reynaud (1987) take the first two agreeing IOIs as defining the pulse. In Allen and Dannenberg's (1990) system, the metrical value of the first event must be given, and the pulse is derived from this value and the first IOI. Chung (1989) derives a number of pulse periods and phases from the event list in a sequential manner. Like Longuet-Higgins and Lee (1982), Chung considers the first two events as potential beats. Subsequent events are considered in the light of this potential pulse: if they do not coincide with the pulse (after allowing some tolerance), a new potential pulse is created, its period being set to the most recent IOI, and the phase being specified by the current event. Limiting the number of pulses is achieved by assigning to each pulse a score depending on: the "importances" (i.e. durations) of its constituent events, the timing deviations of beats from expected pulse positions and the number of syncopations. Solely the two or three highest-scoring pulses are selected. Chung reports that the system usually finds all relevant pulses within the first few bars.

It is also possible to seek periodic behaviors in the feature list by computing a similarity measure between the list and several pulse tracks. This procedure is foreshadowed by the "clock model" of Povel and Essens (1985), where "people perceive, remember and reproduce temporal patterns by structuring their representation according to an internal clock" (McAuley and Semple, 1999, p.178) with a period corresponding to the smallest IOI. This rationale is only suitable for parsed scores and artificially created sequences where IOIs are exact integer multiples of the clock period. In this case, goodness of fit between a pulse and an event list can be estimated by positive evidence (the number of events that coincide with pulse elements), or by negative evidence (the number of pulse elements with no corre-

sponding event), or by combining these two counts (McAuley and Semple, 1999). Similarly, Parncutt (1994, pp.433-436) considers pulse induction on cyclically repeating musical patterns, by expressing pulse period and phase with respect to the shortest IOI, and determining the "pulse-match salience" based on positive evidence rather than negative evidence.

4.2 Computing a periodicity function

The alternative to pulse selection is the computation of a periodicity function: a magnitude (or salience) corresponds to each period (or frequency) in the periodicity continuum. In practice, the range of periods is not continuous but sampled, with typical intervals being 5 or 10 ms. Some systems process several feature lists separately, for example by calculating periodicities in each frequency subband and then integrating the results (Paulus and Klapuri, 2002; Dixon et al., 2003; Scheirer, 1998; Gouyon and Herrera, 2003a). The periodicity function may also be multiplied by a tempo preference (probability) distribution (e.g. Parncutt, 1994, p.439, equation 7), implementing the fact that humans consider tempi with some a priori preference. Some methods also let major periodicities affect rationally-related periods (e.g. a τ-periodicity in the feature list contributing to the raising of several peak magnitudes: at τ, $\tau/2$, etc.), thus encoding aspects of the metrical hierarchy. In some cases, an emphasis is given to the most recent samples, e.g. by multiplying the data with a decreasing window (Desain and de Vos, 1990), (Goto, 2001, equation 7), or by the intrinsic exponential behavior of a comb filter impulse response (Scheirer, 1998). Cemgil et al.'s "tempogram" (2001) also implements this feature in its parameter α.

Periodicity functions are often calculated with standard signal processing algorithms, such as the Fourier transform, which Blum et al. (1999) applies to onset lists and Pampalk et al. (2002) uses on 20 frequency subbands of the audio signal. Wavelets are used to capture temporal organisations at different hierarchical levels in (Smith and Kovesi, 1996; Smith, 1996). The most common signal processing technique for periodicity computation is the autocorrelation function (ACF), which has been applied to subband signals and to onset lists represented as Dirac delta functions (for scores or mechanical performances) or smoothed using e.g. a Gaussian function (to cater for small changes in timing and tempo; see section "Event shift handling").

Brown (1993) computes a sample-by-sample ACF of a sequence of onsets sampled at 200 Hz, weighted by their durations. Her results are best for longer values of the integration time (the time span for the estimation of one correlation coefficient). The integration time is also important because it determines the statistical reliability of the estimate (Desain and de Vos, 1990). Scheirer and Slaney (1997) also compute the ACF of onset trains, and Scheirer (1997) advocates summing ACFs computed over several frequency channels.

The "Narrowed ACF" (NACF) was introduced by Brown and Puckette (1989): the coefficient at lag k is computed as the weighted sum of the ACF coefficients at lags which are integer multiples of k, where the weights decrease for larger multiples of k. The NACF implicitly encodes aspects of the metrical hierarchy (a 2*k-periodicity has an effect on the correlation coefficient of lag k) and gives better period precision at the expense of worse time resolution. Improved precision is a useful feature for signals that contain close periodicities, but this is an unlikely situation in the context of pulse induction. It may be noted that Brown (1993, p.1955) recognises that the NACF is not necessary. Vercoe (1997) proposes the use of the "Phase-Preserving Narrowed Autocorrelation" in order to keep time localisation normally lost in computing an ACF. The computation involves a simplified NACF, i.e. with a very short integration time, which reduces the stability of the

estimate.

Foote and Uchihashi (2001) propose two ways to compute periodicities ("self similarity") in feature lists: they build a similarity matrix and perform either sums or correlations of the matrix diagonal elements. The first of these two options resembles the computation of an ACF: the sum over the ith diagonal is similar to the (normalised) autocorrelation of the signal frame parameters with a lag i. The latter option is similar to the NACF, in that it goes further and accounts for aspects of the metrical hierarchy.

ACFs are also implemented in (Goto, 2001; Gouyon et al., 2000; Gouyon and Herrera, 2003a,b; Dixon et al., 2003; Tzanetakis et al., 2002; Herre et al., 2002).

An alternative approach uses a bank of resonators, each tuned to a possible periodicity, where the output of the resonator indicates the strength of that particular periodicity. Scheirer (1998) uses comb filters as resonators, and performs periodicity analysis separately on 6 frequency subbands of the signal, and then sums the filterbank outputs across the subbands. 150 resonators are used to cover a logarithmically spaced frequency range from 1 Hz to 3 Hz (i.e. 60 to 180 BPM). Scheirer (1997,1998) details similarities and differences between the NACF and comb filter approaches. This method also "encodes implicitly aspects of the rhythmic hierarchy" (Scheirer, 2000, p.91).

The use of histograms of time intervals between similar events is also widespread. These are typically IOI histograms, although Mont-Reynaud and Goldstein (1985) builds histograms of time intervals between temporal *patterns*, resembling somewhat an ACF. Chowning et al. (1984, pp.17-19) and Schloss (1985, p.90) generate a smoothed histogram by associating a Dirac delta function with each IOI, assigning it a weight proportional to its value (i.e. longer IOIs are emphasised) and convolving them with a "bell shaped curve of appropriate bandwidth." Similarly, Rosenthal (1992, p.40) builds a discrete IOI histogram and smears it with a Gaussian curve. Dixon's IOI clustering scheme (Dixon, 2001,1999) is essentially similar to the building of an IOI histogram where the bins are not fixed. Clusters of similar IOIs are given scores based on the number of elements in the cluster and the amplitudes of their onsets. An adjustment of the scores (and cluster representative interval) then favours rationally-related clusters, thus encoding aspects of the metrical hierarchy. Seppänen (2001) and Gouyon et al. (2002) also implement IOI histograms. In the former, the computation is sequential and updated at each new event, emphasis being given to the most recent ones.

Sethares and Staley (2001) propose the "periodicity transform", which projects the signal (here made up of frame energy in a subband) onto a set of basis vectors. Unlike the Fourier and wavelet transforms, the basis vectors are not specified a priori, rather the transform adapts to find the basis vectors which best match the signal.

Cemgil et al. (2001) define the "tempogram" which induces a probability distribution over the pairs {pulse period, pulse phase} given the onsets. Using a Bayesian framework, this probability (*posterior* distribution) is proportional to the *likelihood* of the observed onsets under given period and phase hypotheses, weighted by a *prior* distribution (which in this case is flat, as they consider all tempi to be initially equiprobable). For given periods and phases, the likelihood is computed as the integral, over all the onsets, of the product of a constant pulse track (with appropriate period and phase) and a continuous representation of the onsets (onsets are smeared with a Gaussian curve). It implements the assumption that a good pulse track is one which matches all the onsets well. The tempogram marginal probability function $p(w|t)$ (integral of the tempogram with respect to phase) provides a 1-dimensional representation of periodicities resembling those aforementioned (Cemgil et al., 2001, figure 4).

Recall that the pulse selection method used by Parncutt (1994) and McAuley

and Semple (1999) is based the computation of a similarity measure between event lists and pulse track templates. It can be generalised to deal with musical patterns which are not strictly metronomical and not cyclically repeating. In this case, the "basic time unit" is not known, but it is possible to enumerate all possible pulse track periods and phases, as do Gouyon et al. (2002), who use both positive and negative evidence in the matching of onset lists and pulse tracks. If the feature representation is continuous (e.g. when adding some degree of tolerance for onset times, or when using frame energies), it is no longer meaningful to speak of positive and negative evidence. However, computing the inner product between pulse tracks and the continuous feature list is possible (Laroche, 2003, equation 5). This resembles the aforementioned tempogram, the main differences are that the tempogram accounts for weights on pulse track elements and considers all phase candidates simultaneously.

4.2.1 Parsing the periodicity function

The desired output of the pulse induction process is a discrete pulse period (and optionally its phase) for each periodicity, rather than a continuous periodicity function. Therefore another step is needed in order to produce useful rhythmic information. Usually, this is achieved by a peak-picking algorithm such as an N-point running window method, which defines local maxima as points whose values are higher than those of their direct neighbours (N/2 on the left and N/2 on the right). Peaks must be subsequently interpreted with respect to their musical meaning, e.g. the tick, tactus and measure periods, which are identified using heuristics (Goto, 2001; Smith and Kovesi, 1996; Smith, 1996).

Chowning et al. (1984, pp.17-19) and Schloss (1985, p.90) perform peak-picking on a smoothed IOI histogram, and keep the highest peak, qualifying it as the "important duration." Likewise, Rosenthal (1992, p.41) takes the maximum peak as being the tactus, using a peak-picking algorithm with a bias towards smaller IOIs. In Foote and Uchihashi's (2001) "beat spectrum", the pulse period is determined as the maximal peak, also by peak-picking. In (Brown, 1993), the pulse of interest is the downbeat (measure). All the peaks in the ACF are detected and the measure period is taken from the peak whose height is greater than those of all previous peaks and all subsequent peaks up to twice its period.

Some systems (e.g. comb filter, tempogram) compute the pulse phase (hence all beat positions) jointly with the period. In other cases (e.g. ACF), the computation of the period entails the loss of time localisation, and the phase has to be computed subsequently, either during pulse tracking (e.g. Dixon, 2001) or by enumerating possible phases once the period is known, and calculating the best match (Gouyon et al., 2002).

4.3 Event shift handling

Short term timing deviations always exist in any musical data other than parsed scores and artificial sequences. Feature periodicities are always approximate. This is a problem especially when processing discrete event lists represented as a sum of Dirac delta functions.

One solution is to consider events as having a "tolerance interval" (Longuet-Higgins, 1987). Dixon (2001) uses a fixed tolerance interval of 25 ms (the "cluster width") for IOIs, whereas Dixon et al. (2003) and Chung (1989, p.65) employ tolerance intervals proportional to the IOIs, so that longer IOIs allow for greater variations. Seppänen (2001) quantises the IOI histogram into a specific number of bins, giving a fixed tolerance interval, but does not state the number of bins. A tolerance interval can also be considered in the creation of the feature list, such

as the "summary events" of Rosenthal (1992, p.29) which merge note events into chords if their onset times are within a timing tolerance of 10 ms. Similarly, Dixon and Cambouropoulos (2000) use a tolerance of 70 ms to define onset simultaneity.

The previous procedures can be interpreted as convolving the event list with a rectangular window. This helps in processing music with short term timing deviations, but the resulting representation is still discontinuous (the sum of Dirac functions has been transformed into a step function). This can be improved by using smoother curves for smearing, such as a Gaussian window (Schloss, 1985; Rosenthal, 1992; Gouyon et al., 2002; Cemgil et al., 2001; Chowning et al., 1984), an exponential window (Dannenberg and Mont-Reynaud, 1987, p.245), or a triangular window (Tanguiane, 1994).

5 Pulse tracking

Pulse tracking and pulse induction often occur as complementary processes. Pulse induction models consider short term timing deviations as noise, assuming a relatively stable tempo, whereas a pulse tracker handles the short term timing deviations and attempts to determine changes in the pulse period and phase, without assuming that the tempo remains constant. Another difference is that induction models work bottom-up, whereas tracking models tend to follow top-down approaches, for example, driven by the pulse period computed by the pulse induction module.

Pulse tracking is often implemented with online algorithms, making real-time implementations possible. Previous data is used to compute pulse period and phase that are used as predictions propagated onto incoming data, and tracking is then a process of reconciliation between these predictions and the observed data. An important part of this process is entrainment, adapting the pulse period and phase based on the observations, which must find a good balance between *reactiveness* and *inertia*. Reactiveness determines how quickly the system responds to a change, and reflects the importance given to the incoming data, while inertia determines the stability of the system and reflects the importance attached to the context given by past data.

Diverse formalisms and techniques have been used in the design of pulse trackers: rule-based, problem-solving, agents, adaptive oscillators, dynamical systems, Bayesian statistics and particle filtering. The framework of state models is general enough to describe and compare pulse trackers: they can all be defined by a set of state variables, an initial situation (initial values for these variables), observations (incoming data), a goal situation (finding the best explanation for the observations), a set of actions (adapting the state variables in order to reach the goal situation) and methods to discriminate good and bad actions. In the remainder of this section, we review how diverse models deal with the adaptation of state variables to the observations.

5.1 Observations and state variables

Observed musical events are usually onset features: onset times, durations (or IOIs) and dynamics. Tracking models follow two different rationales regarding observations: they either consider events *sequentially* (i.e. each incoming event is processed and influences the tracker) or they consider *predicted beat positions* (i.e. only events around predicted beats are processed; others are disregarded). State variables usually account for the pulse period or tempo, and the pulse phase, expressed as either the current beat position or the first beat. Some models also include other variables, such as the estimated metrical position or a performance measure indicating the tracker's self-evaluation.

5.2 Actions

Adaptive oscillators predict the next beat position as the current beat position plus the pulse period, and then choose the closest event to this predicted position and adapt the state variables accordingly (McAuley, 1995; Large and Kolen, 1994). For instance in (Large and Kolen, 1994) a simple oscillator, called the "driven" unit, embodies the period and phase variables and adapts to incoming events emitted by the "driver" unit. Each event from the driver perturbs the phase of the driven unit by an amount determined by the coupling strength, which in turn determines the balance between reactiveness and inertia of the model. McAuley (1995) and Large and Kolen (1994) both suggest connecting several oscillators in a network so that they can interact, in order to model several metrical levels jointly (see also Eck et al., 2000; Gasser et al., 1999).

In the rule-based approach (e.g. Desain and Honing, 1999; Longuet-Higgins and Lee, 1982), the state variables are the pulse period and the first and current beats. A set of "if-then" rules adapts these variables as each event is observed, and predicts the next beat. For instance, in (Longuet-Higgins and Lee, 1982), a beat is predicted at the current beat position plus the pulse period, and the pulse period is then adapted by two rules: "conflate" and "stretch." The former achieves a doubling of the pulse period when an onset is observed on the predicted beat, the latter changes the period if an onset is observed before the predicted beat (then the period is set to the distance between this new onset and the penultimate beat). Pulse phase is adapted by the rule "update": if no onset is observed at the predicted beat (nor before it), the first beat is shifted to the current beat and the current beat to the predicted beat (regardless of the fact that there is no onset there). This approach seems biased towards reactiveness rather than inertia.

In (Dannenberg and Mont-Reynaud, 1987), incoming events are considered sequentially and the pulse period is updated as follows. An integer divisor (or multiple) of the pulse period is assigned to the next observation (e.g. 1, 1/2, 1/3, 2, etc.) as the closest metrical position to the actual event position. The resulting deviation then serves to update the pulse period. This updating mechanism depends also on the event position in the metrical hierarchy: events close to multiples of the expected pulse period have a greater impact on the updating mechanism than other events, e.g. half-periods (see Dannenberg and Mont-Reynaud, 1987, "Confidence" parameter). Finally, the balance between reactiveness and inertia is explicitly monitored by the "Decay" parameter.

Allen and Dannenberg (1990) propose to add some flexibility to the previous model by fine-tuning the "Decay" and "Confidence" parameters, depending on the musical style. However, observing that this model does not possess the capability to recover after an error, they introduce the notion of concurrent hypotheses, where a hypothesis is a *sequence of states*. Incoming events are also considered sequentially in this model, but the system does not commit to a decision at each observation. Rather, the evolutions of several concurrent hypotheses are evaluated with some delay with respect to real-time, so that decisions are not taken on the basis of a given state, but on the basis of a sequence of states. In addition to the period and phase variables, a metrical position and a "credibility" (performance measure) are also state variables. In this framework, the number of hypotheses increases with each observation, resulting in a search tree. The tree is pruned to an acceptable size by discarding some hypotheses based on heuristics which implement simple aspects of musical knowledge (e.g. "quarter-notes must start on the downbeat or the upbeat"). Other techniques to reduce the number of hypotheses are by using best-first search, discarding hypotheses which duplicate the current state of other hypotheses, and limiting the number of likely metrical positions. Temperley and Sleator (1999) use dynamic programming to search the solution space of possible mappings of events

to a metrical grid, where the search is guided by a set of preference rules based on GTTM.

Dixon (2001) presents another multiple hypothesis search approach, using an agent paradigm, where each agent has a state (state variables are period and phase of a pulse) and a history ("the sequence of beat times selected to date by the agent"). These agents are comparable to the hypotheses of Allen and Dannenberg (1990), except that observations are only processed if they occur around the predicted beat locations, i.e. "within a window whose width depends on the pulse period."

Cemgil et al. (2001) address pulse tracking through the use of a dynamical system, a "metronome model" that updates state variables at each inferred beat. The system is defined with two hidden state variables: the period and the phase of a metronome. Transition from one metronome beat to the next is modeled by a simple set of state equations. The model is fully determined if the initial state variables are given. To this deterministic model, they add a noise term (a Gaussian random vector whose covariance matrix will be estimated through a training phase) that models the likely tempo variations. Observations to the dynamical system ("noisy metronome beats") are given by the computation of a "tempogram" from incoming onsets. The hidden state variables are estimated by means of a Kalman filter and extensions to the Kalman filter are proposed.

5.2.1 Tracking as repeated induction

Some systems address pulse tracking by repeated induction of the pulse (e.g. Foote and Uchihashi, 2001; Dixon et al., 2002; Scheirer, 1998; Chung, 1989). A pulse is induced on a short analysis window (usually around 5s of data), then the window is shifted in time to include the next event and another induction step takes place. (If the feature list consists of frame features, the hop size is constant.) In this framework, observations to the tracking process are no longer events as used above, but the period and phase of a pulse. Determining the tempo evolution is then reduced to connecting the observations at each step.

In addition to computational overload, one problem that arises with this approach to tracking is the lack of continuity between successive observations. A continuity constraint is implicitly present in the fact that the analysis hop size is usually much smaller than the window size but this results in a strong bias towards inertia rather than reactiveness and an impossibility to model sharp tempo changes.

If each induction step yields several pulse period and phase hypotheses, finding the final tempo curve and beat locations sums up to finding the best path that connects successive hypotheses, e.g. by dynamic programming (Laroche, 2003). 'Best' can be formalised, here again, by costs assigned to the several ways of adapting state variables (i.e. pulse periods and phases and measures of self-evaluation). For Goto (2001) and Laroche (2003), this entails continuity and non-syncopation constraints.

6 Further aspects of metrical structure

Deriving a complete rhythmic transcription of an audio signal (i.e. producing a score) requires the determination of a reference metrical level and its tempo, the time signature, durations for notes and silences quantised with respect to the reference pulse, and measure boundaries (bar lines). We now briefly address the second and third of these points.

6.1 Time signature determination

Few algorithms for time signature determination exist. The simplest approach is based on parsing the peaks of the periodicity function to find two significant peaks, which correspond respectively to a fast pulse, the time signature denominator, and a slower pulse, the numerator (Brown, 1993). The ratio between the pulse periods defines the time signature. Another approach is to consider all pairs of peaks as possible beat/measure combinations, and compute the fit of all periodicity peaks to each hypothesis, using a weighted sum, where the weights represent the likelihood of each metrical unit appearing as a strong periodicity, given the metre (Dixon et al., 2003). The time signature is implicitly calculated by systems that induce a complete metrical structure (e.g., Temperley and Sleator, 1999). Another strategy is to break the problem into several stages: the determination of the time signature denominator (e.g. by tempo induction and tracking), the segmentation of the musical data with respect to this pulse, the definition of features at this temporal scope and subsequently the detection of periodicities in feature lists (Meudic, 2002; Gouyon and Herrera, 2003b). Goto and Muraoka (1999) detect chord changes as indicators of higher level metrical boundaries such as bar lines; however their work is restricted to music with a 4/4 time signature.

6.2 Rhythm parsing (quantisation)

Rhythm parsing can be seen as a by-product of the induction of several metrical levels (e.g. Chung, 1989), which together define a metrical grid. The rhythm of a given onset sequence can be parsed by assigning each onset (independently of its neighbours) to the closest element in this hierarchy. The weaknesses of such an approach are that it fails to account for musical context (e.g. a triplet note is usually followed by 2 more) and tempo changes.

Models by Desain and Honing (1989) and Cemgil et al. (2000) do account for musical context and possible distortions of the metrical structure. However such distortions would in turn be easier to determine if the quantised durations were known (Allen and Dannenberg, 1990). Therefore, rhythm parsing is often considered as a process simultaneous with tempo tracking, rather than subsequent to it. Here also, the joint estimation of tempo and rhythm can be seen as a process of reconciliation between predicted values for state variables and sequential observations. The main difference with tempo tracking lies in the fact that the state variables explicitly specify the interdependency between tempo and rhythm.

For instance, in (Rosenthal, 1992) the state variables account for three metrical levels simultaneously and observations are not defined by sequential events, but only by events which are close to beats at one of the metrical levels. The pruning techniques are comparable to those used by Allen and Dannenberg (1990), but adapted to the fact that states (and therefore hypotheses) are more complex (Rosenthal, 1992, pp.57-68). In (Raphael, 2002) and (Cemgil and Kappen, 2003), events are considered sequentially. Concurrent hypotheses are expressed as posterior probabilities of a probabilistic model whose hidden layers (i.e. state variables) account for score notation and ideal timing in addition to tempo. They implement different strategies for parsing the tree of hypotheses and keeping it from growing exponentially. For instance, particle filters are suitable (see also Hainsworth and Macleod, 2003, for a similar approach using audio data). Temperley and Sleator (1999) also process events sequentially, using dynamic programming and a simple set of preference rules to infer up to 5 metrical levels.

Thornburg (2001) also follows the same rationale, however he includes audio segmentation (onset detection) as a third interdependent process, rather than a preprocessing step before rhythm parsing and tempo tracking. He argues that these

tasks should be considered jointly: polyphonic audio segmentation is necessary to provide data to the rhythm tracker, but rhythm tracking should also orient (i.e. provide prior probabilities to) the segmentation task. This helps to ensure robustness against spurious onsets, which are a common problem in polyphonic audio segmentation. The systems based on MIDI input (Temperley and Sleator, 1999; Raphael, 2002; Cemgil and Kappen, 2003) account inherently for noise in onset timing, but not for spurious onsets.

6.3 Systematic deviation estimation

In the pulse induction process, short term timing deviations can be "smoothed out" or cautiously handled so as to derive patterns of short term timing deviations such as swing. Foote and Uchihashi (2001) suggest that swing could be measured by inspection of a periodicity function (there, the "beat spectrum") within the pulse induction process. This is illustrated by the positions of secondary peaks with respect to some higher ones in (Foote and Uchihashi, 2001, figure 3), but they do not suggest any extraction procedure. Another problem is that periodicity functions do not distinguish the order of events, e.g., the difference between a long-short pattern and a short-long pattern.

Laroche (2001) proposes to estimate the swing jointly with tempo and beats at the half-note level, assuming constant tempo. The procedure is conceptually similar to pulse induction using a pulse track matching function, but enumerating all possible pulse periods and phases, like Cemgil et al.'s tempogram, and searching for the one which best matches the onsets. The number of candidate pulse tracks (the search space) is in fact even larger, as tracks have a third parameter to be estimated (the swing) in addition to the tempo and phase parameters. In this case the pulse tracks are no longer isochronous, but correspond to the long-short timing pattern that we wish to find in the data. The amount of deviation from an isochronous track defines the swing ratio. Gouyon et al. (2003) estimate swing ratio in a comparable fashion.

7 Summary - Discussion

Within any computational modelling paradigm, systematic evaluations of competing models is highly desirable. However, there are several reasons why such an evaluation is not possible for rhythm description. First, there are many models, but few open source implementations, and few models are described completely enough in order to reimplement them. Further, there is no common database of test music labelled with the "ground truth" (but see Temperley, 2004, for a recent proposal regarding MIDI data). Another reason is that there are no precise problem definitions or evaluation criteria, since rhythm description systems have been built for diverse applications using diverse data sets. This review has provided a qualitative comparison of existing systems with respect to the functional units of a general model (Figure 3).

A common aspect of all computational models is the handling of feature lists, either as a starting point (for scores or MIDI data) or as a mid-level representation (for models that process audio). These features (e.g. onsets, amplitudes, pitches, percussive instrument classes, frame subband energy) are assumed to convey the predominant information relevant to a rhythmic analysis. Except in the case of frame-based features, the features are high-level, entailing an "implicit symbolism" (Scheirer, 2000). The first stages of human rhythm perception achieve a comparable parsing of auditory streams into feature lists, however, the actual modelling of these

perceptual processes (the definition of perceptually relevant features) is still ongoing research.

We depicted two procedures for pulse induction: pulse selection and periodicity function computation, and gave examples of various implementations. Computing a periodicity function is usually more powerful than just selecting a pulse. However, there is a trade-off between the amount of data used for induction and the likelihood that the tempo-stability assumption holds. Using few data (typical of pulse selection methods) lowers the reliance on a constant tempo but generates less reliable predictors, whereas using many data (typical of periodicity computation methods) generates more reliable predictors but only when the tempo remains relatively unchanged over this longer duration.

The modelling and processing of short term timing deviations are particularly relevant. Some pulse induction methods encode (implicitly or explicitly) aspects of the metrical hierarchy by letting large time-scale phenomena influence responses at smaller time scales (and inversely), e.g. comb filters. In fact, this encodes the assumption that the perception of high metrical levels, e.g. the measures, orients the perception of lower metrical levels from which they are derived. Parncutt (1994, p.434) questions this assumption, writing "each pair of events in a rhythmic sequence initially contributes to the salience of a *single* pulse sensation" (emphasis ours), and later that "pulse sensations can enhance the salience of other, consonant pulse sensations." One may understand the 'initially' above as an indication not to implement influential schemes between metrical levels in the induction process, but indeed to do it in the tracking process, which is also in agreement with the Dynamic Attending Theory (Drake et al., 2000a; Jones and Boltz, 1989).

A number of diverse formalisms have been used to implement pulse tracking models. An important aspect is the balance between inertia and reactiveness of the model. Models with a sufficient degree of inertia can be built by accounting for several concurrent hypotheses. This seems a must for preventing "garden-path errors" (Rosenthal, 1992, p.11) and keeping the possibility of recovering after an error. Another important aspect lies in the consideration of incoming data on an event by event basis or a predicted beat by predicted beat basis. Following the former strategy is in fact making a first step towards quantising the data, not solely tracking a pulse. AI formalisms have been proposed recently to enhance tempo trackers and address quantisation and pulse tracking jointly.

Very few algorithms for time signature determination exist. They usually entail the computation and parsing of a periodicity function, as in pulse induction. Apart from swing estimation, systematic timing deviation estimation is the object of few computational models. The usual rationale behind swing estimation is to consider that the tempo is constant (i.e. no long term timing deviations) and to seek predefined patterns of short term timing deviations within a pulse induction process.

Current research in rhythm description addresses all of these aspects, with varying degrees of success. For instance, determining the tempo of music with minor speed variations is feasible for almost all musical styles, if we do not insist that the system finds a specific metrical level. Recent pulse tracking systems (Dixon, 2001; Cemgil et al., 2001) also reach high levels of accuracy. On the other hand, accurate quantisation, score transcription, determination of relevant rhythmic features, determination of time signature and characterisation of intentional timing deviations are still open questions. Particularly, it remains to be seen how well recently proposed models generalise to different musical styles.

New research directions include the determination of highly abstract rhythmic features required for music content processing and music information retrieval applications, as tackled by e.g. the European projects SIMAC (www.semanticaudio.org), Semantic HIFI (www.ircam.fr) and GOASEMA (www.ipem.rug.ac.be).

8 Acknowledgements

This research was supported by the EU project FP6-507142 (SIMAC) and the national project Y99-INF sponsored by the Austrian Federal Ministry of Education, Science and Culture in the form of a Start Research Prize. The Austrian Research Institute for Artificial Intelligence acknowledges the support of the Austrian Federal Ministries of Education, Science and Culture and of Transport, Innovation and Technology.

9 Bibliography

Alghoniemy, M. and Tewfik, A. (1999). Rhythm and periodicity detection in polyphonic music. Proc. IEEE Workshop on Multimedia Signal Processing, pp. 185-190.

Allen, P. and Dannenberg, R. (1990). Tracking musical beats in real time. Proc. International Computer Music Conference, pp. 140-143.

Baggi, L. (1991). Neurswing: An intelligent workbench for the investigation of swing in jazz. *Computer*, 24(7):60-64.

Bello, J. (2003). *Towards the Automated Analysis of Simple Polyphonic Music: A Knowledge-based Approach*. PhD thesis, Dept of Electronic Engineering, Queen Mary University of London.

Bilmes, J. (1993). *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm*. MSc thesis, MIT, Cambridge.

Blum, T., Keislar, D., Wheaton, A., and Wold, E. (1999). Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information. USA Patent 5,918,223.

Brown, J. (1993). Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America*, 94(4): 1953-1957.

Brown, J. and Puckette, M. (1989). Calculation of a narrowed autocorrelation function. *Journal of the Acoustical Society of America*, 85(4):1595-1601.

Cambouropoulos, E., Dixon, S., Goebl, W., and Widmer, G. (2001). Computational models of tempo: Comparison of human and computer beat-tracking. In *Proceedings of VII International Symposium on Systematic and Comparative Musicology and III International Conference on Cognitive Musicology*, pp. 18-26.

Cemgil, A., Desain, P., and Kappen, B. (2000). Rhythm quantization for transcription. *Computer Music Journal*, 24(2): 60-76.

Cemgil, A. and Kappen, B. (2003). Monte carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, 18:45- 81.

Cemgil, A., Kappen, B., Desain, P., and Honing, H. (2001). On tempo tracking: Tempogram representation and kalman filtering. *Journal of New Music Research*, 28(4):259-273.

Chowning, J., Rush, L., Mont-Reynaud, B., Chafe, C., Schloss, A., and Smith, J. (1984). Intelligent system for the analysis of digitized acoustic signals. Technical Report STAN-M-15, CCRMA, Stanford University.

Chung, J. (1989). *An agency for the perception of musical beats or If I had a foot...* MSc thesis, MIT, Cambridge.

Clarke, E. (1987). Categorical rhythm perception: An ecological perspective. In *Action and perception in rhythm and music*, pp. 19-34. Royal swedish academy of music.

Clarke, E. (1999). Rhythm and timing in music. In Deutsch, D., editor, *The Psychology of Music, 2nd edition*, pp. 473-500. Series in Cognition and Perception. Academic Press.

Clynes, M. and Walker, J. (1982). Neurobiologic functions of rhythm, time, and pulse in music. In M., C. and J., W., editors, *Music, mind, and brain: The neuropsychology of music*, pp. 171-216. Plenum.

Dannenberg, R. and Mont-Reynaud, B. (1987). Following an improvisation in real-time. Proc. International Computer Music Conference, pp. 241-258.

Desain, P. (1992). A (de)composable theory of rhythm perception. *Music Perception*, 9(4):439-454.

Desain, P. and de Vos, S. (1990). Autocorrelation and the study of musical expression. Proc. International Computer Music Conference, pp.357-360.

Desain, P. and Honing, H. (1989). The quantization of musical time: A connectionist approach. *Computer Music Journal*, 13(3):55-66.

Desain, P. and Honing, H. (1991). Tempo curves considered harmful. a critical review of the representation of timing in computer music. Proc. International Computer Music Conference, pp.143-149.

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

Dixon, S. (1999). A beat tracking system for audio signals. Proc. Conference on Mathematical and Computational Methods in Music (Diderot), pp. 101-110.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39-58.

Dixon, S. and Cambouropoulos, E. (2000). Beat tracking with musical knowledge. Proc. European Conference on Artificial Intelligence, pp. 626-630.

Dixon, S., Goebl, W., and Widmer, G. (2002). Real time tracking and visualisation of musical expression. In *Music and Artificial Intelligence: Second International Conference, ICMAI2002,* pp 58-68.

Dixon, S., Pampalk, E., and Widmer, G. (2003). Classification of dance music by periodicity patterns. Proc. International Conference on Music Information Retrieval, pp. 159-165.

Drake, C. (1993). Reproduction of musical rhythms by children, adult musicians and adult non-musicians. *Perception and Psychophysics*, 53(1): 25-33.

Drake, C. and Bertrand, D. (2001). The quest for universals in temporal processing of music. *Annals of the New York Academy of Science*, pp. 17-27.

Drake, C., Gros, L., and Penel, A. (1999). How fast is that music? the relation between physical and perceived tempo. In Yi, S., editor, *Music, Mind and Science.* Seoul National University Press, pp.190-203.

Drake, C., Jones, M., and Baruch, C. (2000a). The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending. *Cognition*, 77:251-288.

Drake, C., Penel, A., and Bigand, E. (2000b). Why musicians tap slower than nonmusicians. In Desain, P. and Windsor, L., editors, *Rhythm Perception and Production,* 245-248. Swets and Zeitlinger.

Eck, D., Gasser, M., and Port, R. (2000). Dynamics and embodiment in beat induction. In Desain, P. and Windsor, L., editors, *Rhythm Perception and Production,* 157-169. Swets and Zeitlinger.

FitzGerald, D., Coyle, E., and Lawlor, B. (2002). Sub-band independent subspace analysis for drum transcription. Proc. Digital Audio Effects Conference, pp.65-69.

Foote, J. and Uchihashi, S. (2001). The beat spectrum: A new approach to rhythm analysis. Proc. International Conference on Multimedia and Expo, pp.881-884.

Friberg, A. and Sundström, J. (1999). Jazz drummers' swing ratio in relation to tempo. Proc. Acoustical Society of America ASA/EAA/DAGA Meeting Lay Language Papers.

Friberg, A. and Sundström, J. (2002). Swing ratios and ensemble timing in jazz performances: Evidence for a common rhythmic pattern. *Music Perception*, 19(3): 333-349.

Gabrielsson, A. (1973). Similarity ratings and dimension analyses of auditory rhythm patterns. part i. *Scandanavian Journal of Psychology*, 14:138-160.

Gasser, M., Eck, D., and Port, R. (1999). Meter as mechanism: a neural network model that learns metrical patterns. *Connection Science*, 11(2): 187-216.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drums. *Journal of New Music Research*, 30(2):159-171.

Goto, M. and Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. *Speech Communication*, 27(3-4):311-335.

Goto, M. and Muroaka, Y. (1995). A real-time beat tracking system for audio signals. Proc. International Computer Music Conference, pp.171-174.

Goto, M. and Muroaka, Y. (1997). Issues in evaluating beat-tracking systems. Proc. International Joint Conferences on Artificial Intelligence, Workshop on Computational Auditory Scene Analysis, pp. 9-16.

Gouyon, F., Fabig, L., and Bonada, J. (2003). Rhythmic expressiveness transformations of audio recordings: swing modifications. Proc. International Conference on Digital Audio Effects, pp.94-99.

Gouyon, F. and Herrera, P. (2003a). A beat induction method for musical audio signals. Proc. WIAMIS Special session on Audio Segmentation and Digital Music, pp.281-287.

Gouyon, F. and Herrera, P. (2003b). Determination of the meter of musical audio signals: Seeking recurrences in descriptor of beat segment descriptors. Proc. Audio Engineering Society, 114th Convention.

Gouyon, F., Herrera, P., and Cano, P. (2002). Pulse-dependent analyses of percussive music. Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, pp.396-401.

Gouyon, F., Pachet, F., and Delerue, O. (2000). On the use of zero-crossing rate for an application of classification of percussive sounds. Proc. Digital Audio Effects conference.

Hainsworth, S. and Macleod, M. (2003). Beat tracking with particle filtering algorithms. IEEE Worshop on Applications of Signal Processing to Audio and Acoustics, pp.91-94.

Herre, J., Cremer, M., Uhle, C., and Rohden, J. (2002). Proposal for a core experiment on AudioTempo. Report MPEG2001/8415.

Herrera, P., Dehamel, A., and Gouyon, F. (2003). Automatic labeling of unpitched percussion sounds. Proc. Audio Engineering Society, 114th Convention.

Herrera, P., Sandvold, V., and Gouyon, F. (2004). Semantic interaction with music audio content using percussion-related descriptors, submitted.

Honing, H. (1993). Issues in the representation of time and structure in music. *Contemporary music review*, 9:221-239.

Honing, H. (2001). From time to time: The representation of timing and tempo. *Computer Music Journal*, 25(3):50-61.

Jones, M. and Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, 96(3):459-491.

Klapuri, A. (2003). Musical meter estimation and music transcription. Proc. Cambridge Music Processing Colloquium.

Lapidaki, E. (2000). Stability of tempo perception in music listening. *Music Education Research*, 2(1):25-44.

Large, E. and Kolen, E. (1994). Resonance and the perception of musical meter. *Connection Science*, 6:177-208.

Laroche, J. (2001). Estimating tempo, swing and beat locations in audio recordings. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.135-138.

Laroche, J. (2003). Efficient tempo and beat tracking in audio recordings. *Journal of the Acoustical Society of America*, 51(4):226-233.

Lerdahl, F. and Jackendoff, R. (1983). *A generative theory of tonal music.* MIT Press, Cambridge MA USA.

Longuet-Higgins, C. (1987). *Mental processes.* MIT Press, Cambridge.

Longuet-Higgins, C. and Lee, C. (1982). Perception of musical rhythms. *Perception*, 11:115-128.

McAuley, J. (1995). *Perception of time as phase: Towards an adaptive- oscillator model of rhythmic pattern processing.* Thesis/dissertation, Indiana University, Bloomington.

McAuley, J. and Semple, P. (1999). The effect of tempo and musical experience on perceived beat. *Australian Journal of Psychology*, 51(3):176-187.

Meudic, B. (2002). Automatic meter extraction from midi files. Proc. Journees d'informatique musicale.

Mont-Reynaud, B. and Goldstein, M. (1985). On finding rhythmic patterns in musical lines. Proc. International Computer Music Conference, pp 391-397.

Palmer, C. (1997). Music performance. *Annual review of psychology*, 48:115-138.

Pampalk, E., Rauber, A., and Merkl, D. (2002). Content-based organization and visualization of music archives. Proc. ACM International Conference on Multimedia, pp 570-579.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, 11(4):409-464.

Paulus, J. and Klapuri, A. (2002). Measuring the similarity of rhythmic patterns. In *Proceedings of the 3rd International Conference on Musical Information Retrieval*, pp. 150-156. IRCAM Centre Pompidou.

Povel, D. and Essens, P. (1985). Perception of temporal patterns. *Music Perception*, 2(4):411-440.

Raphael, C. (2002). A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, 137(1-2):217-238.

Repp, B. (1992). Probing the cognitive representation of musical time: structural constraints on the perception of timing perturbations. *Cognition*, 44:241-281.

Repp, B. (1994). On determining the basic tempo of an expressive music performance. *Psychology of Music*, 22:157-167.

Rosenthal, D. (1992). *Machine rhythm: Computer emulation of human rhythm perception.* Thesis/dissertation, MIT.

Scheirer, E. (1997). Pulse tracking with a pitch tracker. Proc. IEEE Worshop on Applications of Signal Processing to Audio and Acoustics.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, 103(1): 588-601.

Scheirer, E. (2000). *Music-listening systems.* Thesis, MIT Cambridge.

Scheirer, E. and Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. Proc. IEEE-ICASSP, pp. 1331-1334.

Schloss, A. (1985). *On the automatic transcription of percussive music - From acoustic signal to high-level analysis.* Thesis, CCRMA, Stanford University.

Seppänen, J. (2001). *Computational models of musical meter recognition.* M.Sc. Thesis, Tampere University of Technology.

Sethares, W. and Staley, T. (2001). Meter and periodicity in musical performance. *Journal of New Music Research*, 30(2):149-158.

Smith, L. (1996). Modelling rhythm perception by continuous time- frequency analysis. Proc. International Computer Music Conference.

Smith, L. and Kovesi, P. (1996). A continuous time-frequency approach to representing rhythmic strata. Proc. International Conference on Music Perception and Cognition.

Snyder, J. and Krumhansl, C. (2001). Tapping to ragtime: Cues to pulse finding. *Music Perception*, 18(4):455-489.

Tanguiane, A. (1993). *Artificial Perception and Music Recognition*. Springer, Berlin.

Tanguiane, A. (1994). A principle of correlativity of perception and its applications to music recognition. *Music Perception*, 11.

Temperley, D. and Sleator D. (1999). Modeling Meter and Harmony: A Preference-Rule Approach. *Computer Music Journal*, 23(1):10-27.

Temperley, D. (2004). An Evaluation System for Metrical Models. *Computer Music Journal*, 28(3):28-44.

Thornburg, H. (2001). Bayesian segmentation and rhythm tracking. Draft report CCRMA Stanford University.

Tzanetakis, G., Essl, G., and Cook, P. (2002). Human perception and computer extraction of musical beat strength. Proc. Digital Audio Effects Conference, pp 257-261.

Vercoe, B. (1997). Computational auditory pathways to music understanding. In Deliège, I. and Sloboda, J., editors, *Perception and Cognition of Music*, pp. 307-326. Psychology Press.

Wang, Y. and Vilermo, M. (2001). A compressed domain beat detector using mp3 audio bitstreams. Proc. ACM Multimedia.

Zils, A., Pachet, F., Delerue, O., and Gouyon, F. (2002). Automatic extraction of drum tracks from polyphonic music signals. Proc. International Conference on Web Delivery of Music.