Towards Rhythmic Content Processing of Musical Signals: Fostering Complementary Approaches

Fabien Gouyon
MTG-IUA, Universitat Pompeu Fabra, Barcelona, Spain
fgouyon@iua.upf.es, http://www.iua.upf.es/mtg

Benoit Meudic
Music Representation Team, IRCAM, Paris, France
Benoit.Meudic@ircam.fr, http://www.ircam.fr

This paper is concerned with the handling of rhythm in music content processing applications. Keeping this framework in mind, we briefly report on terminology issues and the interdisciplinary nature of rhythm investigations, we then review approaches to computational modeling of rhythm and rhythm representation schemes. We comment the bottom-up and top-down oriented approaches to computational modeling and the parallel that some authors make with physiological and cognitive views on rhythm perception. We argue that investigations should be listener-oriented, signal-oriented and application-oriented.

1 Introduction

Rhythm is a fundamental musical attribute. In the rapidly growing framework of music content-based processing and music information retrieval, there is an obvious need for a high-level description of the rhythmic aspects of music. On the other hand, there is still a poor understanding of human rhythm perception. There is neither a commonly accepted terminology nor a standard representation of rhythm.

There is a wide spectrum of contributions to the study of rhythm. Nevertheless, few specifically address content-based processing applications. Herein, we intend to give an overview of the requirements of such applications. We report on the interdisciplinary nature of the research in temporal aspects of music and present a review restricted to computational approaches, evaluating what the state-of-the-art models can offer. We also discuss the utility of several existing representation schemes. Finally, we foster complementary approaches to the design of specific prototypes.

The concept of rhythm is often considered indivisible, rhythmic transcription is often considered in tandem with perceptual phenomena. Throughout this paper, we propose a restraint regarding this rationale. As the concept of rhythm –and the way to apprehend it– can be understood in many different ways, we argue that the approach to investigate rhythm should depend strongly on (1) who is producing/reacting to rhythm, (2) the context or application-area, and (3) the type of the signal considered.

1.1 Music content processing framework

The standardization of personal computers and worldwide low-latency networks has spurned developments in digital audio dissemination. Hence, there are ample applications for content-based analysis of music, in particular rhythmic content. It is now technically possible to use PCs to browse and retrieve large databases of music files in audio and symbolic formats, analogous to text databases. In addition, computers are now powerful enough to serve as high-quality home-studios. For Aigrain (1999), in a near future, content-processing technologies will provide "new aspects of listening, interacting with music, finding and comparing music, performing it, editing it, exchanging music with others or selling it, teaching it, analyzing it, and criticizing it."

Accessing and processing musical data would seem much easier (and enjoyable) if one could directly handle meaningful descriptions of the data, rather than the data itself. However, there is a large gap between the features that can currently be computed from musical data and the type of meaningful abstract descriptions that users wish to handle. Indeed, it turns out to be very hard to automatically derive the semantics of musical data (even in symbolic format), much harder than that of *e.g.* text data. In fact, it is hard just to precisely define 'musical contents'. As music affects our minds and our emotions, it is difficult to represent it explicitly. Therefore, aiming at designing content-based applications, the first issue stands to propose representations of the data that would facilitate the apprehension of its semantics. Indeed, "the first dimension of the applications of content processing of music is to support and empower *active* listening" (Aigrain, 1999) –emphasis ours–).

Given a database of musical files (either full music titles, monophonic tracks or more generally sounds), such applications should ideally allow for (see Figure 1):

- Exploring and understanding items
- Database browsing
- Comparing a subset of items

Accepted for publication on May 9th 2002

- Retrieving any desired selection of items
- Transforming items
- Introducing new items

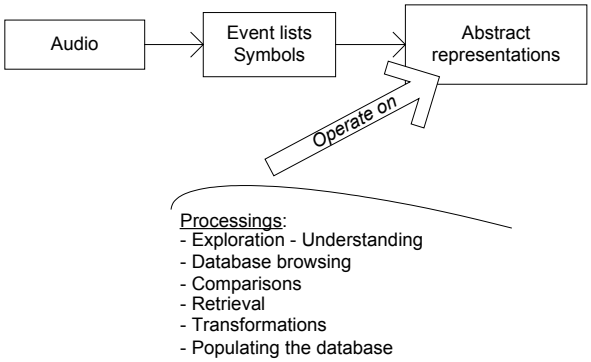

Figure 1: Overview of music content processing. From musical data to meaningful representations.

Further in the article, we will detail the issues of designing relevant representation schemes for musical data (in Section 3) and of identifying elements of these representations in musical data (in Section 2). Prior to these investigations, some additional terminology must be clarified.

1.2 Terminology issues

Even in writings whose very aim is to define rhythmic concepts, it is common to find confusing and sometimes contradictory statements. We will not address here all the aspects of rhythm, nor intend to provide "official" definitions. Our aim is rather to introduce the most salient rhythmic features in the context of content-based processing and to illustrate the usual notions associated to them.

Rhythm

"[...] a precise, generally accepted definition of rhythm does not exist." (Fraisse, 1982, p.149). The same consideration can be found in the very introduction of (London, 2001).

Just as it has a melodic dimension, Music has a rhythmic dimension. Here, the word 'rhythmic' implicitly encompasses all the temporal aspects of music, from small- to large-scale temporal phenomena. It is indeed difficult to draw a clear line regarding the temporal scope of rhythm. Some use this word when referring to the duration of a note, expressed relatively to a reference pulse (as in "this note is an eighth-note").[1] Others refer to the rhythm of a piece, as the tempo and the time signature. Others refer to the rhythm of a pattern of notes, which usually entails the notion of note duration as well as large-scale properties (as the tempo and the meter). Confusion abounds.

Rhythm is commonly defined indirectly. For instance, when stating rhythm is grounded in (1) the architectonic organization of the musical events in time and in (2) the accentuation (or differentiation) of some musical events in opposition to some others. (The architectonic nature of rhythm means that rhythmic features can be assigned to a piece of music according to a nested hierarchy of elements each on a different temporal scale.) Thus, rhythm involves regularity (or organization) and also differentiation (Fraisse, 1982, p.151).

Rhythm is also indirectly defined in its usual opposition to the 'meter' and the 'form'. The three terms involve regularity and differentiation. Yet, the distinction lies in the concept of 'perceptual present'. For London (2001), "rhythm involves the pattern of durations that is phenomenally present in the music, while meter involves our perception and anticipation of such patterns." He also puts it differently: "meter [is] a mode of attending, while rhythm is that to which we attend." London considers that rhythm's proper meaning refers to the "smaller-scale features of musical experience." The reason for this would be that rhythm "is apprehended within the span of the perceptual present", unlike the form and the meter that would "engage one's long-term memory of the piece at hand as well as one's musical background and knowledge." Similarly, Clarke (1999) makes the distinction between "small- to medium-scale temporal phenomena" (rhythm) and "large-scale temporal phenomena" (form). These definitions are foreshadowed in Fraisse's work. Fraisse defines the 'perceptual present' as "the temporal extent of stimulations that can be perceived at a given time, without the intervention of rehearsal during or after the stimulation." (Clarke, 1999, p.474). In other words, it is the dividing line between the *perception* of time (temporal phenomena extending to no more than about 5 seconds), and the *estimation* of time, that relies primarily on the reconstruction of temporal estimates from information stored in memory.

[1] See "Quantized durations" paragraph

Accepted for publication on May 9th 2002

Cooper *et al.* (1960) extend the definition of rhythm to all the temporal scopes of description (from single note to entire movement). They define rhythm as "the way in which one or more unaccented beats are grouped in relation to an accented one." Five basic groupings would be permitted, those of prosody (*iamb, anapest, trochee, dactyl, amphibrach*). In this definition, the underlying meaning of the word 'beat' is very broad. Notes could be grouped together according to their respective accentuations, as could be groups of notes, phrases and so on (p.6).

Accent

'Accentuation' commonly refers to the human ability to perceptually apply a mark on some events in the musical flow, in opposition to other such events. The way we actually carry out this marking process is still not well understood. Pitch, intensity, duration, harmony or timbre perception certainly has an influence on our way to hear out rhythm from music (Thiemel, 2001). But one cannot state unequivocally that one of these factors prevails, nor that these are the sole factors of rhythm perception. London (2001) defines an 'accent' as a "means of differentiating events and thus giving them a sense of shape or organization." One can also find various definitions of accentuation in the literature as reported in (Wittlich, 2001) /illustrations/accent.html.

Pulse / Beat / Tactus / Tempo

Cooper *et al.* (1960) define a pulse as "[…] one of a series of regularly recurring, precisely equivalent stimuli." "Pulses mark off equal units in the temporal continuum." The segmentation of time by a given pulse train defines one of the many coexisting levels of organization. A musical composition usually exhibits a reference temporal unit. The 'beat' corresponds to this preferred (or "most important") pulse. This notion of "most important pulse" is sometimes assimilated with the reference time division value used in standard Western rhythmic notation. It is often represented by the quarter-note (or 'crotchet'), or more generally by the denominator of the score time signature (*e.g.* if the time signature is 12/8, then the beat would be an eighth-note). Given a temporal unit (a beat), the 'downbeat' corresponds to another important pulse: the first beat in a 'measure', which is a special grouping of beats as we detail below. Any other beat in a measure is called an 'off-beat'. The beat that immediately precedes the downbeat is the 'upbeat'.

However, such a purely formal definition of the beat does not hold as it is generally considered that "the sense of pulse may exist subjectively" (Cooper *et al.*, 1960). In addition, humans tend to perceive a pulse in stimuli that were not generated with a "rhythmic intention" (*e.g.* a watch, machines, ocean waves, etc.).[2] In this context, the "preferred beat" corresponds to the pulse that can be apprehended in an immediate, or reactive, fashion (see aforementioned notion of 'perceptual present'), and that leads one to tap his feet accordingly. Human anatomy and motor-behavior naturally account for pulses (walking, heart-beat, breathing). It is commonly thought that there is an intimate connection between these physiological properties and the reactive aspect of rhythm (Fraisse, 1982, pp.151-155), (Lapidaki, 1996, pp.40-47). Lerdahl *et al.* (1983) use the term "tactus" to refer to the perceptually most prominent pulse.

Finally, the 'tempo' describes the musical speed, *i.e.* the rate at which beats pass in time. It can be expressed in BPM (number of beats per minute). Here, 'beat' can correspond to the score temporal unit –in this case, one refers to M.M. tempo (Maelzel metronome)–. It can also correspond to the tactus; as Scheirer (2000, p.56) writes, "The tempo of a sound is the perceptual sense that the sound is recurrent in time at regular intervals, where the interval length is between 250 ms and 2 s."

We believe that the literature concerning computational models –that we review thereafter– lacks an explicit distinction in the terminology. Indeed, the 'beat' is the most commonly used word, and it is often unclear whether this would correspond to the tactus, the reference time division –*i.e.* the time signature denominator–, the downbeat or any pulse. Therefore, in the following, we will stress the implicit concept when not obvious from the context.

Meter / Time Signature / Measure

The notion of meter is quite difficult to define. A recurring idea found in the various definitions is that meter is a basic structure of pulses regularly repeated in time. As opposed to the rhythm, it would be a construct without reality in the stimulus itself, an abstraction from the stimulus properties (Clarke, 1999).

According to Cooper *et al.* (1960), "meter is the number of pulses between the more or less regularly recurring accents." In other words, meter is seen as the regular repetition of a grouping of beats. However, this definition does not specify whether this would mean that the "meter imposes an accent structure on beats" (Brown, 1993), or conversely that the meter would emerge from accentuations of the events (implicit to the excerpt's melody, dynamics, harmony and timbre). In this "chicken and egg" issue, some propose to define the notion of *metric accents* –see 'Rhythmic and metric accent' in (London, 2001)–. Thus, the meter would emerge from a regular recurrence of a certain type of accents (melodic, dynamic, etc.), and would subsequently define positions of metric accents. Once defined, metric accents could complete the existing accents (possibly be opposed to some of them), opening the way to new rhythmical enrichments.

Yeston (1976) defines meter as "an outgrowth of the interaction of two distinct levels (two differently-rated strata), the faster of which provides the elements and the slower of which groups them." This definition seems closer to the usual description of meter that can be found in a score, given by the time signature and the bar lines. The bar lines

[2] See 'Subjective Rhythmisation' (Fraisse, 1982, p.155)

Accepted for publication on May 9th 2002

define the slower of the two levels, the 'measure', and the time signature defines the number of faster pulses that make up one measure. For instance, a 6/8 time signature indicates that the basic temporal unit is an eighth-note (a "note" referring to a "whole", or "semi-breve") and that between two bar lines there is room for six of them. As mentioned above, the 'downbeat' corresponds to the first beat in a measure. Any other beat in a measure is called an 'off-beat'. The beat that immediately precedes the downbeat is the 'upbeat'.

Two categories of meter are generally distinguished: *duple* and *triple*. This notion is contained in the numerator of the time signature: if the numerator is a multiple of two, then the meter is duple, if not a multiple of two but of three, the meter is triple. For instance, 2/4 and 4/4 signatures are called duple, 3/4 and 9/8 are called triple.

Quantized duration

The quantization (or *rhythm-parsing*) task aims at notating the rhythm of each note in a sequence. In accordance with Western notation, the goal is to encode note durations by rational numbers (1, 1/4, 1/6, etc.) relative to the chosen beat interval, which is generally unknown, and must also be identified. (One could roughly interpret the instruction "quantize it" as "make it fit into Western music notation".)

Quantization entails two aspects: (1) seeking the reference value from which the durations will be expressed, the beat, and (2) transforming absolute durations in quantized durations, slightly changing them when they do not match with the beat grid.

Clarke (1987) defines quantization as the separation between two time scales inherent in music: the *discrete* time intervals of a metric structure, and the *continuous* time scales of tempo changes and expressive timing. Accordingly to this definition, Desain (1990) argues that "quantization is the process by which performed time intervals are factorized into abstract integer durations representing the notes in the score and local tempo factors."

If the tempo function is known, a local stretch of the durations can be achieved, in order to make the tempo constant, before the transformation step. Then, the intended tempo changes would be notated on the score (*e.g.* accelerando, ritardando).

For instance, let us consider the following sequence, proposed by Desain (1990):

476 : 237 : 115 : 135 : 174 : 135 : 155
: 240 : 254 : 118 : 112 : 118 : 138 : 476

The initial score sequence is:

500 : 250 : 125 : 125 : 167 : 167 : 167
: 250 : 250 : 125 : 125 : 125 : 125 : 500

The given reference value is: 500

The result of the quantization is:

1 : 1/2 : 1/4 : 1/4 : 1/3 : 1/3 : 1/3
: 1/2 : 1/2 : 1/4 : 1/4 : 1/4 : 1/4 : 1

Many quantizers can be found in the literature (see (Desain, 1992) for more details on the algorithms).

1.3 Temporal aspects of Music – Interdisciplinarity of the field

Temporal aspects of music have been studied for many centuries, within many disciplines and by means of different methodologies. Hence, we do not attempt to present a comprehensive review; instead, we aim at providing a few selected bibliographical pointers.[3] In order to roughly order the bulk of research, the following methodologies can be identified:

- Theoretical analysis
- Empirical investigation (experimentation over subjects)
- Computational modeling

Let us now briefly detail the disciplines concerned. It is obviously difficult to provide references that concentrate on one particular discipline; again, we solely intend to provide some pointers we believe of importance.

Naturally, Music Theory accounts for many contributions to the study of rhythm and musical time, see *e.g.* (Cooper *et al.*, 1960) (London, 2001) (Yeston, 1976) (Wittlich, 2001).

Some researchers point out that the rhythm concept should not refer to intrinsic properties of the musical notation system, but should rather refer to the experience aspects of listeners. As written by Honing (1993), "Most theories agree that there is more in music than what is written in the score. […] The question here is whether a piece of music resides in the notation, in the air, or in the people's minds, or in other words, whether music is cognitive or not." Hence, a very important body of work is concerned with psychological studies of time apprehension. The aim in this field of research is to explore the role of memory and perception and their interaction in time apprehension. This usually entails empirical investigations, see *e.g.* (Fraisse, 1982) (Clarke, 1999) (Desain *et al.*, 2000, Part III). In this type of work some of the central notions are:

- 'Perceptual present', 'subjective rhythmisation' (Fraisse, 1982).

[3] A following section reviews a specific part of rhythm researches.

Accepted for publication on May 9th 2002

- Just-noticeable difference in perceived rhythmic deformation, *e.g.* (Friberg *et al.*, 1995), (Perron, 1994).
- Temporal masking, *e.g.* (Turgeon, 2000).
- Rhythmic similarity, *e.g.* (Gabrielsson, 1973b and 1973a).
- Tempo consistency over time, *e.g.* (Lapidaki, 1996 and 2000), (Levitin *et al.*, 1996).
- Differences in perception according to the subject's musical background, age or sex, and the musical style of the stimulus, *e.g.* (Lapidaki, 1996, pp.53-64, chapter IV and 2000), (Drake, 1993), (Drake *et al.*, 2000), (Gabrielsson, 1973b and 1973a).

Other researchers focus on expert performance analyses, *e.g.* (Gabrielsson, 1999) and (Desain *et al.*, 2000, Part IV) or on providing rules regarding music performance (Friberg, 1997).

Physiology and studies of movement (motor control) are also concerned with rhythm, *e.g.* (Fraisse, 1982, pp.151-155), (Desain *et al.*, 2000, Part I), (Friberg *et al.*, 1999).

Rhythm studies can also be found in Speech processing (Lee *et al.*, 2000), Philosophy, Theology, Anthropology, (Desain *et al.*, 2000, p.xi), and Medicine (Kristiansen *et al.*, 1995).

Finally, the Signal Processing and Artificial Intelligence disciplines usually entail modeling approaches. In the former field, researchers focus principally on properties of the musical stimuli (and not so much on the musical notation system or the experience aspects of listeners). In the latter field, most researchers tackle the issue of rhythm perception understanding by building computer models rather than performing experimentation over subjects –with some exceptions that explicitly consider both aspects: the model and its empirical validation (Desain *et al.*, 1998)–.

We will provide a review of the computational approach only (thus focusing on synthetic approaches rather than analytic ones), and leave aside the theoretical musical analyses and the experimental psychological investigations.

2 Review of computational models

In the framework of music content processing, computational models of rhythm provide the representational elements over which content processing is achieved (as depicted by the arrows between blocks in Figure 1). In this section, we begin to detail what existing computational models can offer.

There are many models in the form of computer programs that seek the extraction of rhythm features from musical data. Descriptors are of different sort, from low-level ones describing in some way the "rhythmic strength" of musical excerpts, *e.g.* (Tzanetakis *et al.*, 2001), to highly abstract concepts: meter (Brown, 1993) or rhythmical patterns (Gasser *et al.*, 1999). But the beat and the quantized durations are by far the central notions in the computational modeling literature.

In different approaches, various models

- handle symbolic data (*i.e.* starting point is the second block in Figure 1) or audio data (starting point is the first block);
- focus solely on temporal features or also on other features (timbral, harmonic, melodic or dynamic);
- are stream-based or off-line (*i.e.* need the whole data to proceed);
- entail mainly signal processing techniques or mainly artificial intelligence ones.

One might be interested in a review of computational models for scientific or more pragmatic reasons. In particular, one may wish to develop novel algorithms for the emulation of human rhythm perception, or, one may simply wish to determine which of the existing models to implement for a specific application. In the latter case, the specificities of the application indicate which model to use. Indeed, the application might handle solely MIDI data, all the items might pertain to the same musical style (*e.g.* percussive music), or it might be designed for a restricted kind of user, etc. Hence, we provide details on the context of use of the models, as specified by their authors, and the performances (when originally provided).

In order to satisfy also the former interest, we believe that it is important to distinguish between (1) the aim to extract what would be phenomenally present in the signal vs. (2) the aim to identify in the signal the musical constructs that could exist in our minds.[4]

2.1 Models processing symbolic data – knowledge-based approach

Part of the models deal with symbolic data, such as MIDI or manually parsed scores –*e.g.* in the format of files containing solely onset times and durations (Brown, 1993)–. All models handle timing data by means of onset times, inter-onset intervals or note durations. Few also take into account other event features (harmonic, dynamic, timbral or melodic).

[4] This difference echoes the discussion existing in the psychological researches regarding whether the apprehension of rhythm would be a physiological or a cognitive process (Fraisse, 1982), (Clarke, 1999), (Levitin *et al.*, 1996), (Lapidaki, 2000).

Accepted for publication on May 9th 2002

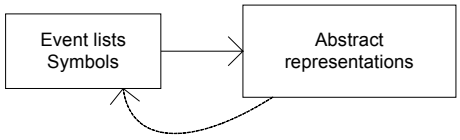

Figure 2: Models processing symbolic data – knowledge-based approach.

These models are typically rule-based. Most models process data off-line and make use of AI techniques. Some preprocessing of the symbolic data yields diverse hypotheses for a rhythmic element (*e.g.* beat, meter or quantized durations). Subsequently, a set of rules permits to rank these hypotheses.

For instance, Longuet-Higgins *et al.* (1982) propose a model of beat induction dealing with onset times of monophonic onset sequences. In this model, a pulse is first assumed and then adapted to the data by a sequence of slight changes that follow a set of "if-then" rules previously defined (*e.g.* 'stretch', 'shift', 'divide' or 'confirm'). Inputs to these rules are the beat assumptions and the time differences between events.

This family of models can also seek other rhythmic features than a single pulse. In fact, any temporal organization of events (*e.g.*, onsets or groups of onsets) one can formalize could be embedded in such models. Depending on the model and on the rhythmic feature sought, the rules make more or less use of musical theories.

Formalizing the process of quantization, some models aim at recovering score rhythmic representations from performance data. For instance, Longuet-Higgins (1987) makes use of a predetermined beat and some knowledge about meter. Basically, the intent is to find the best match to the onset time sequence among several metrical grid templates (Onsets are considered with a "tolerance interval".)

Rosenthal's goal (1992) is similar. He explicitly argues that human rhythm perception would follow a formal quantization process (p.66). His model makes several hypotheses regarding several "rhythmic levels", or pulses. Recurrent time intervals are considered as potential pulses. Here, inter-onset intervals refer to *successive* onsets (p.65); however, Rosenthal also considers intervals between non-successive onsets as potential rhythmic levels (p.67). Different organizations among pulses, or pulse-grouping hypotheses, are considered. Finally, rules (designed by Rosenthal to mimic rhythmic interpretations of human listeners) permit to rank the multiples hypotheses.

More recently, probabilistic approaches have been proposed to the issues of beat tracking and quantization of performance data. In this powerful framework, knowledge is not explicitly constrained by fixed, deterministic sets of rules; rather, it lies in probabilistic models learned from measured data during a supervised training step. Knowledge is thus less subjective and represents more reliably the training data.

Cemgil *et al.* (2000) define a "*metronome model*." Beats of this metronome are supposed to be "noisy", they are considered observations of a hidden variable to be finally estimated: the tempo. A dynamical system is defined with two 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. Their model is fully determined if the initial state variables are given. To this deterministic model, they add a noise term that models the tempo variations (a Gaussian random vector which covariance matrix will be estimated) and a noise term that models the performance timing deviations (a Gaussian random vector which covariance matrix is set). The metronome beats are not directly observed, but *inferred* from a list of onset times. That is, the goal is to compute a probability distribution for the tempo track –random vector {pulse phase, pulse period}– given the observation of a list of onset times. It is achieved by taking into account a Bayesian framework, *i.e.* the tempo track distribution is computed as a posterior distribution from a likelihood distribution and a prior distribution, the former being the probability of the performance given a tempo track and the latter being the tempo track a priori probability. The likelihood distribution is computed from what they call a "Tempogram" (which is computed by making use of locally constant tempo tracks –*e.g.* 5 beat-long– and a continuous, somehow "smoothed", representation of the onset times).[5] The prior distributions are assumed to be flat (*i.e.* all tempos are equiprobable). Once the metronome beats inferred, they are used as the noisy inputs of the dynamical system. Recall that the (inferred) beats are the "observations" of the model. Then, the parameters of the model remain to be estimated: the tempo – actually log(tempo)– is a hidden state variable which distribution parameters are estimated by means of Kalman filtering. What they finally seek is the pulse train that best explains the inferred (noisy) beats. They choose to orient the inferring of beats on (1) the quarter-note level and (2) the bar level. In both cases, they do not claim that the pulse trains sought would be related to the perceived beat. Reported results for beat tracking are very good.

Also following a probabilistic rationale, Raphael's approach (2001) enhances the dynamical model proposed by Cemgil *et al.* (that uses the output of a "measurement model" as observations of a hidden variable) in adding a discrete hidden layer. This permits to consider the estimation of variations in the beat rate (evolution of a "tempo process") jointly with the note quantization (evolution of a "rhythm process"). Raphael's rhythm parsing model takes as input a sequence of onset times and outputs a complete rhythmic transcription of the piece (*i.e.* it assigns a score position as {measure number, measure position} to each onset). He motivates and addresses the problem of simultaneous estimation of tempo –understood here as the rate of the notated bar pulse, or downbeat– and notated rhythm –estimation understood here as quantization–. In Raphael's opinion, while many authors identify the interdependency of tempo and

[5] Resembling Longuet-Higgins's (1987) "tolerance" parameter

rhythm, most of them estimate these two quantities independently. Either they focus on beat tracking without addressing the quantization issue or they quantize the observed note lengths to their closest note value assuming tempo has already been computed (either as a single stable value or a function of time). One might notice that Raphael's system needs strong prior information about the piece to be analyzed:

- The Markovian transition matrix for the rhythm process must either be chosen manually with strong high-level a priori knowledge about the piece (*e.g.* the score), or learned from a corpus of similar pieces (in the paper, to analyze a Chopin mazurka, Raphael trains his system with a corpus of scores of other Chopin mazurkas).
- The set of possible positions in the measure is also given in advance (*i.e.* strong assumptions are made on the time signature and the smallest possible interval between onsets). The author suggests no way to initialize it in an a priori manner, nor does he discuss effects of bad initialization.

However, the system seems to be very effective and Raphael argues that possible applications for his system are automatic transcription from audio (provided a previous segmentation step), score writing from MIDI performance, musicology (*e.g.* to study expressivity and changes in timing) and retrieval from large databases.[6]

Dixon *et al.* (2000) and Cambouropoulos (1998) propose models in which beat tracking (referring here to M.M. tempo) and rhythmic structure extraction are performed on MIDI sequences of performance data. The novelty in their models is that they account for several accentuation phenomena, by handling simultaneously several MIDI features (dynamic, melodic, etc.). (Interested readers should refer to Parncutt (1994) and Snyder *et al.* (2001) who propose interesting discussions regarding what feature might be considered a valuable cue for pulse finding.)

In Dixon *et al.* model, beat tracking is performed in two steps: first, a beat rate is assumed and then the algorithm tries to propagate the beats along the analyzed sequence (*i.e.* it chooses events in the sequence which could correspond to the beat occurrences). They put a special focus on the detection of the beat phase, arguing that it calls for some prior musical knowledge: accentuation rules. Given a sequence of events (and a set of features like duration, dynamics, pitches, etc.), a weighting system gives more importance to some events. If several events are likely to correspond to the temporal occurrence of a beat, the higher-weighted event will be preferred to the other ones, even if not corresponding exactly to an expected beat position. They provide details of interesting systematic evaluations over a large corpus; their model seems to handle very well tempo variations.

Cambouropoulos (1998) also proposes to differentiate between accented and unaccented events in making use of several accentuation phenomena. In order to extract meter and rhythmic patterns, he proposes to find the best metric grid (duple or triple) that maximizes the number of accented events of the sequence corresponding to the weighted events of the grid. The accentuation of the sequence follows a set of basic rules that give importance to some events whose feature values are compared to their neighbor's.

2.2 Models processing symbolic data – Bottom-up approach

At the extreme, symbolic data processing models could consist of a large look-up table of the possible temporal structures one would a priori know that are characteristic of the data to analyze. This obviously lacks generality. Precisely, Desain (1990) and Desain *et al.* (1991) argue that applying fixed rules in a top-down way would not be a suitable approach to the seeking of human rhythm experience modeling. They criticize AI symbolic models, arguing that they lack generality, and that a good description of human rhythm experience should account for an adaptive (or learning) feature. The connectionist (or sub-symbolic) models they develop can learn from the signal, can produce expectancies and build their own understanding of the temporal structures, focusing on inter-onset intervals extracted from MIDI data.

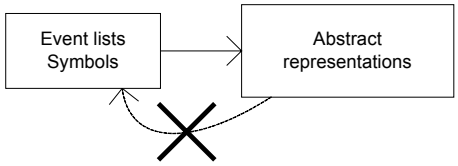

Figure 3: Models processing symbolic data – Bottom-up approach.

Rowe (2001) comments that two main properties distinguish sub-symbolic processes from symbolic ones: "First, sub-symbolic processes learn their behavior from exposure to material; and second, this learning engenders models that do not rely on a fixed set of rules." An interesting review of the connectionist approach to rhythm is provided in Large *et al.* (1994, pp.7-10).

This family of models process data in a streaming paradigm and focus on expectation. Indeed, Large *et al.* (1994) argue that "an advantage of this approach [on rule-based analyses] is that it handles the problem of metrical preferences through real time processing constraints, rather than by global evaluation of alternative constructs." Here also, models handle timing data by means of onset times, inter-onset intervals or note durations.

[6] We will later comment that provided the derivation of transcriptions (in the sense of usual score) from audio signals would be effective, its usefulness in the context of audio browsing and retrieval is not any obvious.

Large *et al.* (1994) focus on the "perception of metrical structure". They aim at constructing networks of basic oscillatory units, or resonators; these units having the principal feature to "embody the notion of metrical pulse, or beat." They investigate the modeling of the metrical structure elements (the diverse coexisting oscillatory units) as a first step to the modeling of the structure itself (the interconnection between units).

The model of Large *et al.* takes inter-onset times as input. A simple oscillator, called the "driven" unit, exhibits an intrinsic oscillation period. The "driver" couples to the driven unit, emitting a series of discrete pulses corresponding to the observed inter-onset sequence. Each pulse of the driver perturbs the phase of the driven to an amount determined by a coupling strength. The resulting instantaneous period of the driven eventually differs slightly from its preferred period without coupling.[7] The stability of such a system is function of the driven/driver period ratio and the coupling strength (the article provides insightful diagram illustrations). In order to prevent the oscillator to return to its former period if the driver stops, a frequency-locking procedure is also needed. Phase and frequency locking is achieved by minimizing the gap between the current next-event-prediction of the oscillator and the actual subsequent pulse of the driving input, according to the method of gradient descent. If this gap is too big, the new pulse will not be taken into account. Several coupling values are tested and the results are detailed in Large *et al.* (1994).

They demonstrate that "individual units can isolate periodic components in complex rhythmic patterns", however they "stopped short on proposing a theory of musical meter" (p.27), as the network construction was proven a very difficult task. The authors remark that their contribution solely addresses the task of pulse-induction.

It should be noted that studies on the issue of connecting several basic oscillators (to perform higher metric level modeling) can be found in (Gasser *et al.*, 1999) and (Eck *et al.*, 2000). Large *et al.* (1994) remark that other musical parameters (*i.e.* pitch, dynamics, etc.) should be considered. Their integration may considerably enhance the model.

McAuley (1995) addresses the issues of tempo induction and tracking. Here also, the model is based on an oscillator adapting its phase to an input made up of discrete pulses. A salient aspect of McAuley's research is the systematic aim to validate modeling by means of comparisons with listening experiments.

McAuley argues that the modeling of meter perception necessarily implies many oscillators. He agrees with Large *et al.* (1994) that a model with several independent oscillators is insufficient, the oscillators should interact in some way. In his opinion, however, it is not clear in what ways the oscillators should interact; possibly, interactions could be "guided by human performance data for rhythms of increasing complexity" (p.127).

The Self-Organizing Network of Oscillators for Rhythm (Gasser *et al.*, 1999) achieves rhythm pattern recognition following a connectionist rationale for learning. This model uses oscillators to model rhythmic structures at higher levels than that of a single pulse. The model assumes that the perceptual accentuation of rhythmic events in music lies in inter-onset intervals. However, they also account for *dynamics* in the input sequence (input pulses have different amplitudes). The model is a network of oscillators that resemble Large *et al.* (1994) or McAuley's (1995) oscillators. In the model of Gasser *et al.*, coupling concerns solely oscillator phases, not their periods. Coupling also exists between oscillators: they influence each other by phase-pulling. The novelty in their approach stands in the fact that connection weights between oscillators are trainable.

The simulations they discuss show interesting behavior of the model, however, it is not clear how other features than dynamics could be considered in this model (indeed, in more realistic inputs, a metrical pattern could emerge from a signal not showing a clear strong-weak relationship in the pulse amplitudes).

Different approaches are based on autocorrelation calculations. Here, the intent is to measure the periodicities inherent to the input signal.

Autocorrelation has been used in some beat rate extraction models, most of them referring to the approach of Brown (1993). Brown proposes to compute a sample-by-sample autocorrelation of a sequence of onsets (with a sampling rate of 200 per second), weighted by their durations. The various maxima of the autocorrelation graph are interpreted as metric positions. However, when intending to derive the score reference beat from these graphs, the experimentations show that the method does not always provide the expected beat, but multiples of it. Moreover, no phase information is extracted.

The autocorrelation method does not detect approximate repetitions. Thus, the beat rate cannot be detected in performance data which contain slight expressive deviations. In order to solve this problem, Tanguiane (1994) proposes the method of variable resolution which changes the definition level of the durations. Considering a sequence of durations, Tanguiane proposes a coding which approximates the values in such a way to enable comparisons. For instance, the duration 14 ms will be seen as: 50% of 14 ms plus 25% of 13 ms plus 25% of 15 ms. This aims at coding the fact that note attack-times and durations in performance signals always slightly differ from their formal values. Adopting this coding, 14ms (*50*14+25*13*+25*15) and 13ms (*50*13*+25*12+*25*14*) show common components that will be detected by auto-correlation.[8]

Desain *et al.* (1990) propose a method using a moving window, in order to simulate a kind of temporal context: a beat is expected, the window is moved to this beat, and a new beat is calculated. This method uses the auto-correlation coefficient in a framework based on expectation.

[7] In this "phase-pulling" scheme, if the driver stops (*i.e.* no more input to the driven), then the driven recovers its preferred period.
[8] Here also, one might think about Longuet-Higgins's (1987) "tolerance interval" for durations.

Accepted for publication on May 9th 2002

These approaches, in the authors' opinion, provide useful intermediary information, but the resulting graphs should be parsed in order to actually derive rhythmic information. That is, peak-picking has to be performed on the autocorrelation function and peaks must be subsequently interpreted with respect to what periodicities one wishes to identify. To this end, an additional set of rules or heuristics may be required.

Finally, another interesting approach is that of Smith (1996) and Smith *et al.* (1996) who handle lists of onset times extracted from monophonic drum MIDI signals, weighted by their amplitudes (*i.e.* MIDI velocity). They perform wavelet analysis to explore the concepts of architectonic rhythmic strata. They show that the wavelet analysis is well-adapted to capture time-organizations at different scales. The choice of wavelet representation is not made to suggest that human perception actually proceeds by means of such signal representation; rather, "the intention is to make explicit that information which is inherent in the rhythm." Thus, this approach succeeds in recovering from the signal what has been intended (hierarchies between the different organizational levels) rather than what is being perceived.

Here also, the authors suggest that subsequent peak-picking algorithms and heuristics would probably be needed to give a musical interpretation to this type of signal representation (*e.g.*, derive the tactus or meter.). ("While it is tempting to draw hypotheses for methods of derivation of the tactus by 'ridge-tracing' or the well-formedness of the global continuation of a voice, further research is required to build a model of tactus in respect of perceptual issues").

Dealing directly with audio data was not thinkable when the first models were designed –*e.g.* (Longuet-Higgins *et al.* 1982)–, hence the handling of symbolic data, presenting musical information in a compressed form. However, at some point in computer hardware progress, it became possible to manage large amounts of audio data as easily as symbolic data. Some attempts to merge the previous models with audio data exist. Analyses are achieved over sequences of symbolic events "filtered" from the audio and assumed to convey the predominant rhythmic information.

2.3 Models processing audio data – knowledge-based approach

Typically, models handling audio data account for several features. Some processing determines onset times. Regions around the onsets are then characterized in terms of note, instrument or chord. Subsequent to this signal processing front-end, AI formalisms permit to achieve musical analyses of the event lists, usually offline.

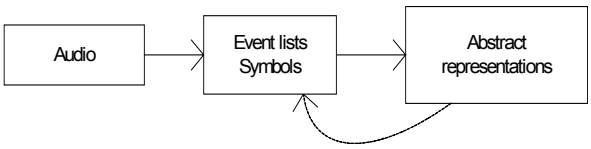

Figure 4: Knowledge-based approaches handling audio signals.

Chowning *et al.* (1984) and Schloss (1985) propose models that focus on monophonic signals: percussion and piano melody lines. Transcription of performance excerpts is their primary goal. On the way to determine quantized duration, their models intend to extract diverse pulsations, the tempo, variations in the tempo and the time signature. This process entails a set of rules that handle lists –"acoustic maps"– of attack times and several types of accents –dynamic and timbral (for the percussion) or melodic (for the piano)– that are derived from acoustic data in a bottom-up manner. It should be noted that their models account a tolerance degree in the onset times –making use of a continuous (smoothed) representation of durations (Chowning *et al.*, 1984, p.18), (Schloss, 1985, p.90)–. Unfortunately, they provide evaluation of their scheme based only on one excerpt.

Goto and Muraoka have contributed many models to the beat-tracking literature. Their models operate in real time and handle polyphonic audio signals.

Goto *et al.* (1997) focus on drumless signals. Events in such signals are grouped by their harmonic roles. A front-end performs onset detection and frequency analysis. Then, a multi-agent architecture makes decisions regarding multiple chord change hypotheses. In this architecture, an assumption is made on the meter, which must be 4/4 (*i.e.* the beat refers to the quarter-note and a measure is made up of four beats). In addition, the tempo range is constrained to 61-120 beats per minute. In this multiple-hypotheses architecture, a set of rules determines the behavior of the agents and the manager. For instance, chord changes are assumed to occur on the first beat of each grouping more frequently than on the third beat.

The aim is to account for different rhythmic levels, *i.e.* find quarter-note positions, half-note positions and measure positions. Reported results on 40 commercial songs are very good.

The BTS system by Goto *et al.* (1995) addresses music where the beat is maintained by drum sounds. Here, in addition to onset detection, the front-end performs a discrimination between drum sounds (bass-drum or snare-drum) –based on a frequency analysis of regions surrounding onsets–. Given these event lists and restrictions on the time signature and the tempo range –similar as above–, templates of bass-drum/snare-drum structures are used to determine the relevance of the detected onsets. Eventually, a multi-agent architecture makes decisions regarding the diverse hypotheses.

Accepted for publication on May 9th 2002

Evaluations were achieved on 44 commercial songs containing drums and whose tactus corresponds to the quarter-note.[9] Here also, results regarding predictions of quarter-note positions seem quite good. Unfortunately, there are neither details nor evaluation of the drum-sound classification issue.

The approach by Laroche (2001) is comparable in that he intends to recover templates of organization of beats in the transient timing data. He also makes the assumptions of 4/4 time signature, constant tempo and tempo range fixed (70-140). He makes use of a probabilistic approach that is somehow reminiscent of that of Cemgil *et al.* (2000). The temporal structure of four successive onset groups is represented by a probability function as the concatenation of four Gaussian distributions. The parameters of this function that are left to be estimated from actual transient timing data are the tempo, the "swing" and the position of the first beat. In order to estimate the parameters that provide the best match between the function and the input data, Laroche makes use of a maximum-likelihood method.

The notion of beat is central to Laroche's article. Indeed, transients are assumed to occur on beats, template structures of groups of four beats are sought, and the swing is defined at the "quarter-beat" level. Unfortunately, in this article, the notion of beat stays ill-defined: on one hand it seems to refer to the perceived tempo (the pulse at which one taps his foot) and on the other hand, it seems to refer to the notated bar –or downbeat– (see figures 2 and 3 in Laroche's paper). Moreover, if one wants to match his definition of swing to the commonly accepted one,[10] the word "beat" should refer to a half-note.

Nevertheless, provided some tuning of the parameters to a given musical style (and provided the data to analyze are consistent regarding the hypotheses made), Laroche's model seems to provide good results in the automatic location of the downbeat and the characterization of the swing.

As described above, Raphael (2001) proposes a probabilistic solution to the problem of quantization and beat-tracking. Initially, his model handles MIDI data, nevertheless, he also reports on an experiment using onset times extracted from an audio signal by the algorithm detailed in (Raphael, 1999). His model outputs a complete rhythmic transcription of the piece (*i.e.* it assigns a score position as {measure number, measure position} to each onset).

2.4 Models processing audio data – Bottom-up approach

Again, the chief goal of the "bottom-up approach" is to obtain lists of onset times and onset features from audio data. In contrast with the previous models, here, the handling of these event lists is rather bottom-up oriented and the methods are signal processing ones. Generally, there are less high-level assumptions on the signal structure. Rather, the aim is to identify periodicities that would be inherent to the signal.

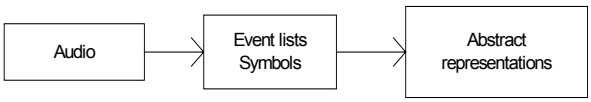

Figure 5: Bottom-up approaches handling audio signals.

In the context of an audio signal classifier based on low-level features, Scheirer *et al.* (1997) propose a feature called the "pulse metric." Its computation makes use of onset train autocorrelations. In a similar context, Tzanetakis *et al.* (2001) compute an 8-dimensional feature vector based on the handling of a "beat histogram" that captures some rhythmic aspects of audio signals. To compute the "beat histogram", peaks are detected in short-time autocorrelation functions (computed over 3-s windows of an amplitude envelope signal); the periodicities corresponding to the peaks of the consecutive windows make up the histogram. In both cases, although the features lack abstraction –they do not refer to explicit rhythmic concepts– the models seem to capture successfully the "strength" of the pulses and seem effective in the task of discriminating among some types of signals (speech/music or Classical/Pop).

Also in the framework of rhythmic similarity between polyphonic audio items, a recent contribution is that of Foote *et al.* (2001). Interestingly, here, the rhythm is not assumed to be conveyed by onset features, no prior symbolic transcription of the audio is sought whatsoever. The approach of Foote *et al.* is based on self-similarity of the audio data. They seek "repetitive events (even silences)", these events must not necessarily be notes, they range from sole 11ms-long frames to collections of such frames, parameterized by the magnitude of the signal's Fourier transform. Then, they build a matrix where each element represents the similarity between two frames and is computed as their normalized product (or cosine distance). From this matrix, they propose two ways to derive a measure of self-similarity, *i.e.* a function called the "beat spectrum": performing either sums or correlations of the matrix diagonal elements. Interestingly, the first of these two options can be seen as a continuation of an autocorrelation-based approach –as (Brown, 1993)–; indeed, the sum over the i^{th} diagonal is similar to the autocorrelation of the signal frame parameters with a lag i (with the difference that the correlation doesn't account for a normalization process). The approach looks promising but some of the authors' claims should be validated with additional, large scale, real-world evidence: Namely

[9] And, presumably, whose bass-drum/snare-drum patterns match those used in the model.

[10] The swing is a musical attribute that originates in jazz music. In scores that intend to notate it, usually, straight eight-notes are written, but an indication is given that in a group of two eight-notes, the first should last longer than the second, the amount being left to the performer.

that (1) the approach would adapt to any music style, (2) the tempo derivation would be similar to that performed by humans, and (3) the approach could provide descriptors of rhythm suitable for similarity and retrieval purposes (*e.g.* global rhythmic similarity (Foote *et al.*, 2002), or "swing").

Dixon (1999) addresses the beat-tracking issue, focusing on onset times. A clustering technique organizes onset time differences (within 5-10 second windows)[11] into groups. This algorithm permits to consider slightly different intervals as instances of the same class. This approach is in some way reminiscent to that of Chowning *et al.* (1984, p.18) and Schloss (1985, p.90) who consider onset times with a tolerance degree, which allows for a continuous (smoothed) representation of durations.[12] The beat is chosen among these interval clusters by simple ranking of the clusters –in a way, this is also comparable to Schloss' (1985, p.89) determination of the "important duration"–.[13] The system is evaluated over 6 polyphonic excerpts and seems to yield quite good results.

In the context of polyphonic signals, some researchers propose a simplification of the types of rhythm to be addressed, and target applications dealing with popular music. They argue that in an important part of the popular music repertoire, the rhythmic feature is mainly given by specific timbres.

Alghoniemy *et al.* (1999) and Blum *et al.* (1999) focus on the low-frequency components of the spectra.

Pachet *et al.* (2000) and Gouyon (2000) –as Goto *et al.* (1995)– base their approach on the automatic extraction from the audio signal of lists of occurrences of percussive timbres. The issue of drum timbres classification within complex signals is addressed by Gouyon *et al.* (2000). The authors argue that a major issue lies in a characterization of percussive timbres that should be relative to musical excerpts rather than absolute and depending on predefined clusters in a "freezed and universal" timbral space. Therefore, they discriminate occurrences of snare-like and bass drum-like timbres by means of non-supervised clustering techniques. These time series of occurrence indexes are subsequently processed by means of correlation computations to design rhythm descriptors. The determination of two series of indexes is based on the progressive identification of the source sound (the percussive sound to find) during the analysis process. (More precisely, templates of synthetic sounds are refined within several iterations, in order to most closely match the actual recurrent percussive timbre of the audio signal.)

In summary, bottom-up models typically adhere to the following approach. First, simple assumptions are made on what element carries the rhythm. This information is isolated from the audio flow (creation of event lists). Handling event lists, some signal processing method permits to highlight periodicities, *e.g.* correlation or Fourier transforms –as in (Blum *et al.*, 1999)–. Then, as mentioned above, an issue in these models is that deriving the beat from the resulting signal representations usually requires peak-picking and ad-hoc procedures. This gets even more necessary when considering more abstract concepts than the beat. Therefore, it seems difficult to design an efficient pure bottom-up procedure.

It is difficult to compare these beat-tracking models as there is no shared benchmark for evaluation. Nevertheless, bottom-up models are generally prone to the same types of errors: the extraction of integer multiples or sub-multiples of the actual beat rate (typically two times as large or small). Here again, it is not clear how one could handle this problem without accounting at some point for heuristics, and therefore without making use of AI formalisms. Precisely, the model by Dixon (2000) is an interesting step towards the integration of a "light" formal aspect in the beat-tracking issue.[14]

2.5 Implicit symbolism (the "transcriptive metaphor")

The previous organization of models parallels the classical "bottom-up vs. top-down" debate in cognitive science. In the field of human auditory perception modeling, part of the researchers espouse the bottom-up approach (Desain, 1990) and others argue that top-down flows of information should also be considered (Slaney, 1995).[15]

Scheirer (2000, p.78) favors interactions between these two types of processes. But the central argument in his discourse is that many bottom-up approaches to auditory perception modeling would still entail an implicit symbolism. Indeed, one might notice that there still is a common element to all the models described above: the handling of symbols, either as a starting point or a mid-level representation.

---

[11] As Rosenthal (1992, p.67), Dixon includes time differences between all pairs of onsets to the definition of inter-onset intervals.

[12] See also (Gouyon *et al.*, 2002)

[13] Or the most frequent inter-onset interval (MFIOI) in (Gouyon *et al.*, 2002)

[14] In the sense that "common-sense" rules are preferred to "heavier" musical knowledge

[15] Bottom-up models assume that human auditory system would be similar to a chain of elements, information being transmitted from one element to the next in a bottom-up flow. At each step, information gets more symbolic, that is, carries more semantic meaning. Quoting Ellis (1996), this approach states that "like human-engineered systems, hearing mechanism [would rely] on some underlying physics to glean information-carrying signals from the environment, which [would] then [be] processed to reveal specific details to be integrated into a simplified representation of the world." In contrast to this point of view, top-down approaches intend to recover in the external world –where are the stimuli– the internal representations, or images, of the perceptual apparatus. That is, for analyzing external stimuli, the perceptual apparatus would use patterns that would have been stored internally; here information flows top-down, from high levels of abstraction of the data (the brain) to lower levels (*e.g.* the cochlea). An analysis following this approach stands in a "process of reconciliation between observed acoustic features and the predictions of an internal model of the sound-producing entities in the environment" (Ellis, 1996).

Accepted for publication on May 9th 2002

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." In his opinion, when building a system that in some way mimic perceptual capabilities, one must put a special focus on the issue of what in this system is chosen to be innate or learned. Honing argues that "a distinction has to be made between […] the existence of possibly innate perceptual mechanisms and learned divisions of continuous time."

Scheirer criticizes the "transcriptive metaphor" according to which music perception would consist in a hierarchy of tasks, the first one being the transcription of audio data into an accurate note-list representation to be analyzed. He argues that solely well-trained musicians hear the music in terms of its conventional musicological structures, and that the "transcription" assumption may not be relevant regarding the actual perception of music by the human auditory system. Scheirer (1998) aims at confirming this hypothesis with a particular musical aspect of interest here, the tempo. He presented to listeners an amplitude-modulated noise constructed by "vocoding a white noise signal by the subband envelopes of the musical signal" instead of the musical signal itself. In this synthesized signal, the notions of onset features are purposely filtered out. From his conclusions, the sense of tempo was similar in both the cases.

He argues that tempo perception would be a low-level perceptual phenomenon, having little to do with cognition; thus, its emulation would not require Top-Down processes. Therefore, he proposes a reactive and representationless model of tempo perception, based on cross-channel integration of the energy levels of a small number –*e.g.* six– of broad bandpass filters. It processes the data in a continuous manner rather than deriving discrete symbols over which would be determined the high-level description, his model does not try to detect onsets.[16]

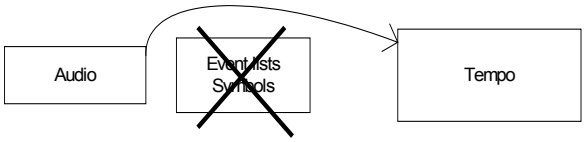

Figure 6: Modeling tempo-induction without resorting to the "transcriptive metaphor".

This model seems to work well for a wide range of musical styles (the 60 publicly-available examples cover monophonic and polyphonic music, styles that differ both in the *époque* and geographic location); it also matches rather well the responses of listeners. (It should be noted that Scheirer (2000, pp.91-92) provides an interesting comparison between his approach and the correlation-based approach.)

2.6 Computational models – Discussion

A scientific perspective

In the aim of designing models of rhythm perception, it seems necessary to account for information that would not be just local (that is, extracted from the signal under direct analysis), explicitly, information accounting for short-term context, as well as long-term context or acculturation. On the other hand, too much reliance on symbolism and musical formalism cause the resultant models to lack generality, and to be heavily biased towards "correct" musical structures, rather than what is actually played. As Desain *et al.* (1998) say, "the on-going nature of music perception or performance often relies on a combination of both bottom-up and top-down mechanisms: one cannot easily pretend that our musical knowledge is static, that it does not influence what we hear and how we play, or that it cannot be revised by our continuing exposure to musical stimuli." Following the "embodied cognitive science" branch of AI –or "new AI" (Brooks, 1991)– Scheirer's critic of the transcriptive metaphor also presents objections to both the pure top-down and the pure bottom-up approaches (2000, pp.74-78).[17]

Indeed, it seems worthwhile to consider interactions between top-down and bottom-up approaches in designing new models (Ellis, 1996).

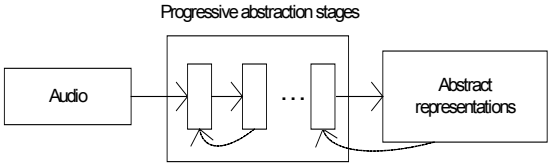

Figure 7: Interactions between bottom-up and top-down processes.

[16] One could see in the amplitude envelope extraction a focus on the onsets, but the important point to focus on is that there is no discretization of events performed on the envelope signals (no thresholding and peak-picking), the model deals with continuous data until the final decision stage.

[17] Eck *et al.* (2000) also embrace the embodied cognitive science framework. However one might remark that their model processes MIDI data.

Nevertheless, the issue of *learning* specifics of the generic models remains. Indeed, both top-down and bottom-up approaches assume that perception would entail several steps of abstraction of the information, organized hierarchically, but one may wonder whether the rules of abstraction, at any level, are evolutive or not. If the mapping between incoming and outcoming data between two layers of abstraction does not account for some kind of learning, or progressive adaptation to a specific type of stimuli, supposed to be recurrent in the external world, then the model still seems to entail symbolism. As Steels (1999) writes when favoring a "bottom-up approach to AI" –in the context of natural language modeling–, "in a way, our powerful engineering methodologies make it too easy to succumb to a strategy of programming directly the human or animal behaviors we observe and interpret as being intelligent. But doing this, we keep simulating the end products of intelligence rather than getting at the heart of intelligence itself. We put our own human concepts explicitly in the machine instead of implementing the mechanisms that enable an artificial agent to acquire new categories itself, implementing by hand a fixed set of predetermined behaviors which we believe the agent should have, rather than supplying mechanisms that allow the agent to acquire new behaviors when faced with unforeseen circumstances [...]."

In sum, it seems that the following research areas could prove fruitful: (1) interactions between bottom-up and top-down processes for the processing of natural stimuli –i.e. audio data– (*e.g.* like in speech recognition systems), (2) the production of rhythmic categories by learning from examples and (3) using pattern-matching techniques that could deal with uncertainty and noisiness in the data.

An interesting recent approach in these directions is that of Thornburg (2001b and 2001a). As Raphael (2001), he proposes to enhance the dynamical model proposed by Cemgil *et al.* (2000) by adding a second hidden layer (but this one explicitly discrete, in terms of notated musical intervals and metrical positions), in order to consider jointly the quantization and the estimation of variations in the beats rate. The additional novelty in his approach stands in the consideration, from the very outset, of the segmentation of audio data. In the framework of rhythm tracking, his model progressively learns how to produce relevant lists of onsets from the audio.

Noting that segmentation is a very difficult task in the context of polyphonic audio, Thornburg considers that uninformed segmentation without accounting for the structure of rhythm usually yields more than onset times.[18] Therefore, in his point of view, segmentation should not solely be seen as preprocessing, it should rather be intertwined with rhythm tracking. Both tasks should be considered jointly: polyphonic audio segmentation (deriving onset times) is a necessary step to be taken to provide data to the rhythm tracker; similarly, rhythm tracking should orient (i.e. provides priors to) the segmentation task. Thornburg refers to this rationale as "closing the loop." As "in the presence of rhythmic structure, the pattern of musically relevant change points, as note onsets, becomes highly regular", exploiting structure might help to "better adapt segmentation to the problem of onset detection", this is referred to "stream filtering." The complementary task of tracking rhythm –notated rhythm–, given a sequence of events, is tackled within a probabilistic framework similar, at first sight, to that of Cemgil *et al.* (2000), but which differs in that it has "multiple layers", similar to the approach of Raphael.

Recall that the "metronome model" by Cemgil *et al.* approaches the task of tempo tracking with a rationale that entails (1) a measurement model (achieving preprocessing of a list of onset times *e.g.* Tempogram) and (2) a dynamical system (using the output of the measurement model as observations of a hidden variable). The measurement model aims at preprocessing the list of events (onset times) and outputting the pulses of a noisy metronome. This model does not infer to what metric level (*e.g.* quarter-note, bar) correspond the metronome pulses. In other words, the measurement model (*e.g.* the Tempogram) provides a set of valuable pulses, but not which would be the best to focus on; at some point, the metronome must be manually oriented towards *e.g.* a half-note pulse.

Thornburg asserts that "a single metronome is insufficient to describe the succession of intervals present in most rhythmic structures." Therefore, he proposes to "short-circuit" the measurement model step, and to replace it by a switching state-space model (SSM): instead of deriving the pulses of one metronome from a temporal sequence of events, he proposes to "model event observations by switching among an array of metronomes, each specialized for a particular rhythmic interval." More formally, he adds an extra discrete-valued hidden layer that models the dynamic probabilistic structure of rhythmic interval successions conditional on previous interval successions and, of greater relevance to Western music, the current metrical positions.

It should also be noted that among other interesting aspects, Chowning *et al.* (1986, pp.15-17, 23-24) provide many insights regarding the implementation of multiple layers of data abstraction ("layered data representation"), the "convergence" of such an architecture and streaming-based learning issues (in opposition to offline learning).

Finally, Desain *et al.* (1998) might provide us with an interesting way to conclude the discussion on the design of new models of rhythm perception. They argue that contributions to the research regarding rhythm should now focus on homogenizing the methodology rather than proposing new models. "A computational model [should] no longer [be] an aim into itself, but a means to compare and communicate theories between different research communities."

[18] Indeed, as this task is usually achieved by detecting abrupt changes in a specific model parameters –the choice of the model depending on how the problem is primarily considered–, segmentation generally yields lists of changes time, some corresponding to onsets and others to *e.g.* timbral or pitch changes, and "only a fraction of detected change times are likely to convey rhythmic information."

Accepted for publication on May 9th 2002

Particularly, they convincingly maintain that computational theories of beat induction should be validated by means of empirical experimentations.

A more practical point of view

In contrast to the preceding, we wish to comment that some pragmatism might be acceptable in the context of specific applications. In the case where applications do not call for generality, it seems suitable, for the sake of efficiency, to embed some a priori knowledge in the algorithms or to focus on symbolic data. For instance, one might be interested in designing an application that would perform music file database browsing; one might know a priori the restrictions of the database, *e.g.* if all items have a binary meter and the styles are restricted. Another application might be specifically targeted towards a reduced set of users, *e.g.* music experts; here it seems necessary to account in some way for established musical formalisms rather than to focus on a learning paradigm. We reviewed many models reporting promising results for particular situations like these ones.

3 Representing rhythm for music cotent processing

In this section, we discuss the usefulness of several existing generic –decoupled from algorithms– representation schemes (*i.e.* the block "abstract representations" in Figure 1).

The first type of representation to be discussed is that of the direct output of signal processing analysis schemes. Such a representation is the basis of investigative material. It is typically a very thorough, low-level description of physical features, the qualities of which can be extracted unambiguously.[19] In our context, this will correspond to a representation of basic elements supposed to convey rhythmic meaning: *e.g.* transient times, dynamics, inter-onset intervals (IOIs), fundamental frequencies, etc. Objective relationships between elements can also be sought and constitute the basis of more sophisticated representations, *e.g.* time series of indexes of a recurring timbre (a snare drum for instance).

In addition to representing the characteristics of the *signal*, one may focus on the representations of the *music* that is conveyed by the signal. The issue here is to differentiate between (1) the *elements* representations and (2) the "*explicit structural representations*" (Honing, 1993) that represent the very structure that exists among elements rather than isolated elements. Here the idea of representing structures that exist in listeners' minds is introduced. This is precisely one of the objectives of a content-based representation, intended to be intuitive for the end-user. In a paper commenting on representational issues in the context of the MPEG-7 standard, Lindsay *et al.* (1999) argue that a perceptual representation can be more useful than a literal transcription. ("The sound is what people *perceive*, and therefore what they describe for search, in many cases.") Thus, they propose multiple representations of music, one of which is labeled "Perceptual".

As Lindsay *et al.* maintain, to be useful, a highly-abstract representation standard (*i.e.* that highlights semantic contents) should permit some *discrimination* between items. As introduced above, other requirements are:

- To enable (1) *exploration and understanding* of the data, (2) *browsing* among items, (3) meaningful *comparisons*, (4) musically- or sonologically-meaningful *transformations*.
- To be made up of elements *automatically derivable* from the data (at least some of them).

These concepts are often linked with the notion of distance between elements, and the idea of "sonic spaces" that would represent items spatially (in 2 or 3 dimensions). That way, exploration, browsing, comparison or transformation have a visual meaning, as in the numerous interactive graphical displays by Tzanetakis *et al.* (2001), or the "Song Surfer" by Cano *et al.* (Cano *et al.*, 2002), see also (Ó Maidín *et al.*, 2000). In other words, representing music contents suggests an association between distances in an intuitive visual space and the actual distances as measured in musically meaningful dimensions.

Naturally, one may wonder what features the above-mentioned computational models provide.

Making simple assumptions about the data, the state-of-the-art models can successfully track the frequency of beats in audio or MIDI files. These beats could correspond to any of the coexisting pulses (*e.g.* the denominator of a score's time signature or the perceived pulse). On the one hand, this feature is considered to be a useful descriptor for specific applications (Cliff, 2000), (Wang, 2001). On the other hand, this is not a discriminative descriptor.

Much stronger assumptions must be made for deriving quantized durations and meter, even from MIDI data.

One may ask what additional features should be sought for representation. As mentioned above, features proposed by Scheirer *et al.* (1997) or Tzanetakis *et al.* (2001) seem useful for classification applications. However, they are tightly linked to algorithmic considerations and do not refer to explicit rhythmic concepts.

Let us now review and comment on some generic proposals of rhythm representation spaces.

[19] Actually, extracting objective physical features is not necessarily that trivial a task. It calls for a previous segmentation in elements to be characterized. And "it is clear that there is still quite a lot of discussion and research needed, especially on the rules of segregation of acoustic signals, before we can decide on the discrete elements of a general representation of music" (Honing, 1993).

Accepted for publication on May 9th 2002

3.1 'Temporal sequence space'

Honing (1993) stresses the difference between the *declarative* and the *procedural* ways of representing knowledge: "declarative being the knowledge *about* something, while procedural knowledge states the knowledge in terms of *how to do* something." "Procedural representations (*i.e.* modeled as processes or procedures) are very powerful in modeling knowledge that is procedural by nature." If one assumes that a sense of rhythm is grounded in human cognition rather than in physiology, then a procedural representation would be more suitable. Precisely, Honing describes temporal musical structure "as a collection of structuring *mechanisms* that have time intervals associated with their components, […] the constraint on these time intervals [being what] specializes the different kinds of structuring" –emphasis ours–. Following this rationale, Desain *et al.* (1991) propose to represent rhythm by connectionist networks. The corresponding rhythm space, or "temporal sequence space" proposed in (Desain, 1990) is a 3-dimensional space of all possible temporal sequences of three inter-onset intervals (four onsets). Their model derives regions in this space that correspond to similarly-quantized sequences. The dimensions are:

- The length of the interval between the first onset and the second
- The length of the interval between the second onset and the third
- The length of the interval between the third onset and the fourth

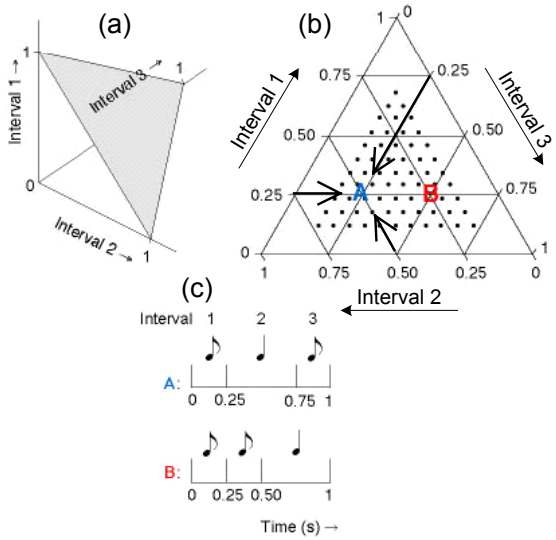

Figure 8: Rhythm space (a), ternary plot (b), and two pattern examples (c) (see http://www.nici.kun.nl/mmm/quantization-demo/index.html, Reprinted with permission. See (Desain *et al.*, 2002) for details).

In more recent works, they assess (by experiments over subjects) the perceptual coherence of the temporal sequence space regions.

Some aspects of this representation space seem to be very useful (*e.g.* it may be useful to check consistency when applying temporal transformations on a sequence). However, it is not really clear how it could be extended to handle more realistic stimuli (*i.e.* other than succession of four onsets). Moreover, as Honing (1993) says, not surprisingly, if "declarative knowledge tends to be accessible", in a procedural representation –as is this one– "deriving semantics is very hard." Thus, this rhythm space does not seem to fulfill all the requirements stated above.

3.2 MIDI representation of time and rhythm

MIDI provides a means of transmitting (and storing) *Tempo* (the number of microseconds per quarter-note), *Meter* (time signature), and the MIDI *Timing Clock*, all of which are dimensions relative to rhythm. The latter permits to represent time by a discrete temporal unit that depends on the notion of tempo (unlike the "MIDI Time Code").[20] This is a message –the status byte F8– sent from 24 to 480 times per quarter note. If the tempo changes, the MIDI Timing Clocks will pass at a faster rate, but the number of messages per quarter-note will stay the same.

This representation standard is rather designed as an efficient real-time communication method between electronic instruments than as an abstract representation standard. As Honing (1993) points out, "a distinction can be made between representations designed for real-time systems that are process-oriented […], and non-real-time systems that have a static global view of the music […]."

[20] The MIDI Time Code –based on SMPTE Time Code– is a representation of *absolute* time in that it follows hours, minutes and seconds and cannot be speeded up or slowed down.

3.3 MPEG7 representation of rhythm

It is surprising to observe that the MPEG7 standard, born 15 years later, which very aim is to provide semantically-meaningful representations of multimedia data, does provide similar dimensions to those of the MIDI standard. They are also very much anchored in a formal approach to music. The elements of this standard that convey a rhythmic meaning are embedded in the melody description:

- The beat
- The meter
- The note relative duration

Here, the beat refers to the pulse indicated in the feature 'meter' (which does not necessarily corresponds to the notion of perceptually most prominent pulse). The beat type is a series of numbers representing the relative positions of the notes in relation to the first note of the excerpt, the positions are expressed as integers, multiples of the measure divisor, the value of which is given in the denominator of the meter. The relative duration of note is the "logarithmic ratio of the differential onsets for the notes in the series" (MPEG7 audio team, 2001). The meter gives in its denominator a reference value for the expression of the beat series. The numerator serves, in conjunction with the denominator, to refer to pre-determined templates of weighting of the events. It is assumed that to a given meter corresponds a commonly-agreed "strong-weak" structure for the events. For instance, in a 4/4 meter, the first and third beats are assumed to be strong, the second and the fourth weak. In a 3/4 meter, the first beat is assumed to be strong, and the two others weak.

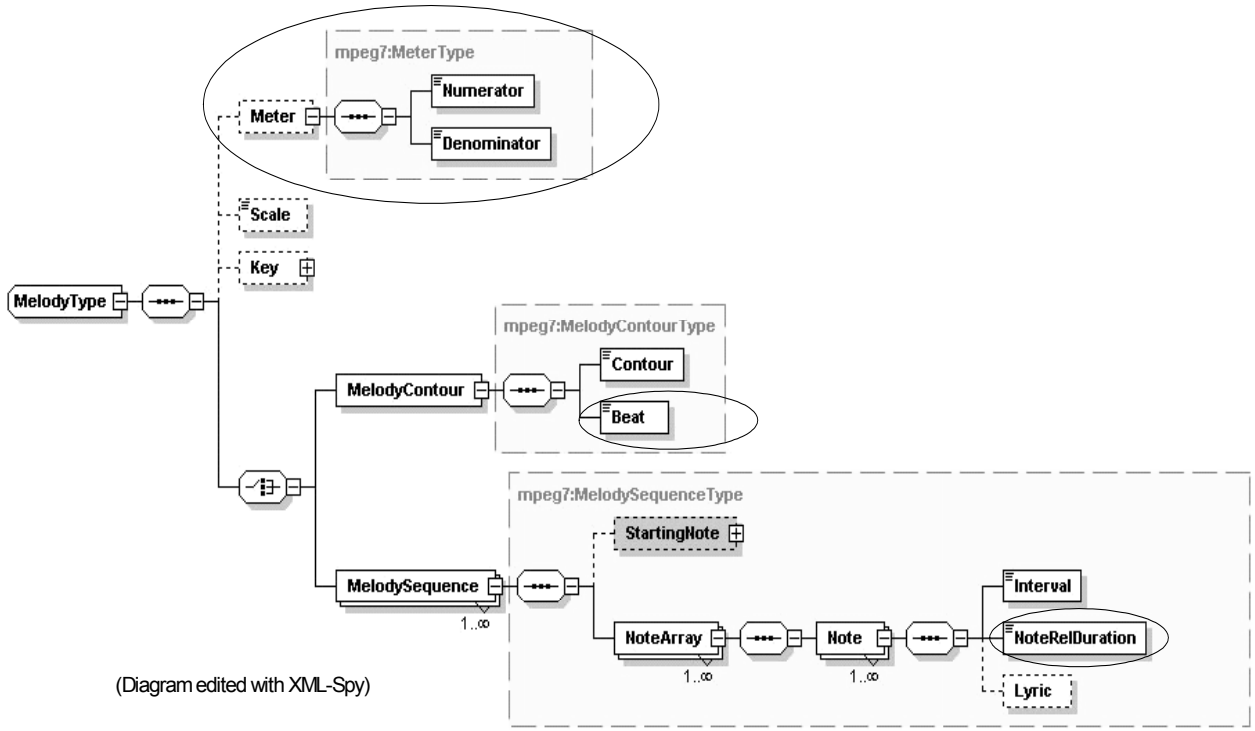

Figure 9: MPEG-7 elements of rhythm, in the melody description, see (Gomez *et al.*, this issue).

Comments that can be made about this representation are as follows: (1) The context is always that of a monophonic melody. (2) There is no direct information regarding the tempo, or the speed at which the pulses pass. (3) When attributing an integer value to an event (*i.e.* the position of the closest beat), there is a rounding towards -∞, thus in the case where an event is slightly before the beat (as happens in expressive performances) it is attributed to the preceding one. (4) This representation cannot serve for exploring fine deviations from the structure. Furthermore, as events are characterized by beat values, it is not accurate enough to represent already-quantized music where sub-multiples are commonly found (see quantization example previously given). (5) It is extremely sensible to the determination of the meter, which is still a difficult task for the state-of-the-art rhythm computational models. (6) There are no algorithms suggested for the determination of these features (even though this is not mandatory in the official MPEG references, informative extraction procedures are provided for many other descriptive features).

3.4 Formal representation

By 'formal representation', we are referring to the elements of a score: the time signature, bar lines, notational durations and tempo indication (Figure 10). Just as text data is not sufficient for text retrieval in large databases,[21] the use of score

[21] See *e.g.* (Baeza-Yates *et al.*, 1999)

data (that is, thorough scores written manually, in contrast to data derived from automatic audio analysis) does not allow for direct retrieval of musical information. Indeed, before doing retrieval, the score must be further parsed. For instance, the data should be segmented into coherent indexes structures over which "themes" can be sought –see *e.g.* (Melucci *et al.*, 2000) and (Smith *et al.*, 2001). Lemström *et al.* (2001) also propose to address the issue of "transposition invariance." Specific algorithms should also be developed in order to make the link between scores and sequences provided in a format defined by the user, *e.g.* MIDI or audio –see *e.g.* (Mazzoni *et al.*, 2001)–.

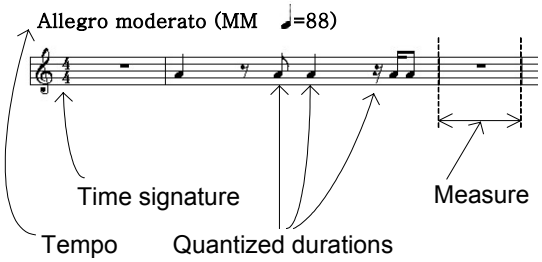

Figure 10: Score representation.

Moreover, it is commonly agreed that the automatic transcription of polyphonic audio is still far from being solved. Nevertheless, "[even if] musical listening systems cannot transcribe automatically (other than in the most simple cases), the majority of western classical music is scored and many of these scores exists in electronic form" (Lindsay *et al.*, 1999). Hence, in their proposal of musical representation, in addition to the already-mentioned "Perceptual" description scheme, Lindsay *et al.* take into account a description scheme labeled "Transcription."

3.5 Bilmes' rhythmic elements

Bilmes (1993) proposes to describe rhythm with the following elements:

- Metric structure (*i.e.* time signature and tempo)
- Tempo variation (*i.e.* changes in execution speed with time)
- Deviations (*i.e.* small time shifts from a beat grid)
- Ametric phrases (*i.e.* phrases that do not have an associated beat rate)
- Tatums (or "temporal atom", *i.e.* lowest level of the metric hierarchy, that is, pulse grid that most highly coincides with all note onsets)

This approach seems quite complete and effective for obtaining the information required to transform quantized musical phrases into musical phrases that are expressive. However, the author focuses on percussive music and does not claim that his algorithm is entirely automatic. It is given "the time signature, the number of tatums per beat, the number of beats per measure and where the beginning of the measure is." (p.59) Moreover, it is assumed that either the score or a reference instrument (one which is not subject to deviation) is available, from which a tempo function can be derived.

3.6 CNMAT 'Rhythm space'

This research group –see (CNMAT RRG)– is mostly interested in musical interactions (Wright *et al.*, 1998) and focus on expressivity features in rhythmic performances.

Iyer (1998, chapter II) proposes original definitions for rhythmic elements; he begins with Bilmes' elements of rhythm (see above) and from that derives what he calls a "Cell." The basic element in the representation of rhythm is "a data structure containing a duration, a tempo curve, and any number of "note layers." A given "note layer" contains either a discrete, regular tatum grid whose elements contain notes, or a list of notes occurring at fractional points of the cell duration" (Iyer *et al.*, 1997). Cells can have architectonic structures, *i.e.* they can be combined into larger Cells.

Accepted for publication on May 9th 2002

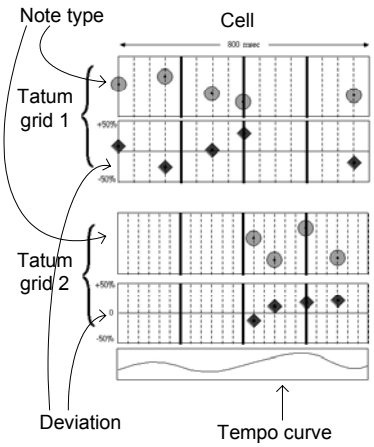

Figure 11: Rhythmic cells (See (CNMAT RRG), Reprinted with permission).

Although the approach seems to provide a powerful tool for dealing with rhythm while composing or performing (Wright *et al.*, 1998), the representation of rhythm they propose is used in real-time *production* systems and does not account for a global *analytical* view of the music –referring to the aforementioned dichotomy by Honing (1993)–.

However, concerning an analytical view of rhythm, they propose some dimensions and distances for a metric space –see (CNMAT RRG) /rhythm-space.html–. The rhythm of a monophonic musical signal would be representable on the following dimensions:

- "Duration"
- "Number of notes"
- "A set of note-on times"
- "A set of note-off times"
- "An underlying pulse, when applicable [a tatum or beat train]"

Hence, the number of dimensions for representation might depend on the excerpt analyzed. Moreover, one might ask whether these dimensions can provide an efficient way to discriminate rhythmic contents.

3.7 Representing Rhythm – Discussion

An important dimension for the representation of rhythm is the rate at which beats pass in time. Yet, this is not sufficient. In addition, the formal representation, though clearly a useful one, cannot be automatically derived and does not account for all the semantics of the data (e.g. no performance information). Likewise, other representation schemes provide interesting features, but none seem to completely satisfy the aforementioned requirements.

In seeking other descriptors of rhythm, it is important to consider the widespread idea that the perception of rhythm would be an analytical process grounded in cognition, *i.e.* using local feature computations in the context of short-term memory, as well as long-term knowledge (acculturation). Precisely, Scheirer (1998) argues that tempo induction would be a low-level perceptual process *separable* from the general perception of rhythm that would rather concern cognition, memory and learning.[22] We wish to push this argument further: Rhythm representation should take into account the fact that knowledge representations evolve as people acquire music skills, or get used to new musical styles (Drake, 1993), (Drake *et al.*, 2000).[23] Just as the idea of listener-dependent rhythm representation is widespread in the experimental psychology field, it should also be noted that Gabrielsson (1973b and 1973a) introduced the notion of signal-dependent rhythm representation (namely mono- or multi-instrumental).

In sum, the representation of rhythm should be manifold, as are actual scenarios of rhythm experience.

At the center of the content-based processing framework still lies an ill-defined issue. The disassembly of music into a rhythmic part, a melodic part, a harmonic part and an orchestration part (*i.e.* repartition of timbres) is an artificial framework that does not necessarily correspond to the human experience of music. Nevertheless, it is an analytical rationale suitable (and necessary) for designing investigation tools and applications. In the context of an application, it is probable that we won't ever be able to perfectly decorrelate rhythmic aspects from –say– harmonic ones, but on the other hand it does not seem suitable to aim at analyzing, representing, modeling, and generally processing Music as a whole. A more achievable objective is to determine and isolate some features, commonly experienced as "having something to do with rhythm", that will make sense from a *practical* point of view. Similarly, to dismantle rhythm into its constituent features should be done with caution. Indeed, musical rhythm is an abstract entity; the very understanding

[22] For the interested reader, we mentioned in the introduction investigations specific to tempo perception and memory.
[23] Many works focus on the perception of rhythm by young children, their music representation being somehow more objective.

of the concept of rhythm is highly subjective and cannot be sought as an objective descriptor that would be present in the signal, waiting to be "dug out" by an appropriate technique.

One of the facets of a music content representation is to focus on structures that exist in listeners' minds. Hence, in our opinion, content-based processing is subject to the following dichotomy:

- At the very outset, analyses should be anchored in semantic concepts, and take into account cognitive methodologies (Ellis, 1996). This option entails that the output of analyses would provide directly meaningful representations. Or,
- A mapping has to be achieved between
  - output of analyses schemes that process the signal in *bottom-up manners*, describing *physical contents*, and
  - representations of the signal respecting dimensions that have musical meanings.

Our aim is not to design models that would follow a cognitive approach in their very implementation. Rather, we address the latter option. The features of a description scheme for rhythm will therefore be sought as pertaining to two substantially different representations:

- A low-level representation of signal features: Low-level, as thorough and objective as possible.
- Diverse representations of musical rhythmic concepts, anchored in the *contexts* of the representation of rhythm. That is, *subjective to the application and the users*.

Then the issue will be to provide a relevant mapping between these two types of representations. See (Amatriain *et al.*, this issue).

4 Discussion – two virtual prototypes

Many approaches intend to cover a global understanding of rhythm making use of music theory, psychology, cognitive science and signal processing. We object that, in the context of rhythmic content processing, this may be misleading. One could judge inappropriate seeking models that are able to (1) transcribe the hierarchical organizations of the multiple temporal patterns imposed by the composer, (2) describe the subjective interpretation of *any* listener, and (3) describe the expressive interpretation of the performer. In seeking this tremendous objective as a whole, previous approaches have led to a poor agreement regarding the description and representation of rhythm. Employing such a rationale forgets the reality that there are many possible understandings of the rhythm.

Starting with the rationale that rhythm perception is grounded in cognition and that there is no "absolute" study of rhythm, the descriptive features of rhythm are dependent on the context of the study.[24] Moreover, it seems logical to consider that the process of understanding rhythm calls for a *production* phase (it is commonly agreed that rhythm involves movement), similar to understanding speech and other tasks involving perception and cognition. Therefore, rhythm analysis could be anchored in the development of complementary prototypes –or production/exploration interfaces–. These prototypes would differ by the types of signals to process, the types of listeners to consider, and the description to focus on. Our hope is that the functionalities of such prototypes could demonstrate that there is more to discover in "relative" studies of rhythm than in an "absolute" one.

So, let us indulge in a brief digression, directly inspired by the explanations of the European project CUIDADO[25] of Vinet *et al.* (2002):

The context is that of general audio retrieval and processing. The related issues range from the analysis process (extraction of descriptors), through the navigation process (retrieval methods and interfaces), and up to the creative process (editing/transforming tools). One could imagine addressing these issues through two interactive tools.

The first prototype would be an authoring tool for retrieving, editing, transforming and mixing isolated sound samples and phrases of reduced polyphonic complexity. It would be designed for music professionals (musicians or producers). On the other hand, the second one would handle issues specific to general polyphonic audio recordings, such as comparison and retrieval, it would be designed for music consumers.

The usefulness of rhythm descriptors would therefore depend on the context of the user. The *functionalities* would greatly differ, as would the *features* supposed to convey rhythm.

4.1 Ideas regarding a first prototype

Dealing with the music professional user type, the specific signals to consider are sound samples and phrases of reduced polyphonic complexities. What would this type of user need rhythm descriptors for? Below is a list of functionalities that could be useful:

- Synchronize two monophonic tracks.
- "Smooth" sequencing of tracks (in the sense of rhythmically coherent).

[24] This presents a pleasing parallel with the fact that the definition of 'rhythm' refers to different notions, depending on the temporal scope of what is being characterized.

[25] To which both the authors are participating.

Accepted for publication on May 9th 2002

- Determine time indexes in audio signals that would stand as "looping" points, or references for "cut and paste" operations.
- Navigate among specific instrument occurrences within a track (entails notes segmentation and timbral description at the tone level).
- Classify a rhythmic pattern with respect to categories (either templates as *e.g.* "salsa" or user-defined).
- Navigate by rhythmic patterns categories within a database of patterns, compare them, retrieve satisfying instances.
- Apply transformations on tracks of reduced polyphonic complexities –see also (Amatriain *et al.*, this issue)–:
  At the *track level*:
    - Time-stretching driven by the knowledge of the tempo of another audio source
    - Quantization to a grid: slightly move events to the closest time indexes in a pre-defined grid (this entails the common meaning of quantization, but also the application of slight temporal transformations using template patterns –*e.g.* specific-performer-like quantization–).
    - "Human touch", that is, slight deviations from the perfect time indexes of the theoretical rhythm pattern
    - Tempo-synchronous audio effects.

  At the *events level*:
    - Application of effects on the occurrences of a specific timbre (mute, equalize, apply an effect on, stretch, etc.)
    - Transformation of the occurrences of a specific timbre by means of synthesis models
    - Substitution of a specific timbre by an alternative one present in a database of sounds.
    - Quantization to a grid: slightly move occurrences of a specific timbre to the closest time indexes in a pre-defined grid

What would then be the features to focus on that would presumably convey rhythm?

Some of the following proposed features are objective in that they can be extracted unambiguously from the signal; others refer to more abstract concepts. Some –if not all– are of course related.

- Accurate timings of transients.
- *Local* features of events (either audio frames or events which limits are derived from a previous segmentation step) such as spectral centroid, temporal centroid, energy, etc. Useful descriptors can be extracted from any signal representation (*e.g.* temporal, spectral, wavelets, etc.), or modelings of the signal (*e.g.* LPC, ARMA, etc.).
- *Evolution* features of events (previously segmented) such as evolution of the spectral centroid over its constituent frames, etc.
- Accurate temporal lists (series) of instrument onset occurrences.
- Inter-onset intervals (either regardless of a specific instrument or taking it into account).
- A weighting for each series to indicate its importance as for the communication of rhythm (probably, this would depend on the type of sounds playing, and on the structures of the series).
- Some measure of complexity of the series. Shmulevich *et al.* (2000) clarify this concept: "In general, an object's complexity reflects the amount of information embedded in it, the representation of the object's information is achieved via coding." They argue that the efficiency of the coding is directly related to the complexity of the object itself.
- Different pulses rates (Tempo –as perceived pulse–, Tatum and a reference pulse –*e.g.* quarter-note–).
- Time signature
- Tempo variations.
- Timing deviations of events respecting to the Tatum or other pulse grids.

With these in mind, the features to focus on are onset times, durations, intensities and timbres (under the assumption that timbral regularity affects rhythmic aspects, it seems important to also focus on timbre characterization –see (Herrera *et al.*, this issue)–. The mapping between low-level features and the representation elements that would be available to the users is here supposedly quite simple. Indeed, someone apprehending music in an analytical manner is probably interested in almost any feature representing objectively the physical contents of sound (*i.e.* low-level features –*e.g.* onsets– or objective organizations of low-level features –*e.g.* onsets of a specific instrument–).

The targeted application and user-type justify the underlying transcriptive approach. Indeed, as this prototype would provide very accurate functionalities to music professionals, usually trained in the task of hearing structures in music, the point of view to be taken should be that of accurate symbolic musical description. Although it seems convincing that music sheet-notations should not represent precisely human perception of rhythm, it can be argued that musicians do use a formal approach when listening and playing. The point could also be made that musicians precisely try to hear the systematic timing deviations between the perfect structures and the actual occurrence instants of the events (Baggi,

Accepted for publication on May 9th 2002

1991). In that respect, it could be argued that a formal approach is relevant as a first step towards the understanding of rhythm perception and expressivity in musicians.

(Some of the functionalities and descriptors above are being implemented by the CUIDADO consortium in a prototype called "Sound Palette".)

4.2 Ideas regarding a second prototype

The management of large music databases would concern a second prototype. Here sounds to handle would be whole polyphonic music titles, in audio format and users would be music consumers.

As above, one might first wonder why would the user of this prototype need rhythm descriptors:

- Perform similarity measures: here, thorough editing and transformation functionalities do not seem the major points to focus on anymore. Rather, the notion of rhythmic similarity appears central. An interesting aspect would concern browsing and retrieval functionalities. A second aspect would be to allow the user introduce new instances in the database.
- Playlist Generation: in addition to proposing sets of individual titles to users by means of a similarity measure, this prototype should also allow to build sequences of music titles satisfying particular properties (or constraints). For instance, one could wish a playlist with the following preferences: 'no slow or very slow tempos', 'medium to heavy beats', 'groove similar to that of "Gimme some more" by Busta Rhymes', etc. The song sequence would then be selected by satisfying constraints on rhythm descriptors –among others–. For a description of such an application, see (Pachet et al., 1999) or (Aucouturier et al., 2002).

These functionalities entail the notion of *interactivity*. The prototype should be able to adapt to the users' tastes and wills. Indeed, let us consider a specific database, music titles would be presented to the user with a given organization –*e.g.* visual– making use of predefined rhythmic similarities. If the user wants to add a new title, declaring it similar to –say– songs A and B, then, the prototype must learn what is in that case the implicit meaning of 'rhythmically similar'. This feature is indeed a challenge.

Keeping in mind these functionalities and the interactivity requirement, special focus should be put on the two following points:

- First, aiming at global, or "unary", descriptors, that describe music titles in a global fashion (one value for one song).
- Secondly, focus on defining similarity rules between different pieces, based on distance measures in *adapted* representation spaces (in contrast with declarative definitions of rhythm descriptors). Here, the analyses would be anchored in the non-expert listener's experience of rhythm.[26]

An interesting approach to the identification and validation of relevant rhythm descriptors and similarity metrics in songs is that of Sony CSL in Paris. They make use of the following techniques:

- Web-based games: these are musical games available on the web, such as the "Discrimination game", that asks the user to choose a descriptor that would permit to discriminate two music titles; or the "Recognition game", that consists in finding which music title of a panel is described by a specific descriptor's value. A comparable effort to collect subjective data (here about artist similarities) via web-based games can be found at http://www.musicseer.com.
- Data-mining: Pachet *et al.* (2001) report on a method of classification based on two techniques: co-occurrence and correlation analysis. By studying large corpora of textual information on web pages about music titles or artists, they are able to cluster interesting groups that can reveal useful similarities. These similarities are partially grounded on rhythm, and serve to motivate or validate new rhythm descriptors.

Finally, here are some rhythm descriptors that may be of interest for such a prototype:

- *Overall tempo* (a constant for the whole piece, it would probably not be of interest here to present the slight variations of tempo to the user).
- Global measure of the *swing* or *groove*, computed from timing deviation data. See (Madison, 2001) and (Laroche, 2001).
- Some measure of "rhythmic regularity", making use of tempo fluctuations and timing deviation data. Rhythmic styles or producing habits can be clustered to some extent with the use of such an attribute. It seems natural to consider that, independently of the rhythmic structure, a Techno music production will have much less "approximate timings" than an excerpt produced for –say– a "roots-Cuban" music disc.

[26] Gabrielsson's methodology (1973a and 1973b) should be of special interest here. He identifies subjective dimensions of rhythm from various perceptual tests. Multidimensional scaling permits to reduce the inherent dimensionality of similarity ratings, and verbal descriptions are used to qualify the resulting dimensions.

- *Percussivity*: the detection and classification of percussive timbres in polyphonic music –*e.g.* (Gouyon *et al.*, 2000)– could be a first step towards the determination of a global measure of the "percussivity" of a title. Accounting for brushes or heavy synthetic bass drums certainly has rhythmic importance.
- *"Danceability"*: this is a very important aspect of rhythm perception in everyday music listening.
- *"Drum Track"*: for music titles where percussions convey the rhythm, some techniques –*e.g.* (Goto *et al.* 1995), (Gouyon, 2000)– can provide the position of the onsets for snare and bass drum-like sounds. A more compact representation of these sequences would allow building a visual space (2- or 3-dimensional) where music titles could be projected.

Here, the mapping between low-level features and the representation elements available to the users seems less simple than in the previous case. Under the assumption that non-expert listeners experience musical rhythm in a non-analytic manner, low-level features are no more candidates for elements of the end-representation. Moreover, the prototype should account for an adaptive feature: the end-representation would be adapted to each user to some extent. (Some of the functionalities and descriptors above are being implemented by the CUIDADO consortium in a prototype called "Music Browser".)

5 Conclusion

Musical rhythm understanding is a challenging domain of investigation whose concepts and limits are still somehow fuzzy. Nevertheless, in the rapidly growing framework of content-based processing of music and music information retrieval, there is an obvious need for describing the rhythmic aspects of music from a point of view that can entail a high level of abstraction.

In this article, we reviewed computational modeling approaches to musical rhythm and proposals of rhythm representation schemes. Finally, we depicted a pragmatic framework for further investigation. We argue that analyses of the rhythmic contents of musical signals should be anchored in the development of specific applications that would imply specific types of signals to be processed, specific types of listeners to consider and specific descriptions to focus on. Our hope is that the functionalities of prototypes developed with this rationale in mind could demonstrate that there is more to discover in relative studies of rhythm than in an absolute one.

Acknowledgment
The authors are thankful to Harvey Thornburg, Tamara Smyth, Bob L. Sturm and Stefan Bilbao in CCRMA, Perfecto Herrera, Pedro Cano, Emilia Gomez, Martin Kaltenbrunner and Alvaro Barbosa of the MTG. The authors also wish to thank the members of Sony CSL Music Team in Paris. The work reported in this paper has been partially funded by the IST European project CUIDADO.

References

Aigrain, P. (1999). New applications of content processing of music. In *Journal of New Music Research* 28(4).
Alghoniemy, M. & Tewfik, A. (1999). Rhythm and periodicity detection in polyphonic music. In *Proceedings of the IEEE third Workshop on Multimedia Signal Processing*.
Amatriain, X., Bonada, J., Loscos, À., Arcos, J.L. & Verfaille, V. (2002). Addressing the Content level in audio and music transformations. In *Journal of New Music Research* (this issue).
Aucouturier, J.-J. & Pachet, F. (2002). Scaling up music playlist generation. *Submitted to IEEE International Conference on Multimedia Expo.*
Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison Wesley.
Baggi, L. (1991). Neurswing: An intelligent workbench for the investigation of swing in jazz. In *Computer* 24(7).
Bilmes, J. (1993). *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm*. M. Sc. Thesis MIT, Media Lab, Cambridge.
Blum, T., Keislar, D., Wheaton, A., & Wold, E. (1999). *Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information*. USA Patent nº 5,918,223.
Brooks, R. (1991). New approaches to robotics. In *Science* 253
Brown, J. (1993). Determination of the meter of musical scores by autocorrelation. In *Journal of the Acoustical Society of America* 94(4).
Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. Ph.D. Thesis Faculty of Music and Department of Artificial Intelligence, University of Edinburgh.
Cano, P., Kaltenbrunner, M., Gouyon, F., & Batlle, E. (2002) On the use of FastMap for audio retrieval and browsing. In *Proceedings of the International Symposium on Music Information Retrieval*.
Cemgil A., Kappen B., Desain P., & Honing H. (2000). On tempo tracking: tempogram representation and Kalman filtering. In *Proceedings of the International Computer Music Conference*.

Accepted for publication on May 9th 2002

Chowning, J. & Mont-Reynaud, B. (1986). *Intelligent analysis of composite acoustic signals*. Technical report STAN-M-36, CCRMA, Stanford University.
Chowning, J., Rush, L., Mont-Reynaud, B., Chafe, C., Schloss, A., & Smith, L. (1984). *Intelligent system for the analysis of digitized acoustic signals*. Technical report STAN-M-15, CCRMA, Stanford University.
Clarke, E. (1987). Levels of structure in the organization of musical time. In *Contemporary music review* 2(1).
Clarke, E. (1999). Rhythm and Timing in Music. In *The Psychology of Music, 2nd edition*, edited by Deutsch D., Series in Cognition and Perception. Academic Press.
Cliff, D (2000). *Hang the DJ: Automatic Sequencing and Seamless Mixing of Dance-Music Tracks*. Technical report Hewlett-Packard HPL-2000-104.
CNMAT RRG. *CNMAT Rhythm Research Group*. http://cnmat.cnmat.berkeley.edu/Rhythm/
Cooper, G.W. & Meyer, L.B. (1960). *The rhythmic structure of music*. University of Chicago Press.
Desain, P. (1990). A connectionist and a traditional AI quantizer, symbolic versus sub-symbolic models of rhythm perception. In *Proceedings of the Music and the Cognitive Science Conference*.
Desain, P. (1992). The quantization problem: traditional and connectionist approaches. In *Understanding music with AI: Perspectives on Music Cognition*, AAAI Press.
Desain, P. & de Vos, S. (1990). Autocorrelation and the study of musical expression. In *Proceedings of the International Computer Music Conference.*
Desain, P. & Honing, H. (1991). The quantization of musical time: a connectionist approach. In *Music and Connectionism*, edited by Todd P.M. & Loy D.G., MIT Press, Cambridge.
Desain, P. & Honing, H. (2002). Modeling the Effect of Meter in Rhythmic Categorization: Preliminary Results. In *Journal of Music Perception and Cognition. Sapporo: JSMPC*.
Desain, P., Honing, H., van Thienen H., & Windsor L. (1998). Computational Modeling of Music Cognition: Problem or Solution? In *Music Perception* 16(1).
Desain, P. & Windsor L. (2000). *Rhythm Perception and Production*. Swets & Zeitlinger.
Dixon, S. (1999). A beat tracking system for audio signals. In *Proceedings of the Conference on Mathematical and Computational Methods in Music.*
Dixon, S. (2000). A Lightweight Multi-Agent Musical Beat Tracking System. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*.
Dixon, S. & Cambouropoulos, E. (2000). Beat tracking with musical knowledge. In *Proceedings of the European Conference on Artificial Intelligence.*
Drake, C. (1993). Reproduction of musical rhythms by children, adult musicians and adult non-musicians. In *Perception and Psychophysics* 53(1).
Drake, C., Penel, A., & Bigand, E. (2000). Why musicians tap slower than nonmusicians. In *Rhythm Perception and Production*, edited by Desain P. & Windsor L., Swets & Zeitlinger.
Eck, D., Gasser, M., & Port, R. (2000). Dynamics and embodiment in beat induction. In *Rhythm Perception and Production*, edited by Desain P. & Windsor L., Swets & Zeitlinger.
Ellis, D. (1996). *Prediction-driven computational auditory scene analysis*. Ph.D. Thesis MIT Media Lab, Cambridge.
Foote, J. & Uchihashi, S. (2001). The Beat Spectrum: A New Approach to Rhythm Analysis. In *Proceedings of the International Conference on Multimedia and Expo.*
Foote, J., Cooper, M. & Nam, U. (2002). Audio retrieval by rhythmic similarity. In *Proceedings of the International Symposium on Music Information Retrieval.*
Fraisse, P. (1982). Rhythm and Tempo. In *The Psychology of Music*, edited by Deutsch D., Series in Cognition and Perception. Academic Press.
Friberg, A. (1997). *A Quantitative Rule System for Musical Performance*. http://www.speech.kth.se/music/publications/thesisaf/
Friberg, A. & Sundberg, J. (1995). Time discrimination in a monotonic, isochronous sequence. In *Journal of the Acoustical Society of America* 98(5).
Friberg, A. & Sundberg, J. (1999). Does music performance allude to locomotion? A model of final ritardandi derived from measurements of stopping runners. In *Journal of the Acoustical Society of America* 105(3).
Gabrielsson, A. (1973a). Similarity ratings and dimension analyses of auditory rhythm patterns, Part II. In *Scandanavian Journal of Psychology* 14.
Gabrielsson, A. (1973b). Similarity ratings and dimension analyses of auditory rhythm patterns. Part I. In *Scandanavian Journal of Psychology* 14.
Gabrielsson, A. (1999). The Performance of Music. In *The Psychology of Music, 2nd edition*, edited by Deutsch D., Series in Cognition and Perception.
Gasser, M., Eck, D., & Port, R. (1999). Meter as mechanism: a neural network that learns metrical patterns. In *Connection Science* 1.
Gomez, E., Klapuri, A., & Meudic, B (2002). Melody description and extraction in the context of music content processing. In *Journal of New Music Research* (this issue).

Accepted for publication on May 9th 2002

Goto, M. & Muraoka, Y. (1995). A real-time beat tracking system for audio signals. In *Proceedings of the International Computer Music Conference*.
Goto, M. & Muraoka, Y. (1997). Real-time Rhythm Tracking for Drumless Audio Signals: Chord Change Detection for Musical Decisions. In *Proceedings of the International Joint Conferences on Artificial Intelligence, Workshop on Computational Auditory Scene Analysis*.
Gouyon, F. (2000). *Extraction automatique de descripteurs rythmiques dans des extraits de musiques populaires polyphoniques*. DEA ATIAM Thesis, IRCAM, Paris.
Gouyon, F., Herrera, P., & Cano, P. (2002). Pulse-dependent analyses of percussive music. In *Proceedings of the AES 22 International Conference on Virtual, Synthetic and Entertainment Audio*.
Gouyon, F., Pachet, F., & Delerue, O. (2000). On the use of zero-crossing rate for an application of classification of percussive sounds. In *Proceedings of the Digital Audio Effects conference.*
Herrera, P., Peeters, G. & Dubnov, S. (2002) Automatic Classification of Musical Instrument Sounds. In *Journal of New Music Research* (this issue).
Honing, H. (1993). Issues in the representation of time and structure in music. In *Contemporary music review* 9.
Iyer, V. (1998). *Microstructures of Feel, Macrostructures of Sound: Embodied Cognition in West African and African-American Musics*. Doctoral dissertation, University of California, Berkeley.
Iyer, V., Bilmes, J., Wright, M., & Wessel, D. (1997). A novel Representation for Rhythmic Structure. In *Proceedings of the International Computer Music Conference.*
Kristiansen, D., Husøy, J., & Eftestøl, T. (1995). Rhythm detection in ECG signals. In *Proceedings of the Norwegian Signal Processing Symposium.*
Lapidaki, E. (1996). *Consistency of tempo judgments as a measure of time experience in music listening*. Ph.D. Thesis Northwestern University, Evanston, Illinois.
Lapidaki, E. (2000). Stability of tempo perception in music listening. In *Music Education Research* 2(1).
Large, E. & Kolen, E. (1994). Resonance and the Perception of Musical Meter. In *Connection Science* 6.
Laroche, J. (2001). Estimating tempo, swing and beat locations in audio recordings. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.
Lee, C., Todd, P.M., Foster, A., & Lomlu, S. (2000). Preliminary investigations of French and English speech rhythm: are cross-linguistic differences in rhythmic organisation primarily metrical in origin? In *Rhythm Perception and Production*, edited by Desain P. & Windsor L., Swets & Zeitlinger.
Lemström, K., Wiggins, G., & Meredith, D. (2001). A three-layer approach for music retrieval in large databases. In *Proceedings of the International Symposium on Music Information Retrieval*.
Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press, Cambridge.
Levitin, D.J. & Cook, P. (1996). Absolute memory for musical tempo: Additional evidence that auditory memory is absolute. In *Perception and Psychophysics* 58.
Lindsay, A. & Kriechbaum, W. (1999). There's more than one way to hear it: Multiples representations of music in MPEG-7. In *Journal of New Music Research* 28(4).
London, J. (2001). Rhythm. In *The new Grove Dictionary of Music and Musicians, 2nd edition*, edited by Macmillan, London.
Longuet-Higgins, C. (1987). *Mental processes*. MIT Press, Cambridge.
Longuet-Higgins, C. & Lee, C. (1982). Perception of musical rhythms. In *Perception* 11.
Madison, G. (2001). Different kinds of groove in jazz and dance music as indicated by listeners' ratings. In *Proceedings of the Seventh International Symposium on Systematic and Comparative Musicology and Third International conference on cognitive musicology*.
Mazzoni, D. & Dannenberg, R. (2001). Melody matching directly from audio. In *Proceedings of the International Symposium on Music Information Retrieval.*
McAuley, J.D. (1995). *Perception of time as phase: Towards an adaptive-oscillator model of rhythmic pattern processing*. Ph.D. Thesis Indiana University, Bloomington.
Melucci,M., Orio, N., & Gambalunga, M. (2000). An evaluation study on music perception for music content-based information retrieval. In *Proceedings of the International Computer Music Conference.*
MPEG7 audio team (2001). *ISO working draft - Information Technology - Multimedia Content Description Interface - Part4: Audio*. Tech. report ISO/IEC CD 159384 version 1.0.
Ó Maidín, D. & Fernström, M. (2000). The best of two worlds: retrieving and browsing. In *proceedings of the Digital Audio Effects conference.*
Pachet, F., Delerue, O., & Gouyon, F. (2000). *Automatic extraction of rhythmic structure from music*. Technical report Sony Research Forum.
Pachet, F., Roy, P., & Cazaly, D. (1999). A Combinatorial Approach to Content-based Music Selection. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*.
Pachet, F., Westermann, G., & Laigre, D. (2001). Musical Data Mining for Electronic Music Distribution. In Proceedings of the *International Conference on Web Delivering of Music*.

Accepted for publication on May 9th 2002

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. In *Music Perception* 11.
Perron, M. (1994). Checking Tempo stability of MIDI sequencers. In *Proceedings of the 97th Convention of the AES.*
Raphael, C. (1999). Automatic segmentation of acoustic musical signals using Hidden Markov Models. In *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence.*
Raphael, C. (2001). Automated rhythm transcription. In *Proceedings of the International Symposium on Music Information Retrieval.*
Rosenthal, D. (1992). Emulation of human rhythm perception. In *Computer Music Journal* 16(1).
Rowe, R (2001). *Machine musicianship*. MIT Press, Cambridge.
Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. In *Journal of the Acoustical Society of America* 103(1).
Scheirer, E. (2000). *Music-Listening Systems*. Ph.D. Thesis MIT Media Lab, Cambridge.
Scheirer, E. & Slaney, M. (1997). Construction and evaluation of a robust multifeature Speech/Music discriminator. In *Proceedings of the IEEE ICASSP.*
Schloss, A. (1985). *On the automatic transcription of percussive music - From acoustic signal to high-level analysis*. Ph.D. Thesis CCRMA, Stanford University.
Shmulevich, I. & Povel, D. (2000). Measures of Temporal Pattern Complexity. In *Journal of New Music Research* 19(1).
Slaney, M. (1995). A critique of pure audition. In *Proceedings of the International Joint Conference on Artificial Intelligence, Workshop on Computational Auditory Scene Analysis*.
Smith, L. (1996). Modelling rhythm perception by continuous time-frequency analysis. In *Proceedings of the International Computer Music Conference.*
Smith, L. & Kovesi, P. (1996). A continuous time-frequency approach to representing rhythmic strata. In *Proceedings of the International Conference on Music Perception and Cognition.*
Smith, L. & Medina, R. (2001). Discovering themes by exact pattern-matching. In *Proceedings of the International Symposium on Music Information Retrieval*.
Snyder, J. & Krumhansl, C. (2001). Tapping to Ragtime: Cues to pulse finding. In *Music Perception* 18
Steels, L. (1999). *The Talking Heads Experiment - Volume I. Words and Meanings*. pre-edition for Laboratorium, Antwerpen.
Tanguiane, A. (1994). A principle of correlativity of perception and its applications to music recognition. In *Music Perception* 11.
Thiemel, M. (2001). Accent. In *The new Grove Dictionary of Music and Musicians, 2nd edition*, edited by Macmillan, London.
Thornburg, H. (2001a). *Bayesian segmentation and rhythm tracking*. Draft report, CCRMA, Stanford University.
Thornburg, H. (2001b). *The Bayesian Approach to Segmentation and Rhythm Tracking*. http://www-ccrma.stanford.edu/~jos/mus423h/Bayesian_Approach_Segmentat.html
Turgeon, M. (2000). *Cross-spectral grouping using the paradigm of rhythmic masking release*. Ph.D. Thesis McGill University.
Tzanetakis, G., Essl, G., & Cook, P. (2001). Automatic Musical Genre Classification of Audio Signals. In *Proceedings of the International Symposium for Audio Information Retrieval.*
Vinet, H., Herrera, P., & Pachet, F. (2002). The CUIDADO Project. In *Proceedings of the International Symposium on Music Information Retrieval*.
Wang, Y. (2001). A Beat-Pattern based Error Concealment Scheme for Music Delivery with Burst Packet Loss. In *Proceedings of the International Conference on Multimedia and Expo.*
Wittlich, G. (2001). *Rhythm and Meter, a bibliographic glossary*. http://www.music.indiana.edu/som/courses/rhythm/
Wright, M. & Wessel, D. (1998). An improvisation environment for generating rhythmic structures based on north indian "Tal" patterns. In *Proceedings of the International Computer Music Conference*.
Yeston, M. (1976). *The stratification of musical rhythm*. Yale University Press, New Haven.

Accepted for publication on May 9th 2002