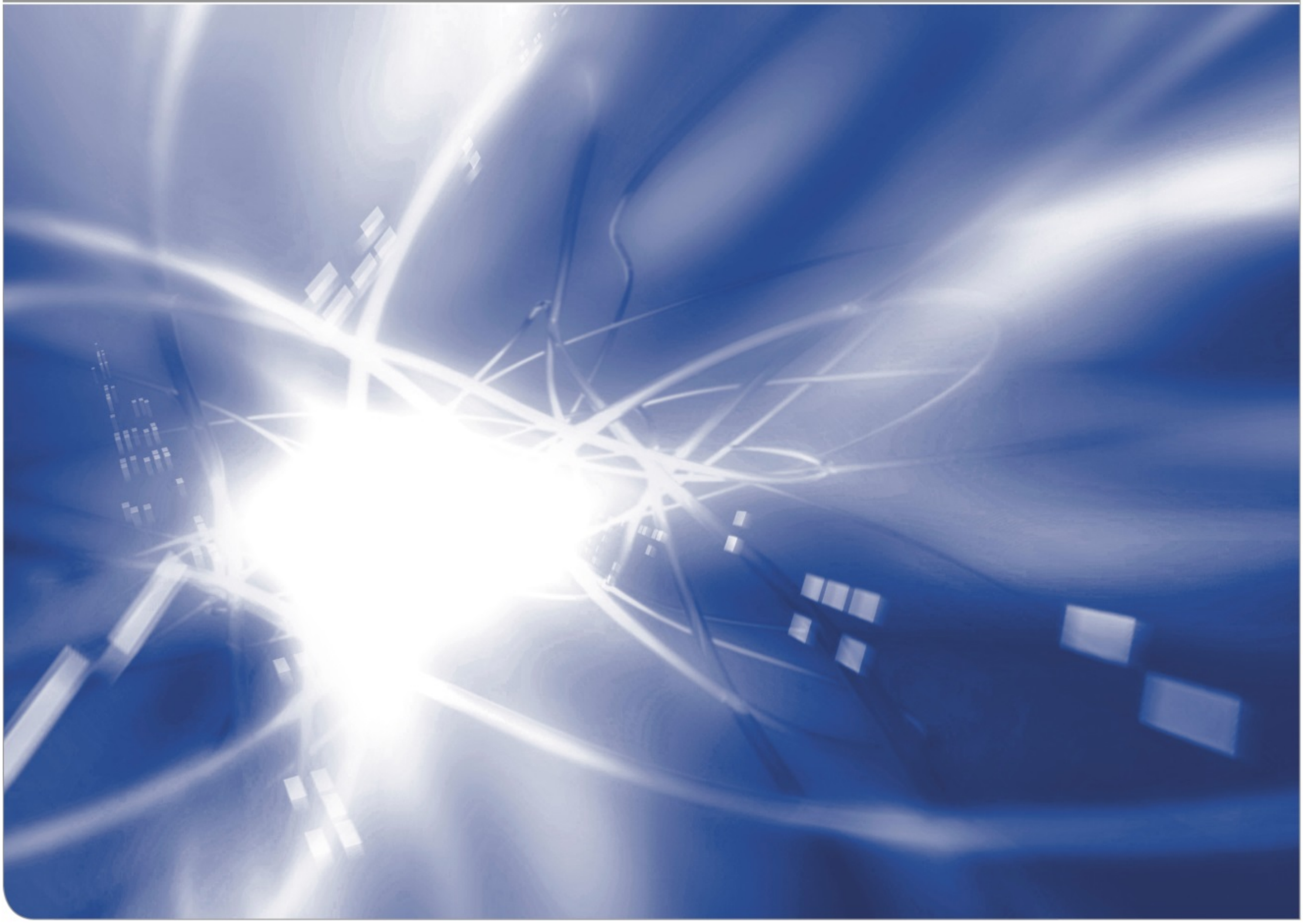
Karlsruhe Institute of Technology

Breaking the vicious circle of rhythm–tempo definitions

by Andranik S. Tangian[1]

KIT SCIENTIFIC WORKING PAPERS 168



[1] Institute of Economic Theory and Operations Research

KIT-ECON
Blücherstraße 17
76185 Karlsruhe
Germany
Tel: +49 721 6084 3077
micro.office@econ.kit.edu
andranik.tangian@kit.edu; andranik.tangian@gmail.com

Impressum

Karlsruher Institut für Technologie (KIT)
www.kit.edu

2021

ISSN: 2194-1629

Institute of Economic Theory and Operations Research
Karlsruhe Institute of Technology

Breaking the vicious circle of rhythm–tempo definitions[1]

Andranik S. Tangian

Scientific paper Nr. 168

June 2021

E-mail: andranik.tangian@kit.edu
andranik.tangian@gmail.com
Tel: +49 721 6084 3077

Blücherstraße 17 76185 Karlsruhe Deutschland

[1] This paper is the second and improved edition of my 1998 article 'An operational definition of rhythm and tempo', *General Psychology*, special issue *Temporal Dynamics and Cognitive Processes* ed. by M. Olivetti-Belardinelli and R. Di Matteo, pp. 55–89, which is not available on the internet.

Abstract

In music literature, rhythm is defined relative to a certain tempo, and tempo is defined relative to a certain rhythm. This vicious circle implies that any sequence of time durations can be regarded as either (a) a sequence of these durations at a constant tempo or (b) a sequence of equal durations at a varying tempo or (c) a sequence of unequal durations at a varying tempo in numerous ways. Most listeners, however, perceive rhythm and tempo in the same way, which we explain as the result of a close interaction of the grouping and simplicity laws of Gestalt psychology.

Operationally, the complexity of a data representation is defined as the amount of memory that is required for the algorithm of the data generation. Each rhythm-tempo representation includes rhythmic patterns and the tempo curve that 'generates' their augmentations and diminutions in time. The complexity of such a representation is split between the rhythmic patterns and the tempo curve, and the representation with the least total complexity is selected. Rhythm and tempo are thus complementary structures that mutually adapt according to the criterion of simplicity, which leads to an optimal rhythm-tempo perception.

In addition to general provisions, we consider a few rules for grouping time events into patterns, a directed search for optimal representations of time events, and the influence of the musical context on the perception of rhythm and tempo.

Keywords: Rhythm; Tempo; Rhythmic grammar; Gestalt psychology; Principle of correlativity of perception; Artificial perception.

Contents

1 Introduction

The nature of rhythm, tempo and meter (time) is one of the first questions posed by music theory. Many music theorists have contributed to understanding the related perception mechanisms. For instance, rhythm is said to be the order and the proportion of durations [Porte 1977]; tempo is explained as a characteristic of executive movement in time with respect to measures and melodic, harmonic, rhythmic, or dynamical cues [Pistone 1977]; meter is considered a form of determining the rational proportions of rhythm [Viret 1977].

The formulations cited are not only vague, but also logically interdependent. On the one hand, tempo is defined relative to a certain rhythm (if there are no events, no movement can be perceived). On the other hand, in order to measure the proportion of durations, rhythm is defined relative to a certain tempo, which leads to a vicious circle. However, as [Dumesnil 1979] summarized, one can hardly find better formulations. Even special psychomusicological publications such as [Fraisse 1982, Povel and Essens 1985, Clarke and Krumhansl 1990] do not offer unambiguous definitions.

In recent decades, this problem has been considered in the context of computerized rhythm recognition and tempo tracking. A 'strategic' approach is proposed by [Longuet-Higgins 1976, 1987] and [Longuet-Higgins and Lee 1982, 1984]. Rhythm recognition is understood as finding a strategy of 'listening' to music: a hypothesis on the rhythmic structure is formed according to the first events, then it adapts to the incoming data and, finally, a hierarchical rhythmic structure is developed. This approach was continued by [Povel and Essens 1985] and [Desain 1992]. The former study deals with simulating an internal clock that is activated by time patterns. The latter is based on expectations that are continuously tested.

A noteworthy approach to modeling rhythm perception is due to [Bamberger 1980] and [Rosenthal 1988, 1989, 1992]. The rhythmic structure of a melody is divided into simple patterns, which are combined into repeating segments to give the entire structure a certain symmetry. Thus, rhythm is understood as a means of organizing data.

[Desain and Honing 1989] and [Desain et al. 1989] use neuron networks. Each time interval between tone onsets is put into correspondence with a neuron whose activation level is proportional to the given time interval. Through the mutual transmission of the activation, the neurons filter the rhythm and represent it without minor inaccuracies. The result is a stable state of the network with simple activation ratios of neighboring neurons, which is interpreted as the rhythm filtered. The new aspect is the understanding of rhythm as a simplified conceptual description of a sequence of time relations. A substantially similar approach, but without reference to neural networks, is discussed by [Clarke 1987].

In the computer studies cited, the periodic arrangement of time events is aimed at finding a steady beat. This approach gives good results when analyzing music with a more or less constant pulse train, but it fails in tracking *rubato* performances with significant tempo variations. For example, Skrjabin's performance of his *Poem* Op. 32 No. 1 transcribed from a piano roll [Skrjabin 1960] demonstrates tempo variations from ♩. = 19 to ♩. =110 (5.5 times faster). Listeners perceive this adequately but the programs that smooth out timing deviations cannot correctly interpret such a radical tempo change.

The main difficulty with rhythm processing is the ambiguity in representing time intervals as either duration changes or tempo changes or both. For example, the time events in Figure 1a are notated with the same nominal duration while accelerating the

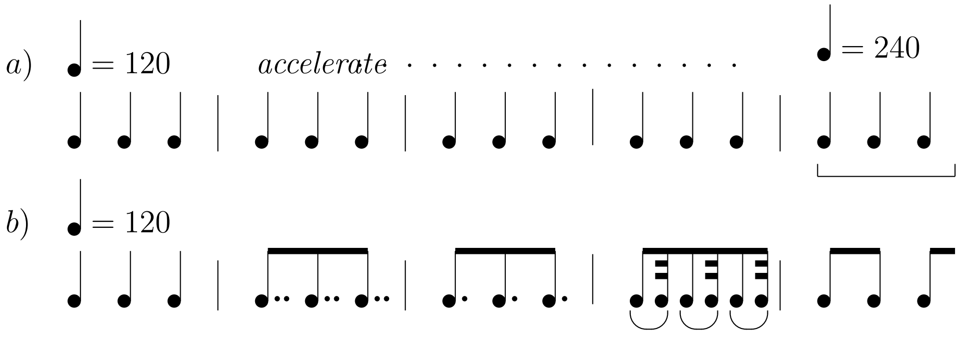


Figure 1: Interpretation of time events as either tempo changes or duration changes

tempo. In Figure 1b, these events are notated with decreasing nominal durations at a constant tempo (this type of transcription is typical for computer notators).

In the given paper, this ambiguity is overcome by applying the principle of correlativity of perception [Tangian 1993, 1994], which is a close interaction of the principles of grouping and simplicity of Gestalt psychology [Wertheimer 1923]. Time events are represented using repetitive rhythmic patterns and a tempo curve in the least complex way in the sense of [Kolmogorov 1965], that is, requiring the least memory storage. The complexity of such a representation is split between the rhythmic patterns and the tempo curve, and the representation with the least total complexity is selected. Rhythm and tempo are thus complementary structures that mutually adapt according to the criterion of simplicity, which leads to an optimal rhythm-tempo perception.

From this viewpoint, the representation in Figure 1a is preferred as less complex. Indeed, the rhythmic patterns in Figure 1a are trivial quarter notes, and the tempo curve is quite simple. In Figure 1b, the constant tempo is trivial but the rhythmic patterns are varied and quite complex.

In Section 2, 'Principle of correlativity of perception', the principles of grouping and simplicity of Gestalt psychology are considered in a close interaction, which is illustrated using examples of visual and audio perception.

In Section 3, 'Rhythm and correlative perception', the principle of correlativity of perception is applied to rhythm. We suggest a hierarchical model of embedded rhythms, which enhances the rhythmic redundancy.

In Section 4, 'Recognizing periodicity', a directional search for quasi-periodicity in a sequence of time events is described.

In Section 5, 'Accentuation', we define strongly and weakly accentuated events exclusively with the help of timing cues.

In Section 6, 'Rhythmic segmentation', the concept of a rhythmic syllable is introduced. It is a sequence of time events that has its own unique accent in the end. A simple psychoacoustic experiment shows that rhythmic syllables are perceived as indivisible rhythmic units.

In Section 7, 'Operations on rhythmic patterns', a kind of rhythmic grammar is developed. The elaboration of a rhythmic pattern is defined as dividing its durations while maintaining the pulse train. The union of rhythmic syllables is the elaboration of their concatenation, which in turn is a rhythmic syllable. In this way, possible rhythmic transformations, as opposed to tempo changes, are described and specific properties of the rhythmic organization are explained.

In Section 8, 'Definition of meter and rhythm complexity', we determine the musical time (meter) and the rhythm complexity, using the root patterns of rhythmic syllables — their simplest prototypes from the point of view of elaboration.

In Section 9, 'Example of analysis', the snare drum part from Ravel's *Bolero* is represented using elaborations of a few rhythmic syllables with a common root, which makes it possible to structure the rhythm and determine its time.

In Section 10, "Summary", the most important provisions of the paper are recapitulated and put into context.

2 Principle of correlativity of perception

By *correlativity of perception* we mean the ability to discover similar stimulus configurations, i.e. their structurally ordered groups, and to build up configurations of a higher level from them. The configurations of stimuli themselves are called *low-level patterns*, and the configurations of the relationships between the low-level patterns are called *high-level patterns*. This hierarchical schema of data representation is provided with feedback that guides the process of data representation in the least complex way. The *complexity* is understood in the sense of Kolmogorov, that is, as the amount of memory storage required for the algorithm of the data generation [Kolmogorov 1965, Calude 1988].

For example, Figure 2a displays pixels (stimuli), which constitute symbols A (low-level patterns) that in turn compose the contour of B (high-level pattern). Instead of storing all the pixels, it is more efficient to save their configuration for one symbol A and then save the contour of B. An important property of such representations is the recognizability of high-level patterns regardless of the recognizability of their carriers: replacing A with an unknown symbol (Ш in Figures 2b–c) has no influence on the recognizability of B.

We focus on identifying similarities for two reasons. First, it allows similar objects to be separated, and the separated objects can be more easily recognized than in data flows. Second, the relationships between similar objects reveal the data structure. Of

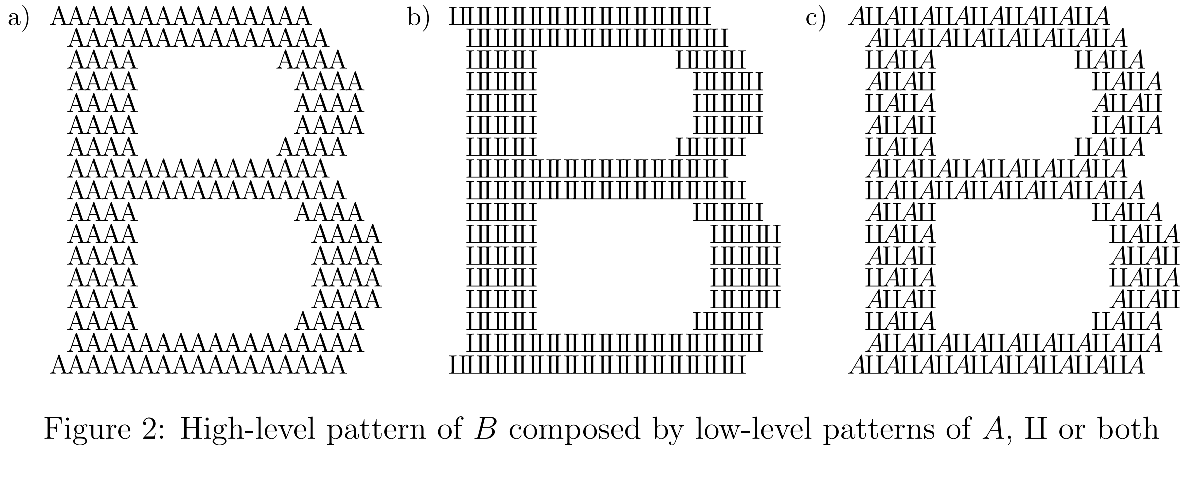


Figure 2: High-level pattern of B composed by low-level patterns of A, Ш or both

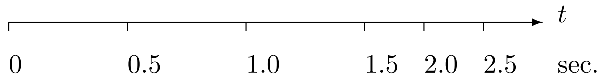
course, similarity is not reduced to identity; there may be similarity in size — as with different symbols in Figure 2c, color, etc. Thereby, the data structure is recognized without knowing the structure, in line with the remark by [Witkin and Tenenbaum 1983]:

> People's ability to perceive structure in images exists apart from the perception of tri-dimensionality and from the recognition of familiar objects. That is, we organize the data even when we have no idea what it is we are organizing. What is remarkable is the degree to which such naively perceived structure survives more or less intact once a semantic context is established: the naive observer often sees essentially the same things an expert does, the difference between naive and informed perception amounting to little more than labeling the perception primitives. It is almost as if the visual system has some basis for guessing *what* is important without knowing *why* . . .
>
> . . . The aim of perceptual organization is the discovery and description of spatio-temporal coherence and regularity. Because regular structural relations are extremely unlikely to arise by the chance configuration of independent elements, such structure, when observed, almost certainly denotes some underlying unified cause or process. A description that decomposes the image into constituents that capture regularity or coherence therefore provides descriptive chunks that act as "semantic precursors," in the sense that they deserve or demand explanations.
>
> — Witkin A.P. and Tenenbaum J.M. (1983) On the role of structure in vision, pp. 482–483

Thus, we are looking for data representations in the form of repetitive (correlating) messages or generative elements and their transformations. (In Figure 2, the transformations are displacements of low-level patterns that form the contour of B.) However, whether similarities are used or not used to represent data may depend on additional arguments. In our model, this is the overall complexity of data representation, which includes the complexity of low-level patterns and the complexity of high-level patterns (= complexity of transformations of low-level patterns). The following example illustrates the contextual dependence of the least complexity criterion that 'decides' whether similarities should be used to represent data.

Example 1 (Contextual dependence of rhythm-tempo perception) *Figure 3 displays a sequence of six time events. Most listeners perceive it as a single rhythmic pattern (Representation A) rather than as the first three events repeated twice faster (Representation B). Here,* **R012** *denotes calling the repetition algorithm* **R** *with three parameters: return to time* **0***, play* **1** *time, play* **2** *times faster). However, if the events have pitch, as in Representation C, then the repetitive melodic contour enhances the sensation of a rhythmic repetition (cf. with the recognizability of a fugue theme in diminution). This means that Representation D is preferred over Representation C.*

To explain the perception of the same rhythm as non-repetitive or repetitive, we estimate the complexity of these representations. We assume that each duration is encoded in one byte, a duration with pitch — in two bytes, and calling the repetition algorithm with parameters — in four bytes. For the rhythm alone, Representation A is less complex than



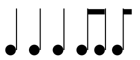
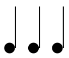
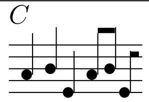
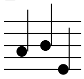
Representation of time events	*A*	*B*	*C*	*D*
		 **R012**		 **R012**
Complexity of rhythmic pattern	6	3	12	6
Complexity of its transformation	0	4	0	4
Total complexity	6	7	12	10

Figure 3: Complexity of representations of six time events, in number of bytes

B (6 bytes versus 7), resulting in the perception of a single rhythmic pattern at a constant tempo. In the melodic context, Representation D is less complex than C (10 bytes versus 12), which leads to the perception of repetition at a double tempo.

In the terminology of [Bregman 1990], we describe *audio scenes* before recognizing the percepts' meaning. No learning or special constructs like conceptual frames [Minsky 1975] or meaningful settings [Palmer 1975] are required. Perceptual structures arise 'by themselves' according to the general laws of similarity and simplicity. This pseudo-semantic self-organization of data is the most important feature of the correlative perception. Of course, semantic cues for structuring data cannot be neglected, but they are not needed at an early stage and can be used later.

To find similarities, it is necessary to perform correlation analysis of given data under their various transformations. Directional search for correlated data blocks in the data transformed can be implemented using the *method of variable resolution*, which we explain in the following example.

Example 2 (Method of variable resolution [Tangian 1993]) *Figures 4a–b display two similar 2D configurations of 1s against a background of 0s indicated by dots. To establish their similarity, we find a simple transformation that makes them equal. For this purpose, configurations a) and b) are superimposed and the number of matching 1s is counted. Since the configurations are unequal, the matching 1s are too few (= no correlation). Then the 'resolution' of both images is reduced by replacing 1s with clusters of 1s as shown in Figures 4c–d. Now the framed areas in Figures 4c–d coincide (= correlate). Once the correlation is captured, the original resolution is gradually restored while keeping control over the areas correlated: if they lose some of the matching 1s (= correlation decreases) then these 1s are shifted to retain the correlation high. These local adjustments determine the required image transformation. In the given case, the transformation is simple, which confirms the similarity of configurations a) and b). If necessary, reducing and restoring the resolution can be done gradually in several steps.*

Thus, the search for the required *global transformations* of data is reduced to *local adjustments*. All of these resemble the operation of perceptrons [Minsky and Papert 1988] and pyramidal data structures [Hummel 1987]. Resolution reduction (= filtering) is common when identifying similarities [Palmer 1983, Witkin 1983, Bouman and Liu 1991], but

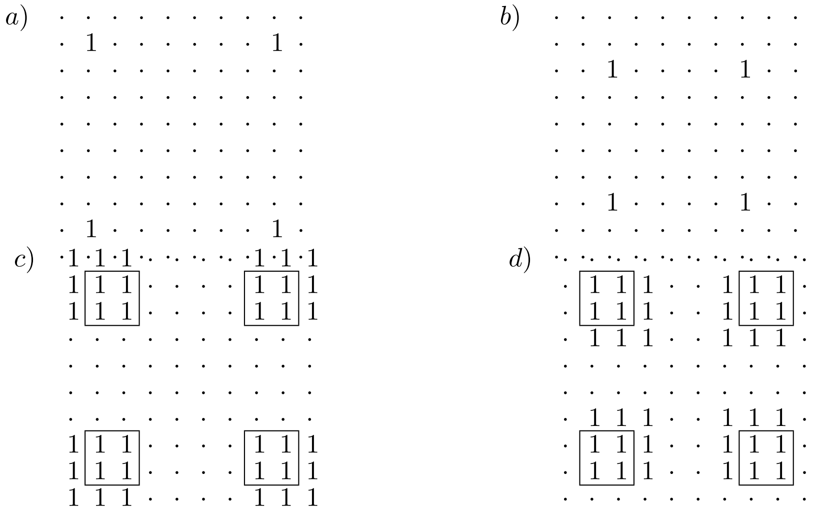


Figure 4: Illustration to the method of variable resolution

we apply it as an intermediate step to find correlated elements, which are then slightly modified to create simple data representations. The method of variable resolution differs in that both filter schemes and correlation analysis are equipped with double feedback that guides the directional search for the 'suitable' image transformations. On the one hand, the data transformations of interest must lead to high data correlations. On the other hand, the least complexity criterion rejects the transformations that are too complex. Obviously, both correlation analysis and the method of variable resolution can be implemented in neuron networks with parallel processing.

3 Rhythm and correlative perception

Now we apply the principle of correlativity of perception to rhythm and tempo. In this case, the low-level patterns are better or worse correlated repetitive *rhythmic patterns* whose time relationships are associated with the high-level pattern of *tempo curve.* In other words, repetitive rhythmic patterns are considered reference units for tempo tracking. Drawing analogy to vision, similar rhythmic patterns correspond to instantaneous states of an object (like in cinema frames), and the tempo curve corresponds to the object's trajectory ([Desain and Honing 1991] do not assume a continuous tempo curve). Table 1 shows the analogy between rhythmic patterns and visual patterns in Figure 2.

Table 1: Analogy between visual and time data

	Visual data	Time data
Stimuli	Pixels	Time events
Low-level patterns	Symbols A, or II, or both	Rhythmic patterns
High-level pattern	Symbol B	Tempo curve

As mentioned in Section 2, high-level patterns are recognizable without explicitly identifying their carriers — low-level patterns. This means that tempo is perceived without associating the rhythmic patterns with waltz, march, etc. We suppose that in most cases there exists a 'compromise' representation with a few fairly simple generative rhythmic patterns (recognizable but not necessarily identifiable) and a fairly simple tempo curve. Their choice is interdependent: the rhythm recognition depends on tempo recognition and vice versa. As complementary structures, rhythm and tempo can be recognized not separately but only together, using the 'external' optimality criterion of the least complex representation of time data with respect to the given musical context; see Example 1 illustrated in Figure 3.

Rhythmic pattern transformations are not limited to tempo variations. Some transformations are what [Mont-Reynaud and Goldstein 1985] call *elaborations* — inserting additional time events in the given rhythm (= dividing pattern durations into shorter ones). Let us explain the role of elaboration in some detail. In Western music, the regularity of time organization is observed at several levels: measures, couples of measures, etc., up to the level of musical form [Lerdahl and Jackendoff 1983]. Composed of patterns of lower levels, patterns of higher levels break down into smaller segments, i.e. are divisible. Accordingly, the perception of rhythm is multilevel, and each level is characterized by its own rhythmic patterns, determined by the relationships between the patterns of the lower levels. Such a multilevel rhythmic structure can be pictured as a tree with indecomposable rhythmic segments at the first level and branches that unite (group) them into patterns of higher levels. This is illustrated by the following example.

Example 3 (Multilevel rhythmic structure) *Figure 5 shows the snare drum part from Ravel's Bolero. The regularity of the rhythm of the first level is determined by the root eighth durations r and their elaborations $E(r)$, which are braced in the first subscript row. The pulse of the second level is determined by the root pattern R, consisting of two eighths, and its elaborations $E_1(R)$ and $E_2(R)$, which are braced in the second subscript row. The third-level pulse is caused by the combination of the second-level patterns into the pattern S and its elaboration $E(S)$ that together produce the fourth-level pulse of the full rhythm T.*

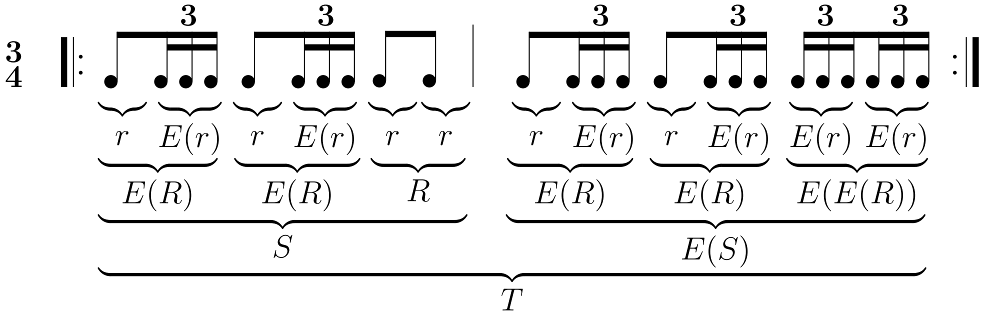


Figure 5: Multilevel rhythmic structure of the snare drum part from M.Ravel's *Bolero*

Embedded pulse trains are inherent in so-called *divisible rhythms*, which are characteristic of Western music. Due to their structural redundancy, tempo variations are not perceived as rhythmic changes. The role of structural redundancy for recognizing repetitions with tempo variations is explained in Example 1. There, the redundancy is created by additional melodic rather than rhythmic cues, but the nature of the effect is the same. The perceptibility of significant tempo variations with a factor of 5.5 in Skrjabin's performance [Skrjabin 1960] is also due to the contextual redundancy created by embedded rhythms, melodic intonation, similarity of the accompaniment, harmonic pulsation, etc.

In some cultures, rhythms do not have multiple embedded levels, which limits tempo variations. Tempo changes are strictly forbidden in Bulgarian or Turkish music with so-called *additive rhythms*, which are characterized by complex duration ratios. From the viewpoint of the correlativity of perception, if a rhythm is not structurally redundant then tempo variations are perceived as rhythmic changes. In order not to violate the musical meaning, the tempo is kept strictly constant.

Thus, the recognizability of tempo variations depends on the rhythm redundancy, which, among other things, presupposes rhythmic elaboration. Therefore, when considering transformations of rhythmic patterns for rhythm-tempo recognition, aspects of both tempo and elaboration should be taken into account.

4 Recognizing periodicity

By *periodicity* we mean the recurrence of time events or their groups. The structure of repeating segments of time events is called *rhythm*. For instance, tempo changes can break the sensation of periodicity but not of rhythm, especially if the musical (rhythmic) structure is sufficiently redundant (cf. with Example 1).

Musical practice deals however not so much with periodicity as with quasi-periodicity caused by minor inaccuracies, tempo fluctuations, etc. To identify quasi-periodicity, the method of variable resolution can be used.

Example 4 (Recognition of quasi-periodicity) *Let the quasi-periodic sequence of time events in Figure 6 be digitalized with an accuracy of 0.1 sec as follows*

$$s(n) = \underbrace{1000000000}_{10}\underbrace{100000000}_{9}\underbrace{10000000000}_{11}10 \tag{1}$$

whose autocorrelation function

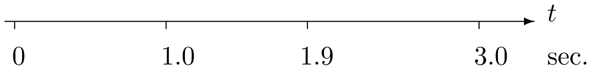
$$R_s(p) = \sum_n s(n-p)s(n)$$

returns the number of matching 1s for the string s *self-superimposed with a shift (period) of p positions. The application of the method of variable resolution is traced step by step in Figure 6:*

- *Since* s *does not have a strict period,* $R_s(p)$ *shows no significant autocorrelation.*

- *Reducing the resolution of* s *by duplicating 1s, we obtain the string* s_1*, whose autocorrelation* $R_{s_1}(p)$ *peaks at* $p = 10$ *with the matching 1s shown in frames.*

- *Restoring the initial resolution, we obtain the original string* $s_2 = s$*, where the position of the lost correlation is denoted by* 0^**. To restore high autocorrelation, we transform* s_2 *into* s_3 *by moving* $\overrightarrow{1}$ *to the position of* 0^**.*



$$
\begin{aligned}
s &= \underbrace{1\ 000000000}_{10}\underbrace{1\ 00000000}_{9}\underbrace{1\ 0\ 000000000}_{11}\ 1\ 0\\
s_1 &= \underbrace{\boxed{1}100000000}_{10}\underbrace{\boxed{1}10000000}_{9}1\underbrace{\boxed{1}000000000}_{11}\boxed{1}1\\
s_2 &= \underbrace{1000000000}_{10}\ \underbrace{100000000}_{9}\ \underbrace{\vec{1}\ 0^*\ 000000000}_{11}1\ 0\\
s_3 &= \underbrace{\boxed{1}000000000}_{10}\underbrace{\boxed{1}00000000\ 0}_{10}\underbrace{\boxed{1}000000000}_{10}\boxed{1}0
\end{aligned}
$$

p	$R_s(p)$	$R_{s_1}(p)$	$R_{s_2}(p)$	$R_{s_3}(p)$
...	0	0	0	0
8	0	1	0	0
9	1	3	1	0
10	1	4	1	3
11	1	3	1	0
12	0	1	0	0
...	0	0	0	0

Figure 6: Autocorrelation $R_s(p)$ of time events

- *The peak of $R_{s_3}(p)$ at $p = 10$ suggests that s is either quasi-periodic or periodic under the following variable tempo:*

$$
T(n) = \begin{cases} 10/10, & 1 \le n \le 10 \quad \textit{(initial tempo)} \\ 10/9, & 11 \le n \le 19 \quad \textit{(acceleration)} \\ 10/11, & 20 \le n \le 30 \quad \textit{(deceleration)} \end{cases} .
$$

If the tempo curve of T looks too complicated, then s should be considered as a rhythmic pattern at a constant tempo, otherwise s is considered a regular pulse train at a variable tempo, which in the notation of Example 1 looks as follows:

$$
s = \{1000000000\}\ \mathbf{R\ 0\ 2\ 10/9\ 10/11}\ ,
$$

where $\{1000000000\}$ *is the generative unit and* **R 0 2 10/9 10/11** *means* **R***epeat from moment* **0**, **2** *times, at tempi* **10/9, 10/11**, *respectively. Of course, the final decision about the least complex representation depends on coding conventions and complexity measures.*

The tempo curve can be encoded by capturing the points in time when the tempo deviates from its current value by, say, more than 5% (= 1/20 of the reference duration). This heuristic corresponds to logarithmic scaling and zonal nature of perception. It's also pretty simple, which is practical in computer experiments.

The described approach has been tested in computer experiments, where the *Bolero* rhythm (Figure 5) has been performed on a computer keyboard. The experiments have revealed the periodicity at the four rhythmic levels shown in Figure 5, as well as the

0-level periodicity of the sixteenth triplets even if they were imprecisely performed with the duration ratio of 10 : 8 : 9. Such a great inaccuracy hasn't been observed for longer durations. It looks that very short sixteenth triplets (of about 0.13 sec. at the *Bolero* tempo ♩ = 76) are less important for perceiving periodicity than longer durations like eighths and quarters. This suggestion is consistent with experimental evidence that tempo fluctuations are best noticeable if the reference duration is in the range of 0.2–1.0 seconds [Michon 1964]. These durations are considered fundamental for rhythm perception, and we will give them a certain priority.

Thus, the model of correlative perception reveals the quasi-periodicity of imprecise performance. However, recognition of periodicity is not yet rhythm recognition. To recognize rhythm, time events must be 'conceptually' structured, that is, segmented, classified, and represented symbolically using standard notation. These tasks are discussed in the following sections.

5 Accentuation

As mentioned in the previous section, a periodic sequence of time events can be considered a rhythm if it is segmented, that is, the segment ends are indicated. This task is not as simple as it seems. In the strictly periodic African percussion music in Figure 7, almost every event can be selected as the beginning of a period [Schloss 1985]. In this case, we say that we do not recognize the rhythm, but only the periodicity.

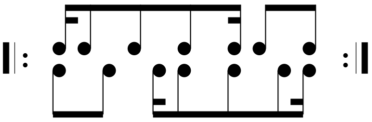


Figure 7: Ambiguous segmentation of a periodic sequence of time events

To find the ends of rhythmic segments, if any, we distinguish between accentuated and non-accentuated events. By analogy with speech, where emphasis is made by lengthening the vowels (and not by increasing the volume!), longer durations are considered more accentuated than shorter ones. The following rules embody the ideas of [Boroda 1985, 1988, 1991].

Rule 1 (Durations) *A time event is characterized by the* duration *of the time interval between the beginning of the event (e.g. tone onset) and the beginning of the next event. The duration of the last event in the sequence is not fixed and can be assumed arbitrarily long.*

According to Rule 1, we only focus on timing cues. The duration of a sound event (e.g. tone) includes both the duration of the sound and the subsequent pause.

Rule 2 (Accentuation Premise) *Accentuation with respect to timing requires at least two types of durations.*

According Rule 2, a sequence of equal durations has no accents. By Rule 1, only the last event can be accentuated.

Rule 3 (Strong and Weak Accents) *A duration d_i is* strongly accentuated *if it is not shorter than the preceding duration d_{i-1} and longer than the next duration d_{i+1}:*

$$d_{i-1} \leq d_i > d_{i+1} \ .$$

A duration d_i is weakly accentuated *if it is longer than the preceding duration d_{i-1} and the next duration is of the same length, being not strongly accentuated (= the second next duration is not shorter):*

$$d_{i-1} < d_i = d_{i+1} \leq d_{i+2} \quad (d_{i+1} \not> d_{i+2}) \ .$$

The idea is that a 'break' in a succession of events (long duration) marks the end of a segment, resulting in an accent. Since successive accents make no sense, adjacent durations are not accentuated simultaneously. For fine segmentation, we distinguish between weak and strong accents, giving priority to the latter. A 'stop' — a long duration followed by a shorter one — is perceived as a 'definite' segment end, resulting in an strong accent. 'Slowing down' — moving to longer durations that are not strongly accentuated — is perceived as a 'less definite' segment ending, resulting in a weak accent. If the durations gradually increase or decrease, only the last event (indefinitely long duration) is accentuated.

Example 5 (Accentuation by timing cues) *The first quarter note in Figure 8 marked by the symbol '>' is weakly accentuated because it lies between a shorter duration and an equal one that is not strongly accentuated. Indeed, the second quarter note is not followed by a shorter duration: the following eighth note and the eighth rest constitute a quarter duration. The last quarter note, which is also marked with '>', is strongly accentuated because it lies between an equal and a shorter duration. The last note of the sequence (associated with an indefinite duration) can be considered either accentuated or not accentuated.*

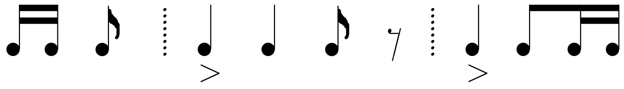


Figure 8: Accentuation by timing cues

In musical notation, the metric accent follows the bar line. By inserting the bar lines before the accentuated events in Figure 8, we thereby identify the meter 3/4.

6 Rhythmic segmentation

The introduced accentuation rules are insufficient for rhythmic segmentation. Indeed, the periodic sequence of time events in Figure 9a can be segmented with respect to the strong accents marked by '>' in two equally justified ways shown in Figures 9b–c. Experiments with humans have confirmed this ambiguity: the playback of this sequence in a loop with fading it in and out (to mask the ends of the sequence) has been segmented in both ways with almost equal probabilities.

As we can see, the accentuation alone is not sufficient for the rhythmic segmentation. Therefore, we introduce Rules 4–6 that suggest additional segmentation cues.

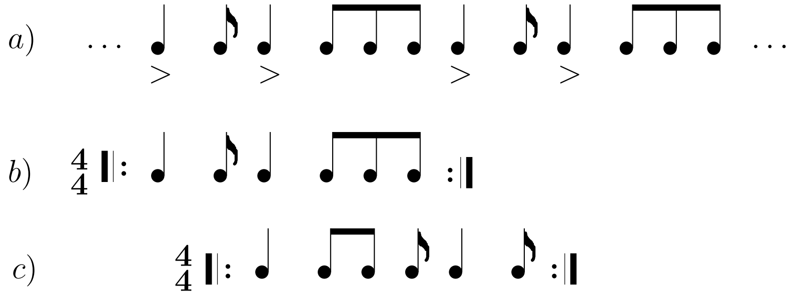


Figure 9: Rhythmic segmentation by timing cues

Rule 4 (Phrases and Syllables [Katuar 1926]) *A segment of time events that begins immediately after an accented durations and ends at an accentuated duration is called a* rhythmic phrase*. A rhythmic phrase with a single accent (at its end) is called a* rhythmic syllable*.*

Each rhythmic phrase is made up of rhythmic syllables, which are the simplest of rhythmic phrases perceived as indecomposable units. This is confirmed by the following audio experiment. Figure 10 shows a rhythmic syllable, the two eighths of which have a fixed absolute duration of 0.2 seconds. This syllable is reproduced repeatedly with variable delays divisible by 0.2*sec*, e.g. 0.8, 1.0, 1.2, 0.8, . . . *sec.* — to keep the pulse of eighths. (Tempo determination by the common time divisor was suggested by [Messiaen 1944]). However, the participants in our experiment did not perceive a constant tempo with changing pause lengths, but a variable tempo, i.e. the rhythmic syllable was perceived as a 'non-rhythmic' reference mark for tempo tracking. All of these speak for the perception of rhythmic syllables as whole and not as composite units.

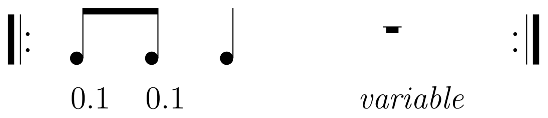


Figure 10: Rhythmic syllable as an indecomposable rhythmic unit

The experiment results can be easily explained from the viewpoint of the principle of correlativity of perception: it is simpler to represent the time events using a single small rhythmic syllable and a tempo curve than to store a long complex rhythmic structure at a constant tempo.

7 Operations on rhythmic patterns

Rule 5 (Elaboration [Mont-Reynaud and Goldstein 1985]) *Rhythmic pattern A is an* elaboration *of rhythmic pattern B (denoted as $A = E(B)$) if the durations of A result from dividing the durations of B.*

A possible elaboration of a quarter duration is shown in Figure 11.

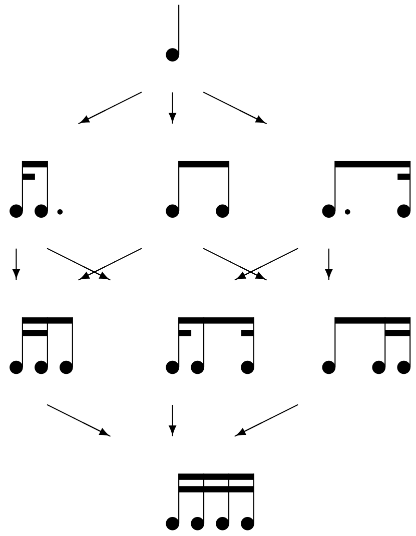


Figure 11: Elaboration of a quarter duration

Rule 6 (Sum and Union of Rhythmic Syllables) *A concatenation of rhythmic syllables is called their* sum*. A* union *of two rhythmic syllables is an elaboration of their sum with a unique strong accent at the end.*

Since the last duration of a syllable is not fixed, the intermediate duration between two concatenated syllables can be arbitrary, meaning that rhythmic syllables can be summarized in different ways. If a sum of rhythmic syllables has a unique accent (at the end) then it is already their union. If the sum has internal accents, they can be suppressed by dividing (= elaborating) the durations of the sum.

Example 6 (Sums and Unions of Rhythmic Syllables) *Figure 12 displays two sums of rhythmic syllables $A + B$ differing in the intermediate duration shown by dots. To suppress the accent at the long intermediate duration, the latter is divided into eighths, transforming the sum into union $E(A + B)$.*

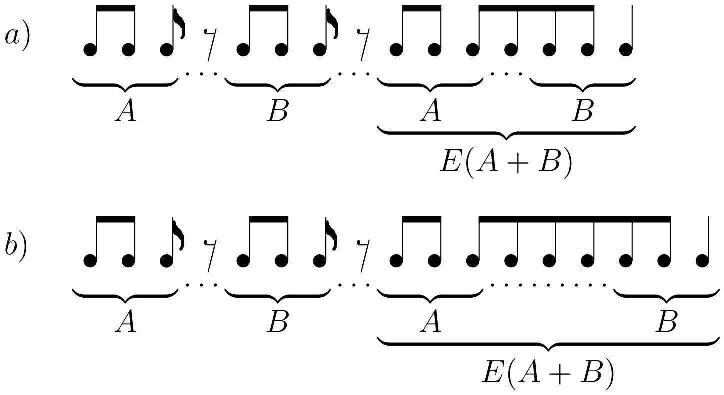


Figure 12: Two different unions of rhythmic syllables A and B

The above rules explain the 'naturalness' of rhythmic structures with the segment ratio $1 : 1 : 2 : 4 : \ldots$. Such a structure contains a segment, its repetition or elaboration, the elaboration of their sum, etc., with each next segment being the elaboration of the sum of all preceding segments. Thereby, some music material is always first exposed and then elaborated, resulting in a progression of easily perceptible variations.

8 Definition of meter and rhythm complexity

The rhythmic patterns generated by elaboration of a certain root pattern constitute an *ordered directed set.* Such a set is illustrated in Figure 11, where the generative pattern, the quarter note, is at the top and its successive elaborations are indicated by arrows.

The patterns of the same total duration that are not elaborations of each other (as in the second row of Figure 11) are of particular interest. If a sequence of time events contains these patterns then the only common pulse is that of the root (in our example, it is the pulse of quarters) without embedded pulses. This 'minimal' pulse together with accents is used to determine the time (meter) of the sequence of events. In other words, time is defined by the rhythm of the generative roots.

Rule 7 (Determination of Time) *If a sequence of time events is generated by elaboration of certain root patterns then the time (= meter) of the given sequence is determined by the duration ratio of these roots. In hierarchical representations of time events, the time patterns form the middle level, being superior to low-level rhythmic patterns but inferior to the high-level pattern of tempo curve.*

The number of 'small' rhythmic patterns that are not elaborations of one another characterizes the *rhythm complexity.* For example, Figure 13 shows the rhythm, the pulse of which is generated by patterns I and II, which are elaborations of a quarter duration, but not elaborations of each other. Therefore, its *complexity index* could be assumed 2.

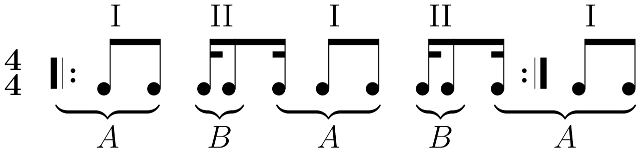


Figure 13: Rhythm with the complexity index 2

A definition of rhythm complexity, which requires the *a priori* knowledge of the pulse, is not self-sufficient. This disadvantage is overcome by referring instead to well-defined rhythmic syllables. Decomposing the rhythm in Figure 13 into two syllables A and B, we get the same complexity index 2, but requiring no *a priory* knowledge of the pulse.

Both definitions of rhythm complexity meet the ideas of [Messiaen 1944] who has characterized the rhythm diversity by the number of non-commensurable patterns used.

9 Example of analysis

Taking into account two approaches to rhythm segmentation — with respect to the (known) pulse train or by phrasing in the sense of Rules 3–4, Figure 14 displays two

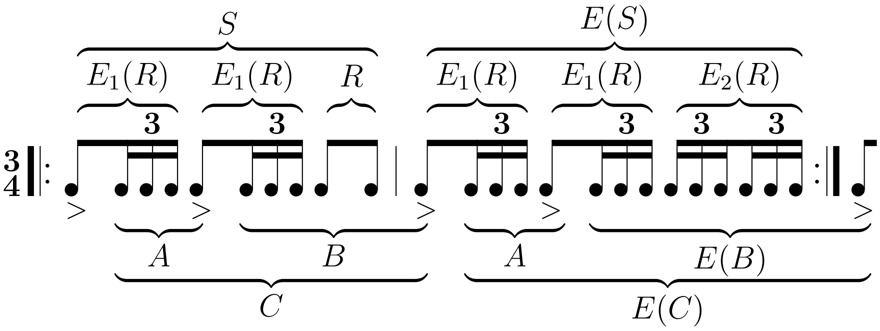


Figure 14: *Bolero's* rhythm segmented with respect to the pulse and rhythmic syllables

segmentations of the snare drum part from Ravel's *Bolero*. The segmentation with respect to the pulse of quarter durations, the same as in Figure 5, is shown by the upper braces. The three segments used — root pattern R and its two elaborations $E_1(R)$ and $E_2(R)$ — determine the rhythm complexity index 3. The six-step period consists of two similar (equal to within elaboration) segments S and $E(S)$ with three pulse durations each, implying the rhythm's three-step meter 3/4 — as in Ravel's score.

The bottom braces show the rhythmic syllables that end at the strong accents marked with '>'. The three syllables used — A, B and the elaboration $E(B)$ of the latter — determine the rhythm complexity index 3. The six-step period falls into two similar phrases (equal to within elaboration) $C = A + B$ and $E(C) = A + E(B)$, with B's duration being twice longer than that of A, which prompts the three-step meter 3/4.

10 Summary

Let us summarize the main points of the paper.

1. Our approach to rhythm recognition is based on some general provisions (correlativity of perception, optimal data representation), some heuristics (coding conventions for the tempo curve and estimation of the complexity of rhythm), and some particularities of hearing (priority of longer durations in the rhythm and tempo perception).

2. Rhythm and tempo are *complementary structures* that characterize time events. Rhythmic patterns are generative units, and the tempo curve describes their time relationships. In a sense, rhythmic patterns are reference marks for tempo tracking. The interdependence of rhythm and tempo is overcome by the 'external' criterion of the least complex data representation. The optimal data representations are found by the method of variable resolution.

3. The meter is the 'root representation' of the rhythmic structure, which consists of segments of time events. To find and classify small rhythmic patterns, rules of accentuation and elaboration are introduced. Accents are associated with longer durations that end rhythmic syllables, and the similarity of syllables is explained in terms of their elaboration.

It should be noted that our model based on characterizing musical rhythms solely by the durations between sound onsets is not universal. It does not take into account dynamics, melody, harmony and other indications. In particular, the definition of a rhythmic syllable as ending at an accented event (long duration) neglects feminine endings that are caused by melodic suspensions and resolutions (non-accented event after an accented one). For example, our definition fails to recognize a feminine rhythmic syllable when an open hi-hat is struck on the first beat, which gives a 'sustained, suspended tone', and closes on the second beat, which gives a short 'resolution'. In order to recognize 'rhythmic suspensions' and some other effects, one would take into account the 'articulation' of rhythm, when not only sound onsets but also sound ends are considered 'time events'.

References

[Bamberger 1980] Bamberger J (1980) Cognitive structuring in the apprehension of simple rhythms. *Archives de Psychologie,* 48, 171–199.

[Boroda 1985] Boroda MG (1985) On some rules of rhythmic recurrence in folk and professional music. *Kompleksnoe Izutchenie Muzykalnogo Tvortchestva: Konzepziya, Problemy, Perspektivy.* Nauka, Tbilisi, 135–167 (Russian).

[Boroda 1988] Boroda MG (1988) Towards the basic semantic units of a musical text. *Musikometrika*, 1. Brockmeyer, Bochum, 11–68.

[Boroda 1991] Boroda MG (1991) The concept of "metrical force" in music with bar structure. *Musikometrika*, 3. Brockmeyer, Bochum, 59–94.

[Bouman and Liu 1991] Bouman Ch, Liu B (1991) Multiple resolution segmentation of textural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 13(2), 99–113.

[Bregman 1990] Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound.* MIT Press, Cambridge MA.

[Calude 1988] Calude C (1988) *Theories of Computational Complexity.* North-Holland, Amsterdam.

[Clarke 1987] Clarke E (1987) Categorical rhythm perception, an ecological perspective. In: Gabrielsson A (Ed) *Action and Perception in Rhythm and Music.* Publication of Royal Swedish Academy of Music No. 55, Stockholm.

[Clarke and Krumhansl 1990] Clarke E, Krumhansl CL (1990) Perceiving musical time. *Music Perception*, 7, 213–252.

[Desain 1992] Desain P (1992) A (de)composable theory of rhythm. *Music Perception,* 9(4), 439–454.

[Desain and Honing 1989] Desain P, Honing H (1989) Quantization of musical time: a connectionist approach. *Computer Music Journal,* 13(3), 56–66.

[Desain and Honing 1991] Desain P, Honing H (1991) Tempo curves considered harmful. *Proceedings of the International Computer Music Conference'1991.* Faculty of Music, McGill University, Montreal, 143–149.

[Desain et al. 1989] Desain P, Honing H, de Rijk K (1989) A connectionist quantizer. *Proceedings of the International Computer Music Conference'1989.* Computer Music Association, San Francisco, 80–85.

[Dumesnil 1979] Dumesnil R (1979) *Le Rythme musical.* Slatkine Reprints, serie "Ressources", Paris–Genève.

[Fraisse 1982] Fraisse P (1982) Rhythm and tempo. In: Deutsch D (Ed) *The Psychology of Music.* Academic Press, Orlando FL, 149–180.

[Hummel 1987] Hummel R (1987) The scale-space formulation of pyramid data structures. In: Uhr L (Ed) *Parallel Computer Vision.* Academic Press, Boston, 125–146.

[Katuar 1926] Katuar G (1926) *Muzykalnaya Forma. 1. Ritm.* Moscow (Russian).

[Kolmogorov 1965] Kolmogorov AN (1965) Three approaches to defining the notion "quantity of information". *Problemy Peredatchi Informatsii,* 1(1), 3–11. Reprinted in: Kolmogorov AN (1987) *Theory of Information and Theory of Algorithms.* Nauka, Moscow, 213–223 (Russian).

[Lerdahl and Jackendoff 1983] Lerdahl F, Jackendoff R (1983) *A Generative Theory of Tonal Music.* MIT Press, Cambridge MA.

[Longuet-Higgins 1976] Longuet-Higgins HC (1976) The perception of melodies. *Nature,* 263, 646–653.

[Longuet-Higgins 1987] Longuet-Higgins HC (1987) *Mental Processes.* MIT Press, Cambridge MA.

[Longuet-Higgins and Lee 1982] Longuet-Higgins HC, Lee CS (1982) The perception of music rhythms. *Perception*, 11, 115–128.

[Longuet-Higgins and Lee 1984] Longuet-Higgins HC, Lee CS (1984) The rhythmic interpretation of monophonic music. *Music Perception*, 1, 424–441.

[Messiaen 1944] Messiaen O (1944) *Technique de mon langage musical. Vol. 1.* Leduc, Paris.

[Michon 1964] Michon JA (1964) Studies on subjective duration. *Acta Psychologica*, 222, 441–450.

[Minsky 1975] Minsky M (1975) A framework for representing knowledge. In: Winston PH (Ed) *The Psychology of Computer Vision.* McGraw-Hill, New York, 211–277.

[Minsky and Papert 1988] Minsky M, Papert S (1988) *Perceptrons, 2nd ed.* MIT Press, Cambridge MA.

[Mont-Reynaud and Goldstein 1985] Mont-Reynaud B, Goldstein M (1985) On finding rhythmic patterns in musical lines. *Proceedings of the International Computer Music Conference'1985.* Computer Music Association, San Francisco, 391–397.

[Palmer 1975] Palmer SE (1975) Visual perception and world knowledge: notes on a model of sensory-cognitive interaction. In: Norman DA et al (Eds) *Exploration in Cognition.* Erlbaum, Hillsdale NJ, 279–307.

[Palmer 1982] Palmer SE (1982) Symmetry, transformation, and the structure of perceptual systems. In: Beck J (Ed) *Organization and Representation in Perception.* Erlbaum, Hillsdale NJ, 95–107.

[Palmer 1983] Palmer SE (1983) The psychology of perceptual organization: a transformational approach. In: Beck J, Hope B, Rosenfeld A (Eds) *Human and Machine Vision.* Academic Press, New York, 269–339.

[Pistone 1977] Pistone D (1977) Tempo. In: Honnegger M (Ed) *Dictionnaire de la Musique. Science de la Musique. Formes, Technique, Instruments.* Bordas, Paris.

[Porte 1977] Porte D (1977) Rythme. In: Honnegger M (Ed) *Dictionnaire de la Musique. Science de la Musique. Formes, Technique, Instruments.* Bordas, Paris.

[Povel and Essens 1985] Povel DJ, Essens P (1985) Perception of temporal patterns. *Music Perception,* 2(4), 411–440.

[Rosenthal 1988] Rosenthal D (1988) A model of the process of listening to simple rhythms. *Proceedings of the 14th International Computer Music Conference.* Feedback-Studio-Verlag, Köln, 189–197.

[Rosenthal 1989] Rosenthal D (1989) A model of the process of listening to simple rhythms. *Music Perception*, 6(3), 315–328.

[Rosenthal 1992] Rosenthal D (1992) Intelligent rhythm tracking. *Proceedings of the International Computer Music Conference'1992.* Computer Music Association, San Francisco, 227–230.

[Schloss 1985] Schloss WA (1985) *On the Automatic Transcription of Percussive Music. From Acoustical Signal to High Level Analysis.* Stanford University, Dep. of Music Report STAN-M-27, Stanford CA.

[Skrjabin 1960] Skrjabin A (1960) *Poem for Piano. Op. 32 No. 1. The text of author's performance by recording on "Velte-Mignon". Transcribed by P. Lobanov.* Gosudarstvennoye Muzykalnoye Izdatelstvo, Moscow (Russian).

[Tangian 1993] Tanguiane AS (1993) *Artificial Perception and Music Recognition.* Springer, Berlin (Lecture Notes in Artificial Intelligence No. 746).

[Tangian 1994] Tanguiane AS (1994) A principle of correlativity of perception and its applications to music recognition. *Music Perception,* 11 (4), 465–502.

[Viret 1977] Viret J (1977) Mesure. In: Honnegger M (Ed) *Dictionnaire de la Musique. Science de la Musique. Formes, Technique, Instruments.* Bordas, Paris.

[Wertheimer 1923] Wertheimer M (1923) Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung*, 4, 301–350. Condensed transl in: Ellis WD *A Source Book of Gestalt Psychology, Selection 5.* New York: Humanities Press, 1950. Also in: Beardslee DC, Wertheimer M (eds) *Readings in Perception, Selection 8.* Princeton NJ: Van Nostrand Reinhold, 1958.

[Witkin 1983] Witkin AP (1983) Scale-space filtering. *Proceedings of the 8th International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany,* 1019–1024.

[Witkin and Tenenbaum 1983] Witkin AP, Tenenbaum JM (1983) On the role of structure in vision. In: Beck J, Hope B, Rosenfeld A (Eds) *Human and Machine Vision.* Academic Press, New York, 481–543.

KIT Scientific Working Papers
ISSN 2194-1629

www.kit.edu