Chapter 6

Applications to Rhythm Recognition

6.1 Problem of Rhythm Recognition

The nature of time, rhythm, and tempo is one of the first questions posed by music theory. Many distinguished music theorists have contributed to the understanding of related perception mechanisms. Recently, a series of works has been published where the problem of rhythm and tempo is formulated with regard to computer rhythm recognition and tempo tracking (see Section 1.2 for references).

However, further progress in modeling rhythm and tempo perception is constrained by the lack of explicit definitions of these concepts. The known definitions are rather ambiguous. For example, rhythm is defined as the order and the proportion of durations (Porte 1977); tempo is explained as a characteristic of execution motion with respect to measures and melodic, harmonic, rhythmic, or dynamical cues (Pistone 1977); time is considered as a form of determination of rational proportions of rhythm (Viret 1977). The incompleteness of these definitions is obvious; however, according to Dumesnil (1979), one can hardly find better definitions, even in special psychomusicological publications like (Fraisse 1983; Povel & Essens 1985).

The definitions of tempo and rhythm are not only incomplete but also interdependent. On the one hand, the tempo is defined with respect to a certain rhythm (if there are no events, no motion can be perceived). On the other hand, in order to measure the proportion of durations, the rhythm is defined with respect to a certain tempo. This makes a kind of logical circle.

Attempting to overcome such an ambiguity in defining interdependent concepts, we apply the principle of correlativity of perception. According to this principle, time data are represented in terms of repetitious low-level configurations and a high-level configuration of their relationships. The grouping of

time events into low-level configurations is realized with respect to the simplicity principle, meaning that the representation of time events by high-level and low-level configurations requires least memory.

In our case, the low-level configurations are correlated *rhythmic patterns*. The high-level configuration, being determined by time relationships between the rhythmic patterns correlated, is associated with the *tempo curve*. In other words, we consider repetitious rhythmic patterns as recognizable reference units for tempo tracking. Drawing analogy to vision, similar rhythmic patterns correspond to instant states of an object, and the tempo curve corresponds to the object trajectory in time. Table 6.1 shows the analogy between rhythmic patterns and visual patterns in Fig. 2.1. Note that the tempo curve is not supposed to be continuous, which meets Desain & Honing's (1991) view at tempo phenomenon.

As mentioned in Section 2.2, the advantage of such a representation is the recognizability of high-level patterns regardless of the recognizability of low-level patterns. This means that we may have difficulties in identifying a rhythm as waltz, march, etc., while being capable to recognize its tempo. On the other hand, if the tempo has been determined, the rhythm recognition becomes easier.

The goal of this chapter is developing a technique for finding rhythmic patterns. First of all we consider the problem of rhythm recognition in general, and then restrict our attention to the segmentation of rhythmic progressions with respect to timing accentuation. For this purpose some formal rules of timing accentuation and classification of rhythmic patterns are introduced. Then the rules of segmentation are illustrated with an example of rhythm recognition.

In Section 6.2, "Rhythm and Correlative Perception", we formulate the problem of rhythm recognition from the standpoint of the principle of correlativity of perception. In case of rhythm, the two-level scheme of data representation is generalized to a multi-level scheme, since the rhythm is represented as a hierarchy of embedded rhythms. Such a multiple embedding makes the rhythm redundant and thus easier recognizable under considerable tempo fluctuations.

In Section 6.3, "Correlativity and Recognition of Periodicity", we apply the principle of correlativity of perception to the recognition of quasi-periodicity in a sequence of time events, corresponding to the recognition of repetitions under variable tempo. In particular, we trace the operation of the method of variable resolution with a simple example and compare the results obtained by our model with some conclusions from known psychological experiments.

In Section 6.4, "Timing Accentuation", we pose the problem of rhythmic segmentation. In order to solve this problem we formulate rules for distinguishing certain events which are called accentuated. In the given study we consider

Table 6.1: Correspondence between visual and time data

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

the accentuation with respect to time cues only, ignoring pitch and dynamic cues. We define strong and weak accents and illustrate these definitions with an example.

In Section 6.5, "Rhythmic Segmentation", we show that the accentuation alone is not sufficient for unambiguous segmentation of time events. In order to realize the segmentation and classification of segments, the notion of rhythmic syllable is introduced. A rhythmic syllable is understood to be a sequence of time events with the only accent at the last event. Referring to a simple psychoacoustic experiment, we show that rhythmic syllables are perceived as indecomposable rhythmic units.

In Section 6.6, "Operations on Rhythmic Patterns", we consider a kind of rhythmic grammar. We define the elaboration of a rhythmic pattern as a subdivision of its durations which preserves the original pulse train. The elaboration is also explained in terms of correlation of rhythmic patterns. The junction of rhythmic syllables is defined to be the elaboration of their sum. This way we define possible transformations of rhythmic patterns other than distortions caused by tempo fluctuations. Using the concepts of elaboration and junction we explain certain properties of rhythm organization.

In Section 6.7, "Definition of Time and Rhythm Complexity", we define a root pattern to be the simplest pre-image (with respect to the elaboration) of a generative syllable of the rhythm. In order to recognize such a pattern the above mentioned concepts of rhythmic syllable and their junction are used. The notion of root pattern is applied to time determination and estimation of rhythm complexity.

In Section 6.8, "Example of Analysis", we consider the snare drum part from Ravel's *Bolero* and trace the operation of our model of rhythm segmentation step by step. As a result, the sequence of time events is represented in terms of generative syllables and their transformations. This representation reveals root patterns, their rhythmic elaboration, and thus the structure of the rhythm. Finally, the representation obtained is used to determine the time of the rhythm.

In Section 6.9, "Summary of Rhythm Perception Modeling", the main statements of the chapter are recapitulated.

6.2 Rhythm and Correlative Perception

As mentioned in the previous section, our approach to rhythm and tempo recognition is based on a two-level representation of time events in terms of generative rhythmic patterns and their transformations. The transformations fall into

- distortions which are caused by tempo fluctuations, and
- variations which are caused by musical elaboration.

We shall consider both types of transformations, using special techniques for each type.

Firstly, restrict our attention to distortions caused by tempo fluctuations, supposing that there are no elaboration of rhythmic patterns. The problem is to interpret a sequence of time events either in terms of rhythm, or in terms of tempo, or both. For example, consider the sequence of time events shown in Fig. 2.3. In Fig. 2.4a this sequence of events is interpreted in terms of rhythm, i.e. as a single rhythmic pattern under a constant tempo. In Fig. 2.4b this sequence of events is interpreted in terms of tempo, e.g. as a repeat of the first three durations (generative rhythmic pattern) under a tempo change. Moreover, every sequence of time events can be interpreted as generated by a fixed single duration which is getting shorter or longer because of tempo changes at each event. On the other hand, it can be interpreted as a single complex rhythmic pattern under a constant tempo (cf. Fig. 2.3–2.4 from Section 2.2). However, the two extreme representations are complex, the former owing to the complexity of the tempo curve, and the latter owing to the complexity of the rhythmic pattern. We suppose that in most cases there exists a compromise representation, with a few rather simple generative rhythmic patterns and a rather simple tempo curve, while the total complexity being quite moderate.

The choice of interpretation can be influenced by tradition, context, or some special intentions. In Section 2.2 we have shown that the interpretation may depend on melodic context. For this purpose we have estimated the complexity of alternative representations. In case of a pure rhythm, the complexity is least for the representation of time events as a single pattern, whereas in case of melodic context its representation as a repeat is preferable.

This example illustrates the idea that the recognition of rhythmic patterns determines the perception of tempo. If in Fig. 2.4a and Fig. 2.4c the three quavers are recognized as a repet of the three crotchets, then the tempo is perceived as changing. If the group of the last three durations is not recognized as a repeat of the first three durations, then the tempo is perceived as constant.

At the same time, the recognition of rhythmic patterns depends on tempo. Indeed, rhythmic patterns are perceived as repetitious, only if they are comparable in time, i.e. if they are considered with respect to a certain tempo.

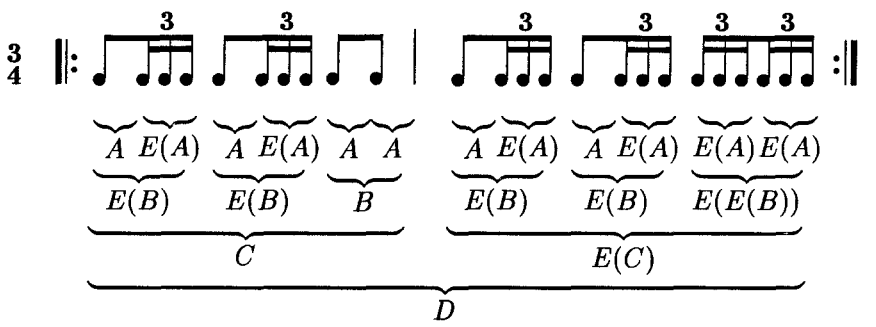


Figure 6.1: Multilevel repetition structure of rhythm from *Bolero* by M.Ravel

Therefore, rhythm and tempo are interdependent concepts, which implies the impossibility of their separate recognition. Nevertheless, in the model of correlative perception their functions are different, and the interaction between them is determined with respect to the criterion of least complex representation.

Thus the task of rhythm and tempo recognition is formulated as finding the least complex representation of a sequence of time events. This representation reveals the regularity of time organization which in case of Western music is usually observed at several levels simultaneously (Lerdahl & Jackendoff 1983), corresponding to simultaneous perception of musical texture at several levels, e.g. at the level of measures, couples of measures, etc., up to constituents of musical form.

Therefore, a model of rhythm perception should be multi-level, while each level being characterized by its own generative patterns determined by generative patterns of lower level. These levels should be commensurable with respect to tempo, otherwise each would have its own tempo curve, which would complicate the representation of data. The desired representation of a given sequence of time events can be understood as a tree with indecomposable rhythmic segments at the low level, and branches which show their grouping into the patterns of higher levels.

In order to recognize the multi-level regularity, we have to consider the second type of transformations, variations of rhythmic patterns caused by musical elaboration. Let us illustrate the idea of multi-level regularity with an example of snare drum part from Ravel's *Bolero* (Fig. 6.1).

At the first level the regularity emerges as the commensurability of the quaver durations which are braced in the first subscript line in Fig. 6.1. Since quaver durations are repeated regularly, the pulse train of quavers is quite evident.

According to Mont-Reynaud & Goldstein (1985), a subdivision of a rhythmic pattern is said to be its *elaboration.* With respect to elaboration, the quaver patterns are classified into *root patterns* and their derivatives. In Fig. 6.1 a root pattern is denoted by a letter, e.g. A, and its elaboration is denoted by $E(\cdot)$, e.g. $E(A)$.

At the second level, the rhythm regularity emerges as the commensurability of crotchet durations and their elaborations, braced in the second subscript line in Fig. 6.1. Similarly to rhythmic patterns of the first level, patterns of the second level are grouped into the patterns of the third level, braced in the third subscript line in Fig. 6.1, which in turn are grouped into the patterns of the fourth level.

The idea of such an arrangement of rhythmic patterns is the representation of the *embedded pulse train of the rhythm.* Thus the rhythm of *Bolero* is characterized by the following four embedded levels:

1. Quavers (elaborations of root pattern A).
2. Crotchets (elaborations of root pattern B).
3. Three-quarter measures (elaborations of root pattern C).
4. Repetitious two-measure segments D.

Such a multilevel representation of rhythm is used further for classification of rhythmic patterns and determination of time.

The embedded pulse train is inherent only in so called *divisible rhythms* which are used in Western music. The rhythmic structure of other cultural traditions is not described by such multi-level schemes.

It is remarkable that the tempo is most free in Western music as well. This can be explained by the fact that several embedded pulse trains make a rhythm redundant (since there are several cues to recognize repetitions) and thus recognizable even under severe distortions. Similarly to the example in Fig. 2.4c–d, where a repetitious rhythm is complemented with a repetitious intonation, a redundant rhythm is complemented with an embedded rhythm of another level, implying its total representation as a repeat to be less complex than its representation as a single rhythmic pattern. Hence, a redundant rhythmic structure is recognizable even under considerable tempo deviations.

The range of tempo fluctuations which do not prevent from perceiving repetitions is larger if there are several complementary cues like melodic intonation, similarity of accompaniment, harmonic pulsation, etc. For example in Skrjabin's performance of his *Poem* Op. 32 No. 1 transcribed from a piano-roll recording (Skrjabin 1960) the tempo varies within limits $\dot{\text{♩}}. = 19 \div 110$, i.e. up to 5.5 times. These fluctuations are unambiguously perceived as tempo changes but not as changes of durations (changes of rhythmic patterns), owing

to several complementary cues in the repetitious texture like accompaniment figures, phrasing, etc.

Tempo changes are strictly prohibited in Bulgarian or Turkish music with so called *additive rhythms* with complex duration ratios. From the standpoint of the correlativity principle it is explained by the fact that if a rhythm is not structurally redundant, then even minor tempo deviations are not perceived as *accellerando* or *ritardando* but rather as changes of durations (rhythmic changes), implying an inadequate perception of musical meaning.

Thus we have shown that the rhythm recognizability under tempo deviations depends on the rhythm redundancy caused by rhythmic elaboration. This dependence means that in a model of rhythm recognition the two types of transformations of rhythmic patterns, caused by tempo fluctuations and musical elaboration, should be taken into account simultaneously.

6.3 Correlativity and Recognition of Periodicity

In the remainder of the chapter we discuss different elements of our approach to rhythm recognition, detection of periodicity, accentuation, segmentation, and time determination.

Further we use the notions of *periodicity* and *rhythm*. By periodicity we mean a repetition of events or their groups with no segmentation into rhythmic patterns. If time events are periodically segmented with respect to the time structure, we obtain the concept of rhythm.

At first, consider the problem of recognizing periodicity. In our model of correlative perception the recognition of repetitious rhythmic patterns is based on autocorrelation analysis of a sequence of time events under various distortions of time scale. The distortions which provide high correlation of patterns can be found by the method of variable resolution introduced in Section 2.4.

Let us trace the operation of the model of correlative perception with a simple example of recognizing quasi-periodicity (generative element) in a sequence of time events in Fig. 6.2.

Consider the following sequence of tone onsets recordered to within one time unit which is equal to 1/20 sec. (Fig. 6.2a):

$$t = 0, 10, 19, 30.$$

Define the *event function*, for all $t = 0, 1, \ldots, 30$ putting

$$s(t) = \begin{cases} 1 & \text{if } t = 0, 10, 19, 30, \\ 0 & \text{otherwise.} \end{cases}$$

The event function $s(t)$ in a form of binary string is shown in Fig. 6.2b. One can try to determine its period p by finding the peaks of autocorrelation function

$$R(p) = \sum_t s(t-p)\, s(t).$$

However, no autocorrelation and, consequently, no periodicity is recognizable (see the second column in Table 6.2). According to the method of variable resolution, the accuracy of representation in Fig. 6.2b should be reduced as shown in Fig. 6.2c. Then autocorrelation function $R(p)$ has salient peaks (see the third column in Table 6.2) and the periodicity is recognizable. After the correlated events have already been known, these events should be locally shifted in order to increase in the correlation. The result of this procedure is shown in Fig. 6.2d. Restoring the resolution, we obtain the sequence in Fig. 6.2e together with the description of tempo fluctuations corresponding to the shifts of time events determined earlier.

Now one has to accept or reject the hypothesis about the representability of this sequence of time events as generated by a repetitious rhythmic pattern, which in the given case is a single duration of 10 units. For that purpose one has to compare the complexity of two alternative representations. The first representation corresponds to storing the period of 10 units with a repeat algorithm (or its call) provided with a tempo curve. The second representation corresponds to coding the given sequence of time events as it is, i.e. as a single rhythmic pattern under a constant tempo. Obviously, the complexity of total representation depends on the way of coding repeat algorithm and tempo curve.

By some reasons, usual algorithms of coding functions (by means of storing the coefficients of their polynomial approximations or Fourier series) are not suitable for our purposes. Usual ways of coding require the knowledge of the function values on the whole domain of its definition, whereas we would like to track the tempo in real time, coding the tempo curve while processing current data.

One of possible ways of coding the tempo curve is fixing the time moments when the tempo deviates from its current value, say, by more than 5%. Such a heuristic algorithm meets zonal properties of perception and its logarithmic sensitivity to absolute values. On the other hand, it is quite simple, which is desirable in computer experiments.

The model of recognition of rhythm periodicity has been tested with a series of computer experiments on the recognition of rhythm from *Bolero* (Fig. 6.1) performed by striking a computer keyboard. The experiments have revealed the repetitious structure of the rhythm at four levels shown in Fig. 6.1. The repetition has been recognized also at the level of sixteenth triplets, even in cases when their duration ratio 1:1:1 has been distorted up to the ratio 7:6:10.

Such a great variance of relative durations observed for sixteenth triplets

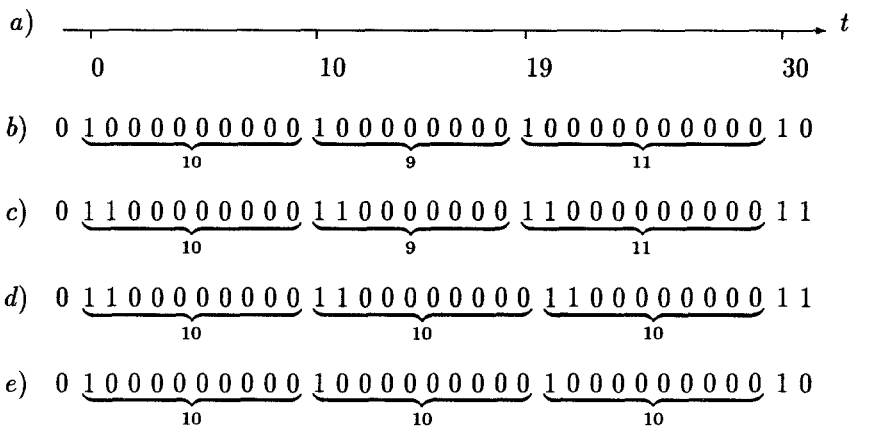


Figure 6.2: Representation of time events with variable resolution

Table 6.2: Autocorrelation $R(p)$ of time events

p	$R(p)$ in Fig. 6.2b	$R(p)$ in Fig. 6.2c	$R(p)$ in Fig. 6.2d	$R(p)$ in Fig. 6.2e
...	0	0	0	0
8	0	1	0	0
9	1	3	3	0
10	1	4	6	3
11	1	3	3	0
12	0	1	0	0
...	0	0	0	0
18	0	1	0	0
19	1	3	2	0
20	1	3	4	2
21	0	1	2	0
22	0	0	0	0
...	0	0	0	0
29	0	1	1	0
30	1	2	2	1
31	0	1	1	0

hasn't been observed for longer durations. This means that these short durations are less important for the perception of periodicity than longer durations, which in our experiments correspond to eighths and crotchets.

The above conclusion meets the experimentally established fact that the perception is most sensible to tempo fluctuations if the durations are about 0.1–1.0 second (Michon 1964). These durations are usually considered as fundamental in rhythm perception, and therefore the model should treat these durations with a certain priority.

Thus the model of correlative perception reveals repetitious rhythmic patterns under variable tempo. However, recognizing a periodicity is not yet recognizing a rhythm. In order to recognize a rhythm, a sequence of time events must be segmented and the resulting segments must be classified. The related procedures are described in the following sections.

6.4 Timing Accentuation

As mentioned at the beginning of the previous section, in order to be considered as a rhythm, a periodical sequence of time events should be segmented, i.e. certain events must be recognized as starting points of periods.

A distinct periodicity can lack an unambiguous segmentation, as in African percussion music where almost every time event may be considered as a start point of a period (Schloss 1985). In such a case, illustrated by Fig. 6.3 we say that we can recognize no rhythm but just a periodicity.

Thus in order to perceive a rhythm unambiguously, one has to recognize both periodicity and start points of periods which are accentuated somehow. This means that we have to distinguish between accentuated and non-accentuated events. All the known ways of accentuation are based on breaking the homogeneity in pitch, harmony, dynamics, timbre, etc.

In the present study we restrict our attention to timing accentuation. In speech the timing accentuation is realized by pauses and longer vowels, i.e. by longer durations. Generally speaking, the same is valid for music.

Summing up what has been said, let us introduce the following rules of timing accentuation which are similar to some rules formulated by Boroda (1985; 1988; 1991).

Rule 1 (Durations) *The only characteristic of a time event is the duration of time interval between its onset and the onset of the next time event. This interonset time interval is said to be the duration associated with the event. If the given time event is the last in the sequence, the associated duration is assumed to be not fixed.*

This rule restricts the attention to timing cues only. An important remark

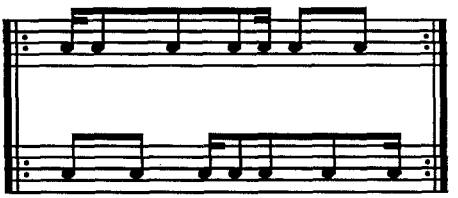


Figure 6.3: Ambiguous segmentation of a periodical sequence of time events

concerns the last time event whose duration can be arbitrarily long. The use of this assumption will be clear in the sequel.

Rule 2 (Accentuation Distinguishability) *In order to distinguish accentuated events in a sequence of time events, at least two types of durations are necessary.*

This rule postulates the case when timing accentuation is possible.

Rule 3 (Accentuated Durations) *The duration associated with an event is said to be strongly accentuated if*

(a) it is longer than its closest neighbors, i.e. it follows a shorter duration and the next one is also shorter;

(b) it follows an equal duration and the next one is shorter.

The duration is said to be weakly accentuated if

(c) it follows a shorter duration and precedes an equal duration which is not strongly accentuated, i.e. the second next is not shorter.

A time event is said to be accentuated (strongly or weakly) if the duration associated with the event is accentuated (strongly or weakly, respectively).

The idea of the third rule is that a longer duration which is adjacent to a shorter one is accentuated. The most evident case (a) concerns a situation when a longer duration is between two shorter ones. If a duration is between shorter and equal one then it is usually accentuated (case b and c), yet in order to avoid simultaneous accents at two equal successive durations between two shorter ones, we assume no accent at the first duration (case c). No accentuation emerges when durations are successively getting shorter or longer.

In order to provide unambiguous segmentation when several accentuated durations emerge in a short phrase, we distinguish between strong and weak accents, with the priority of strong accents. We suppose that a change from a longer duration to a shorter one is immediately recognized, resulting in an accent. Yet after a change from a shorter duration to a longer duration one can expect some further increase in duration, resulting in a weaker sensation of accent; this is the case of weak accentuation.

To illustrate the above rules, consider the sequence of time events shown in Fig. 6.4. The first crotchet marked by symbol ">" is weakly accentuated by virtue of Rule 3c, since it is the duration between a shorter duration and an equal one. The last crotchet which is also marked by ">" is strongly accentuated by virtue of Rule 3b, since it is the duration between an equal and a shorter one. Note that although the second crotchet precedes an eighth, it is not accentuated. Indeed, by virtue of Rule 1 the duration of the event associated with this eighth is crotchet, and no shorter duration is adjacent to it. By virtue of Rule 1 the last note of the sequence can be considered both as accentuated, or not.

To show the accentuation in notation, bar lines are put before accentuated events as shown in Fig. 6.4. In the above example, the segmentation with respect to the accentuation determines the 3/4 time of the given phrase.

6.5 Rhythmic Segmentation

The accentuation defined is not sufficient for rhythmic segmentation. Indeed, consider a periodic sequence of time events segmented with respect to the accentuation defined in two different ways as shown in Fig. 6.5a and Fig. 6.5b. To prove the perceptual ambiguity of segmentation of these events, we have performed the following audio experiment: The given sequence of time events has been recorded and played back in a loop, having been amplified gradually from zero level. A series of audio tests has shown that listeners recognize the two segmentations with almost equal probability.

Thus to recognize a rhythmic segmentation we need some other cues in addition to the accentuation. For that purpose we introduce rules of classification and elaboration of rhythmic patterns. By a *rhythmic pattern* we understand any segment of a given sequence of durations. However, the most important is the case when rhythmic patterns are segmented with respect to the accentuation. Thus we obtain the following definition.

Rule 4 (Phrases and Syllables) *A rhythmic phrase is defined to be a sequence of durations which follows an accentuated duration and ends at an accentuated duration. A rhythmic phrase with the only accentuated duration is said to be a rhythmic syllable (Katuar, 1926).*

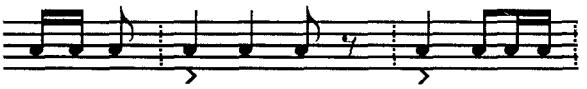


Figure 6.4: Accentuation by timing cues

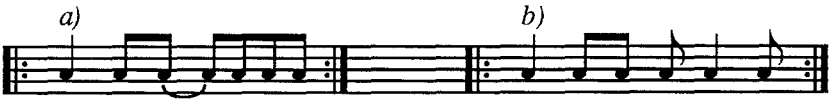


Figure 6.5: Rhythmic segmentation by timing cues

Consequently, a rhythmic syllable is a simplest rhythmic phrase. Any rhythmic phrase is formed by adding syllables to each other.

Note that by virtue of Rule 4 a syllable is determined by the durations which precede an accentuated event. The accentuated duration itself is not included into the syllable. The accentuated event just marks the end of the syllable, and the associated accentuated duration may be not fixed (cf. with Rule 1).

We suppose that rhythmic syllables are perceived as indecomposable time units. In order to prove it we have performed the following audio experiment: The rhythmic syllable shown in Fig. 6.6 with two fixed absolute durations $0.1sec$ has been repeatedly reproduced under variable delays divisible by $0.1sec$, e.g. 0.8, 1.0, 1.2, 0.8, ... *sec*. If the rhythmic syllable was perceived as a composed structure, the common time unit ($0.1sec$) would result in a sensation of constant tempo with changes of rest durations. (Tempo determination with respect to the common time unit was proposed by Messiaen (1944)). However, in our audio experiment listeners have recognized tempo deviations rather than rhythmic changes. This proves that the syllable is perceived as an entirety rather than as composed of smaller units and that the end duration is not important for identifying equal syllables.

Such a way of recognizing tempo by time intervals between the entries of

Figure 6.6: Syllable as an indecomposable unit

similar rhythmic patterns meets the principle of correlativity of perception. In fact, in our experiment we have shown that similar rhythmic patterns are used as reference indivisible units for tempo tracking. Besides, we have shown that the tempo is a percept of another level than the rhythm.

6.6 Operations on Rhythmic Patterns

In order to classify rhythmic phrases and recognize generative rhythmic patterns, we define a reflexive transitive binary relation E, *"is the elaboration of"* on the set of rhythmic patterns X. Recall that a binary relation E on X is reflexive if xEx for all $x \in X$ (a rhythmic pattern is the elaboration of itself), and transitive if xEy and yEz implies xEz for all $x, y, z \in X$ (a successive elaboration of a rhythmic pattern is its elaboration).

Rule 5 (Elaboration) *Rhythmic pattern A is the elaboration of rhythmic pattern B if A preserves the pulse train of B, i.e. if A results from a subdivision of durations of B by inserting additional time events (Mont-Reynaud & Goldstein, 1985).*

Fig. 6.7 illustrates the idea of rhythmic elaboration with an example of subdivisions of a crotchet duration (recall that by virtue of Rule 1 the crotchet duration, in order to be determined, should be followed by a next tone onset which is not shown in the figure).

The idea of elaboration can be explained in correlation terms. Represent the rhythmic patterns in Fig. 6.7 by 0 and 1 within the accuracy of a sixteenth. Then the top pattern which we denote by T is written down as follows

$$T = \{t_1 \dots t_4\} = \{1000\},$$

and the bottom pattern which we denote by B is written down as

$$B = \{b_1 \dots b_4\} = \{1111\}.$$

It is easy to see that pattern B is the elaboration of pattern T if and only if $B \supset T$. This means that B contains all ones of T. Since the number of ones in T is equal to the autocorrelation

$$R_{T,T} = \sum_{i=1}^{4} t_i \cdot t_i$$

(which in the given case is equal to 1), and the number of coinciding ones in B and T equals to the correlation

$$R_{B,T} = \sum_{i=1}^{4} b_i \cdot t_i$$

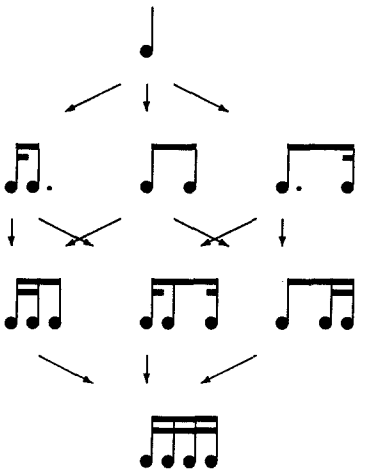


Figure 6.7: The elaboration of a crotchet rhythmic pattern

(which in the given case os equal to 1), we obtain that B is the elaboration of T if and only if

$$R_{B,T} = R_{T,T}.$$

Since the correlation is usually understood as a measure of similarity, the last equation means that the pattern B, being the elaboration of pattern T, is similar to pattern T.

For the patterns of equal duration which are not the elaboration of each other (as in the second line of Fig. 6.7), the correlation is less than autocorrelation. For example, putting

$$L = \{l_1 \dots l_4\} = \{1100\}$$

and

$$M = \{m_1 \dots m_4\} = \{1010\},$$

we obtain

$$R_{L,M} = \sum_{i=1}^{4} l_i \cdot m_i = 1 < 2 = R_{L,L} = \sum_{i=1}^{4} l_i \cdot l_i = R_{M,M} = \sum_{i=1}^{4} m_i \cdot m_i.$$

Now we define the junction of syllables.

Rule 6 (Sum and Junction of Syllables) *The sum of two successive rhythmic patterns is defined to be the rhythmic pattern constituted by the time events of these patterns which are put one after another.*

The junction of two successive rhythmic syllables is defined to be a rhythmic syllable which is the elaboration of their sum.

Note that the sum of two syllables is more than the two syllables in succession. Besides the two syllables themselves, the sum contains the *link*—the accentuated duration after the first syllable. According to the remark following Rule 4, this duration is undefined if the first syllable is considered separately, since instead of the whole duration we consider just an accent. In the sum of syllables, this accent turns to be a duration, linking the two syllables. Therefore, there can be many different sums of the same two syllables, depending on the link duration.

Also note that the sum of two rhythmic syllables is a rhythmic phrase, whereas their junction is a rhythmic syllable. This means that the sum of two syllables can have two accents, at the ends of each syllable, whereas in their junction the internal accent is suppressed by dividing the associated duration into shorter ones which are no longer accentuated. This implies that the junction has a new rhythmic quality, a through tension towards its end.

Fig. 6.8a displays two identical rhythmic syllables and their junction. The total duration of the third syllable is the same as the sum of the two syllables. This results in the symmetry of the whole passage, providing its structure to be $1 + 1 + 2$.

Consider another junction of the two syllables, for instance, obtained by adding two quavers to the third syllable as shown in Fig. 6.8b. This implies that we link the first two syllables not by a crotchet duration but by a half-note duration, i.e. we consider the elaboration of another sum of the syllables.

The connection between the three syllables in Fig. 6.8b is less evident than in Fig. 6.8a. Indeed, in Fig. 6.8a one can see not only the two syllables, but already their sum which is elaborated next. In a sense, the elaboration is already "prepared" for easy perception. On the contrary, in Fig. 6.8b the sum of the two syllables is different from the sum which is elaborated. In Fig. 6.8b the intermediate phase between the two syllables and their junction is missed, breaking the successiveness in their perception.

We could provide the effect of such a successiveness in Fig. 6.8b, making the rest between the first two syllables longer, up to a half-note duration. The duration of the second rest is not so important. Even if we change the duration of the second rest in Fig. 6.8a, the third rhythmic syllable is still perceived as the elaboration of the sum of the first two.

From our standpoint, we can explain the simplicity of rhythmic construction $1 + 1 + 2 + 4 + \ldots$. Such a structure contains a rhythmic pattern, the elaboration of the preceding segment, then the elaboration of two preceding segments, and so on. Therefore, the origin of such a structure is quite simple, adding junctions of all preceding segments. This results in perceiving such rhythms with ease; moreover, the perception is "prepared" to recognize the elaboration since the sum is already exhibited.

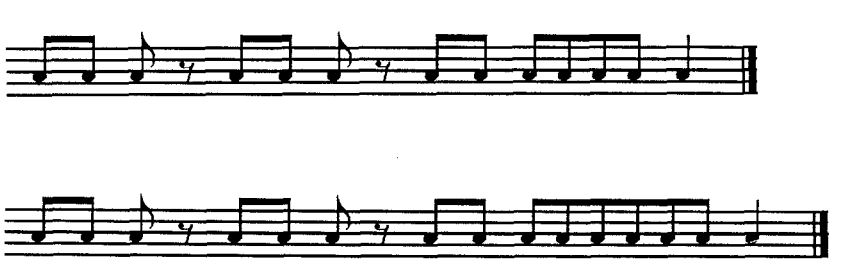


Figure 6.8: Two different junctions of the same rhythmic syllables

6.7 Definition of Time and Rhythm Complexity

Thus we have introduced the rules of representation of a given sequence of time events in terms of generative syllables. Constructing such representations, one can reveal origins of a given rhythm with conclusions concerning its time.

Note that rhythmic patterns of equal total duration constitute an *ordered directed set* with respect to the elaboration, where every two elements have a common superior—their common *root*. An example of such an order is shown in Fig. 6.7 with a common root pattern at the top and its successive elaborations indicated by arrows.

The patterns of the same total duration which are not elaborations of each other (like in the second line of Fig. 6.7) are of particular interest. If a rhythm contains such patterns then this rhythm has no embedded levels of the pulse train and can be represented as a succession of irreducible units whose pulse train becomes predominant.

The idea of a pulse train generated by indecomposable rhythmic patterns can be applied to rhythmic syllables. Since each syllable has the only accent, the accents of syllables determine a pulse train with a certain rhythm. We use this rhythm to determine the time of a given sequence of time events.

Rule 7 (Determination of Time) *If a sequence of time events is representable in terms of elaboration of certain rhythmic syllables (phrases), then the time of the given sequence is determined by the duration ratio of their roots.*

In other words, one has to find a stable preimage (with respect to the elaboration) of generative patterns.

Roughly speaking, the time is defined to be the rhythm of roots of generative syllables. In a sense, time patterns, being superior to generative patterns in the hierarchical representation, constitute an intermediate representation level between rhythmic patterns and the tempo curve (cf. Fig. 6.1).

Besides time determination, rhythmic patterns which are irreducible to each other can be used for estimating the complexity of rhythm. Indeed, their number corresponds to the number of generative patterns required to generate the given sequence.

For example, consider the rhythm in Fig. 6.9 which is constituted by two rhythmic patterns of equal duration. One can see that the crotchet duration is the root for the two rhythmic groups beamed but no rhythmic group is the elaboration of another. This means that the pulse train of crotchets is supported by no pulse train of quavers or some other shorter durations.

Such a rhythm can be considered as less redundant and therefore as more complex. The *complexity of a rhythm* can be identified with the branching index of the graph of the rhythmic patterns used, i.e. by the maximal number of irreducible to each other rhythmic configurations of the same level.

As seen from Fig. 6.1, each rhythmic level of *Bolero* is generated by a single pattern, implying its complexity being equal to 1. The rhythm in Fig. 6.9 is generated by two patterns of equal duration which are not reducible to each other; consequently, its complexity is equal to 2.

Such an understanding of rhythm complexity meets the ideas of Messiaen (1944) who has characterized the diversity of rhythm by the number of non-commensurable patterns used.

Thus finding irreducible (with respect to elaboration) patterns has two applications: time recognition and estimation of rhythm complexity.

6.8 Example of Analysis

Consider the snare drum part from *Bolero* by M.Ravel (Fig. 6.10). Since we use time data only (Rule 1), our method cannot be applied to the rhythms which are based on pitch and dynamic accentuation. Since the chosen rhythm contains two types of durations, by virtue of Rule 2 it is an appropriate object for our analysis. Let us trace the procedure of structurizing this rhythm step by step.

1. Consider Duration 0. The following one is shorter, consequently, by virtue of Rule 3b it is strongly accentuated. Since it is the first event in the sequence, we recognize the first syllable S as constituted by Duration 0 only. To write down the syllables, we shall use the denotations from Section 6.6, with the only difference that a digit will correspond not to

Figure 6.9: A rhythm with the indicator of complexity 2

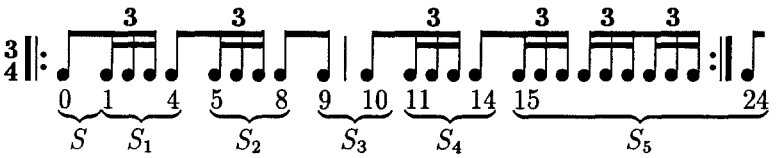


Figure 6.10: Determination of time by recognizing rhythmic syllables

the duration of sixteenth but to the duration of sixteenth triplet. Thus,

$$S = \{100\}, \text{ corresponding to } \text{♩}.$$

Thus up to the current moment our rhythm is represented by the only syllable

$$S.$$

2. Consider Duration 1. It is preceded by a longer duration and succeeded by an equal one. By Rule 3 it is not accentuated. By Rule 4 we don't recognize the end of a syllable at Duration 1.

   Since Durations 2 and 3 are not preceded or succeeded by shorter ones, by virtue of Rule 3 they are not accentuated. Since they are not accentuated, by Rule 4 we don't recognize the end of syllable at these durations.

3. Since Duration 4 is between two shorter durations, by virtue of Rule 3a it is strongly accentuated. By Rule 4 we recognize the end of syllable which we denote

$$S_1 = \{111\ 100\}, \text{ corresponding to } \overset{3}{\text{♬}}.$$

   Now we compare syllable S_1 with the earlier recognized, verifying:

   (a) whether the given syllable is the elaboration of another one;

   (b) whether any other syllable is the elaboration of the given one;

   (c) whether the given syllable is the junction of other syllables;

   (d) whether any other syllable is the junction of the given syllable with another one.

One can see that syllable S_1 is not the elaboration of any other syllable, no syllable is the elaboration of S_1, but S_1 is the junction of two syllables S. Therefore, up to the current moment our rhythm is represented as

$$S \;\; S_1$$

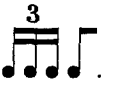
or

$$S \;\; E(2S),$$

where $E(2S) = E(S + S)$ denotes the elaboration of the sum $S + S$ (i.e. the junction of two syllables S).

4. Similarly to Item 2, there is no accentuation at Durations 5–7 and we don't recognize the end of syllable.

5. Similarly to Duration 4 analyzed in Item 3, there is an accentuation at Duration 8, with the only difference that Duration 8 is *weakly* accentuated. By Rule 4 we recognize the end of syllable which we denote

$S_2 = \{111\ 100\}$, corresponding to 

.

Note that syllable S_2 is equal to S_1. Consequently, everything said about syllable S_1 relates also to S_2. Therefore, up to the current moment our rhythm can be represented in the following two ways

$$\begin{array}{llll} S & E(2S) & E(2S); \\ S & S_1 & S_1. \end{array}$$

6. One can see that Duration 9 is not accentuated, and therefore no syllable ends at Duration 9.

7. By Rule 3b Duration 10 is accentuated, and we recognize syllable

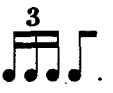
$S_3 = \{100\ 100\}$, corresponding to 
.

Answering the questions (a)–(d) enumerated in Item 3, we recognize that S_1 and S_2 are the elaborations of S_3; besides, S_3 is the sum of two syllables S. Thus we obtain the following equivalent representations of our rhythm:

$$\begin{array}{llll} S & E(2S) & E(2S) & 2S; \\ S & E(S_3) & E(S_3) & S_3. \end{array}$$

8. Since Durations 11–13 are not accentuated, no syllable ends at these durations.

9. Since by Rule 3a Duration 14 is accentuated, we recognize syllable

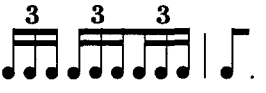
$$S_4 = \{111\ 100\},\ \text{corresponding to}$$ 

Having answered the questions (a)–(d) enumerated in Item 3, we obtain the following representations of the rhythm:

$$\begin{array}{cccccc} S & E(2S) & E(2S) & 2S & E(2S); \\ S & E(S_3) & E(S_3) & S_3 & E(S_3). \end{array}$$

10. Since Durations 15–23 are not accentuated, no syllable ends at these durations.

11. By virtue of Rule 3a Duration 24 (or Duration 0, taking into account the repeat sign) is accentuated. Consequently, we recognize syllable

$$S_5 = \{111\ 111\ 111\},\ \text{corresponding to}$$ 

Having answered the questions (a)–(d) enumerated in Item 3, we obtain that

$$S_5 = E(S_1 + S_3) = E(S_2 + S_3) = E(S_3 + S_3).$$

Hence, we get the following two representations of our rhythm:

$$S\ \|\!:\ E(2S)\ \ E(2S)\ \ 2S\ \ E(2S)\ \ E(4S)\ \ :\!\|\,;$$

$$S\ \|\!:\ E(S_3)\ \ E(S_3)\ \ S_3\ \ E(S_3)\ \ E(2S_3)\ \ :\!\|\,,$$

or

$$S\ \|\!:\ S_1\ \ S_1\ \ S_3\ \ S_1\ \ E(S_1 + S_3)\ \ :\!\|\,. \tag{6.1}$$

If we consider strong accents only, ignoring weak accents, then syllables S_2 and S_3 join into syllable

$$S_{2+3} = \{111\ 100\ 100\ 100\},\ \text{corresponding to}$$ 

Since rhythmic syllable S_5 is the junction of syllables S_2 and S_3, we obtain even more simple representation of the rhythm as follows

$$S\ \|\!:\ S_1\ \ S_{2+3}\ \ S_1\ \ E(S_{2+3})\ \ :\!\|\,. \tag{6.2}$$

With regard to the repetitions of the given rhythm, syllable S can be interpreted as the end of syllable S_5. Finally, we obtain the representation of the

given rhythm as generated by phrase S_1, S_{2+3}. Since S_{2+3} is two times longer than S_1, by virtue of Rule 7 we interpret our rhythm as having triple time: 3/4, or 3/8, etc. The choice of denominator (unit of counting) is a question of convention.

Note that there is a risk to interpret the period in (6.1) as consisting of three equal groups, i.e. instead of "correct" segmentation

$$S \ \|: \ [S_1 \ \ S_1 \ \ S_3] \ \ [S_1 \ \ E(S_1 + S_3)] \ \ :\|\,,$$

on can accept the "wrong" segmentation

$$S \ \|: \ [S_1 \ \ S_1] \ \ [S_3 \ \ S_1] \ \ [E(S_1 + S_3)] \ \ :\|\,.$$

This corresponds to recognizing the time of the rhythm as 2/4. However, the representation (6.2) which is obtained by ignoring local accents leaves no doubts in the triple time basis. Thus distinguishing between strong and weak accents is rather useful.

6.9 Summary of Rhythm Perception Modeling

Summing up what has been said, let us enumerate the main items of the chapter.

1. The proposed approach to rhythm recognition is based on some general principles (correlativity of perception, optimal data representation), on some heuristic methods (the way of coding the tempo curve and estimating the complexity of rhythm), and on some particular properties of hearing (priority of certain durations in rhythm and tempo perception).

2. The interaction of rhythm and tempo is understood as follows. Rhythmic patterns are considered as reference time units for tempo tracking. The interdependence of rhythm and tempo is overcome by the least complex data representation, implying the juxtaposition of rhythm and tempo in an "optimal" way. Since the model is destined for recognizing any forms of repetitions, it is applicable both to divisible and additive rhythms.

3. In order to realize a directional search for generative rhythmic patterns, we suggest formal rules of accentuation and segmentation. Accents are associated with longer durations which determine a segmentation of time events into rhythmic syllables. Then elaboration, sum, and junction of rhythmic syllables are defined. Using this kind of rhythmic grammar, we represent a series of time events in terms of generative syllables (phrases) and their transformations. The method is illustrated with an example of time determination of a given rhythm.