Chapter 1

Introduction

1.1 Formulation of the Problem

By analogy with speech recognition, by *music recognition* we understand studies in computer simulation of music perception which contribute to developing systems for automatic notation of performed music. The performance of a system for automatic notation is supposed to be analogous to that of speech recognition systems: Acoustical data at the input and music score printing at the output.

Music recognition is needed in:

- *Music education* (computer listening to pupils and judging their skills, automatic accompaniment);
- *composition* (automatic music printing);
- *computer-aided performance* (simulating ensemble interaction of the computer and the soloist);
- *music publishing* (acoustical input of musical data);
- *musicology* (automatically transcribing and analyzing live performance, in particular, folk music; facilities for music data input for music data bases);
- *recording engeneering* (visualizing recording and tape editing).

Music recognition systems are also desirable for music lovers and amateur bands who would like to imitate their favorite performers. Music recognition devices could meet the demand for precise transcriptions of musical pieces which are not available from music publishers.

A complete system of music recognition consists of three interfaced subsystems for:

- *Acoustical recognition,*
- *music analysis,* and
- *music printing.*

The task for acoustical recognition is to determine the number of simultaneously sounding parts along with their dynamical and timbral specifications, to recognize instruments, to segment the signal into locally constant and transient segments, to determine periodicity, and to derive pitch trajectories for all voices. The output of this stage of analysis should resemble MIDI messages.

Music analysis includes recognition of time, tempo, tonality, note values and their relative durations, techniques of execution, dynamics, and other musical characteristics which are fixed by notation. At the given stage of data processing the acoustical information is interpreted semantically and represented symbolically.

Music printing, i.e. printing of musical scores, is rather a technical problem which has been solved already. A detailed survey of systems for music printing can be found in (Hewlett & Selfridge-Field 1990; 1991).

Most of the problems arising in music recognition have been considered in different disciplines. For example, pitch and timbre recognition are studied in psychoacoustics, more technical items as signal segmentation are developed in speech recognition, tonality determination is discussed in quantitative musicology, etc. However, there are two specific items inherent in music recognition:

- *Processing compound acoustical signals*, and
- *interpreting acoustical events in terms of elements of musical composition.*

Various particular problems arise both in audio data processing and symbolic coding of semantical information which is carried by the musical signal.

Thus developing a system for automatic notation of performed music requires solving new problems and implementing the models developed in other disciplines. On the other hand, the importance of music recognition for music and musicians is difficult to overestimate. It is similar to that of speech recognition for computer users and computer business. All of this makes music recognition to be a great challenge for computer scientists.

1.2 Brief Survey of Music Recognition

The seek for automatic recognition of music goes back to Bartok and Sieger who have realized its necessity in ethnomusicology studies. The latter have

even used a device called "melograph" which could draw the fundamental frequency curve of a melody (Schloss 1985).

The works in computer recognition of music date back to early 70ies. The first experiments on automatic notation of monophonic music are realized by Sundberg & Tjernlund (1970) and Askenfelt (1976). According to Piszczalsky & Galler (1977) and Roads (1980), Ashton (1971) and Knowlton (1971; 1972) developed a program for transcribing polyphonic music performed on a computer-wired keyboard, having prepared grounds for some later commercially available devices (Roads 1982; 1987; Wyse Carl Disher & Labriola 1985).

The first integral computer system for music recognition was developed by Moorer (1975; 1977) and was, notably, oriented towards polyphony, although its capabilities were severely limited. The music which was to be analyzed was assumed to have two parts, only pitched sounds were admitted (bell-like, or percussive voices were prohibited), certain pitch combinations were avoided (primes, octaves, twelves, etc.), rhythmic structure had to be simple and the tempo had to be constant. In short, only specially selected and specially performed music could be recognized. Nevertheless, Moorer's system proved the realizability of automatic notation and became the starting point for a series of succeeding works.

Considerable progress was achieved in recognition of monophonic music (Piszczalsky & Galler 1977; Piszczalsky Galler Bossemeyer Hatamian & Looft 1981), especially owing to the application of artificial intelligence methods (Chafe Mont-Reynaud & Rush 1982; Foster Schloss & Rockmore 1982; Mont-Reynaud 1985; Mont-Reynaud & Goldstein 1985). It was proposed to use the knowledge about music of a particular style in order to describe all possible combinations of musical elements, among which the recognition system had to choose that combination which most closely resembled the input data. Some comments follow on the strong and weak aspects of this approach.

The use of artificial intelligence methods made it possible to juxtapose rhythm and pitch information so as to take into account rules of elaboration, and, by comparison to earlier works, to widen the admissible variety of recognizable rhythmic figures and melodic passages. A quite complex melody from the end of Mozart piano sonata K.333 performed without any restrictions was recognized in (Chafe et al. 1982). In this case the peculiarities of musical style were taken into account for correcting errors of acoustical recognition.

However, alongside evident advantages, the methods of artificial intelligence exclude applications of the same model to music of different cultural traditions, since stylistic peculiarities which are not anticipated in the model are usually misinterpreted. In other words, general capabilities of music perception are not adequately reflected in the model which is endowed only with particular knowledge.

Another weak point is the impossibility of adapting the model for running

in real time, since only passages of considerable length can be analyzed, a handicap which results from the use of statistical hypotheses whose reliability is dependent on the amount of data processed. That is quite different from human perception; indeed, for example, rhythm is easily recognizable at the very first measures.

Notable among studies concerned with automatic notation of monophony is Niihara & Inokuchi's (1986) system for recognizing Japanese folk singing. A singing voice, being less stable than an instrumental sound, is much more difficult for acoustical analysis. The orientation towards folk music, of practical interest to ethnomusicologists, also distinguishes this study from purely laboratory experiments. However, the folk tunes used in the recognition experiments are too simple for judging the merits of the system.

By about 1985, ten years after Moorer's pioneering work, papers appeared which were devoted to some special topics in music recognition. The problems considered in these publications can be classified into the following two branches:

- *Chord recognition, or voice separation,* and
- *rhythm/tempo recognition.*

In papers on voice separation and note identification (Schloss 1985; Chafe Jaffe Kashima Mont-Reynaud & Smith 1985; Chafe & Jaffe 1986) account is taken of rules of counterpoint, knowledge about the previous context, specifications of the voices used, and so on. It has been proposed also to separate voices by the detection of the asynchronism in tone onsets. This idea is implemented in a system for recognizing Afro-Cuban percussion music, where the asynchronism of tone onsets is quite clear in the sharp attacks in tone envelopes of percussion sounds (Schloss 1985). More traditional material (an excerpt from Frescobaldi) is considered by Chafe & Jaffe (1986) who analyze a piano performance of a two-part passage with contrasting rhythmic patterns of voices.

A few Japanese papers report on the recognition of polyphony by means of advanced signal processing techniques (Katayose Kato Imai & Inokuchi 1989; Katayose & Inokuchi 1989a, 1990). However, the more detailed article (Katayose & Inokuchi 1989b) shows that the system is not yet sufficiently reliable. In a simple chord accompaniment to a melody, for almost every recognized chord there are either missed or false notes.

An interesting idea about the analysis of polyphony stems from the observation that different sounds even of the same instrument show individual characteristics of fluctuations in amplitude and pitch (Vercoe & Cumming 1988; Ellis & Vercoe 1991; Mont-Reynaud & Mellinger 1989; Mellinger & Mont-Reynaud 1991). These authors propose to track such fluctuations in order to separate the spectrum of a multivoice signal into classes of partials united by common

law of motion. The asynchronism of tone onsets and differential development of simultaneous tones afford a basis for voice separation with respect to their dissimilarity. Recently, Mont-Reynaud & Gresset (1990) proposed the use of image processing models for the analysis of multivoice spectra versus time.

Among the studies on rhythm recognition some are of special interest.

A "strategic" approach to rhythm recognition is proposed by Longuet-Higgins (1976; 1987) and Longuet-Higgins & Lee (1982; 1984). Rhythm recognition is understood as finding a strategy in "listening" to music: A hypothesis concerning the rhythmic structure is formulated by the very first events, then it is continuously confronted with the current data, being modified if necessary, and finally a hierarchical structure of the rhythm is developed. A further elaboration of this idea is reported by Povel & Essens (1985) and Desain (1992). The former of the cited works deals with modeling an internal clock which is activated by temporal patterns. The latter is based on the notion of expectancy applied to rhythm perception.

A noteworthy approach to simulation of rhythm perception is proposed by Bamberger (1980) and Rosenthal (1988; 1989; 1992). The rhythmic structure of a melody is separated into simplest patterns which are arranged into repeating segments in order to give the whole structure a certain symmetry. A point of interest is that rhythm is understood as a means of data organization.

An advance in rhythm recognition in application of neuron nets was made by Desain & Honing (1989) and Desain Honing & de Rijk (1989). They suggest that each time interval between tone onsets should be put into correspondence with a neuron whose activation level is proportional to the given time interval. By transferring activation to each other, neurons filter the rhythm and represent it without irregular deviations. The result is a stable state of the net with simple ratios of activation levels of adjacent neurons. This stable state of the net is interpreted as the filtered rhythm. The novel aspect of the model lies in the understanding of rhythm as a simplified conceptual description of a sequence of time relationships. A similar approach to rhythm is discussed by Clarke (1987).

Finally, we mention the development by Dannenberg & Mont-Reynaud (1987) of a system which automatically accompanies a soloist and provides tempo tracking in real time. Their program follows a jazz improvisation, taking into account both time and pitch relationships. A further development is reported by Allen & Dannenberg (1990) and Dannenberg & Bookstein (1991). In spite of restrictions and limited reliability, the program is remarkable as an attempt of rhythm recognition in real time from current data.

The studies cited show that there has been considerable progress in music recognition as well as in understanding its main difficulties. It is recognized that the direct technical approach is nearly exhausted and that new musicological theories of music perception are needed (Richard 1990; Widmer 1990, 1992).

Summing up what has been said, we conclude the following.

1. The problem of music recognition and automatic notation is solvable. However, there are still many difficulties to be overcome.

2. The direct technical approach of the first developments may be replaced by more refined methods. Recent studies in music recognition exhibit two specific problems, namely voice separation and rhythm/tempo tracking. Some original ideas and sophisticated techniques of signal processing are suggested to the end of solving these two problems.

3. Further progress in music recognition is constrained by the lack of clear understanding of the nature of music perception. In particular, it turns out that there are no explicit definitions of notes, chords, rhythm, and tempo, which complicates their recognition. Therefore, in order to move forward, it is desirable to understand the nature of these concepts and the associated mechanisms of music perception.

1.3 Brief Survey of Artificial Perception

In the previous section we have pointed out some principal music recognition difficulties and have shown that common pattern recognition methods are not sufficient to surmount them. Therefore, we consider the artificial perception approach which has been developed mostly in computer vision and which turns out to be useful in music recognition as well. First of all let us characterize artificial perception in general and make references to principal contributors.

We distinguish the following two stages in pattern recognition:

(a) *pattern segregation,* i.e. grouping data into messages;

(b) *pattern identification,* i.e. matching the segregated messages to known memory patterns.

For example, the first stage corresponds to distinguishing lines, spots, etc. in abstract painting, but their associating with common concepts is the task for the second stage.

Some pattern recognition methods are based on direct identification of known objects in data streams without any intermediate processing. Such direct matching of objects to memory patterns can be quite efficient in certain applications (Freeman 1979).

Instead of matching to memory patterns, some objects can be directly identified by so called *invariant property methods*, i.e. by recognizing some invariant

features of the objects which are common to all of their views (Pitts & McCulloch 1947; Ullman 1990a). For example, biological cells can be recognized with respect to a "compactedness measure", which is defined to be the ratio between the cell's apparent area and its perimeter length squared. Cells that are almost round have a high score with respect to this measure, whereas long and narrow objects have a low score. It is important that such a measure is invariant with respect to rotations, translations, and scaling. Certain coefficients of the Fourier transform, object moments, and topological properties are some other examples of invariant characteristics useful for object recognition.

Another example of direct identification is given by the *alignment approach* (Ullman 1990a). This approach consists of two stages. In order to match an object to patterns stored in the memory, the transformations (rotations, translations, etc.) linking each pattern with the object are determined (this is said to be alignment), and then the memory pattern which is the closest to the object is chosen. It is important that the search of the pattern is performed over the patterns but not over the patterns under various transformations, since the transformations (alignments) are already known.

An example of applying similar ideas to character recognition can be found in (Neisser 1966; Preparata & Shamos 1985). In order to recognize a character, every character stored in the memory is matched to the given one, and the allowed transformation which provides the best correspondence between the memory pattern and the given character is to be determined. At the next stage, the memory pattern which provides the best correspondence is chosen.

However, in most practical cases some special techniques are used to segregate patterns before their identification. For this purpose a preliminary stage of data processing is added. The models used at this stage are said to be models of "not-intelligent", "not informed", "naive", or "early" perception. They are aimed at obtaining aggregate representations of input data by means of some description primitives to the end of object segregation, without their identification however. The identification, being based on confronting input data with some previously accumulated knowledge, belongs, strictly speaking, to the domain of artificial intelligence. In order to distinguish the models of "non-intelligent" perception from that of artificial intelligence, we shall refer to them as to *artificial perception* models.

The necessity of such artificial perception models was recognized by Zucker Rosenfeld & Davis (1975). It became clear that the "segmentation-and-label" problem was ill-defined, because the objects to be segregated depend on task and context (Marr 1982; Witkin & Tenenbaum 1983b).

The purpose of artificial perception models is recognizing structure in images. The importance of this task is explained by Witkin and Tonenbaum (1983b, pp. 482–483) as follows:

> 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 relationships 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.

However, recognizing structure is difficult because of the lack of its explicit definition. Therefore, most authors avoid giving strict definitions of structure, preferring to formulate the problem as finding regularity, symmetry, repetitiveness in the data, or, more generally, *perceptual grouping.*

The basic observation about grouping is that the perceptual system has a tendency to put together elements of the visual field in terms of "belongingness" (Palmer 1983, p. 287). Perceptual grouping was studied by Gestalt psychologists and explained as the capability to group elements with respect to their proximity, similarity in color, size, and orientation, and also continuity, closure, and symmetry (Wertheimer 1923). One of most important factors of grouping is the *"common fate"* of some elements which becomes apparent in dynamics when the elements move simultaneously in the same direction at the same rate.

The recognition of "common fate" is the underlying idea of the computer recognition of *shape from motion.* The related works started in 70ies were based on discovering the configurations of the elements which are united by similar displacements, i.e. on recognizing a configuration from the "common fate" of its elements. One can say that the trajectory of motion is recognized first, and the carrier of this trajectory is recognized as an object. Note that object recognition by motion doesn't require any special knowledge about the object, so that no learning is needed (Longuet-Higgins & Prazdny 1984; Shoham 1988).

This approach was applied to the analysis of two- and three- dimensional

scenes, usually for rigid objects (Marr 1982; Gong & Buxton 1992; Meygret & Thonnat 1990; Navab & Zhang 1992; Thibadeau 1986; Ullman 1979, 1990b). Under certain constraints motion cues were used to recognize non-rigid objects as well (Terzopoulos Witkin & Kass 1988; Witkin Kass & Terzopoulos 1990; Pentland & Horowitz 1991).

A physically similar approach, but based on the analysis of optical flow instead of motion field was applied to the recognition of *shape from shading.* The only difference is that the motion field is a purely geometric concept without any ambiguity—it is the projection into the image of three-dimensional motion vectors, whereas the optical flow is a velocity field in the image transformation; hence it needs an additional constraint in order to uniquely determine the image transformation. The works on recovering shape from shading were started by Horn (1975). The related "variational approach" to optical flow was implemented into a computational model by Horn & Schunk (1981) and Ikeuchi & Schunk (1981) see Horn & Schunk (1993) and Ikeuchi (1993) for a retrospective.

The idea of "common fate" was also applied to modeling *textural grouping* in static images. It was notices that similar textural elements, in a sense united by a "common fate," can be grouped together, providing for a segmentation of the image. It was shown that the textural segmentation occurs as a result of differences in the first-order statistics of local features of textural elements rather than as a result of differences in the global second-order statistics of image points (Beck Prazdny & Rosenfeld 1983).

The texture was also used for recognizing local shape in images. The related problem is usually referred to as recognizing *shape from texture.* Gibson (1950) assumed that a homogeneous plane can be considered as covered with textural elements whose density is constant over the plane, whence the density gradient of the image is directly related, via perspective effects, to surface orientation. For example, the shape of a ball is easily perceived from the changes of the density of its ornamentation elements. Two different computer implementations of this idea were developed by Witkin (1981) and Kanatani & Chou (1989) which were generalized by Blake & Marinos (1990).

Another important problem of artificial perception is *recognition of contours* in images and recognition of *shape from contour.* Commonly contours are recognized by their continuity and by some characteristic local environment (change of color, texture density, etc.). It is noteworthy that contours are supposed to be determined by some kind of repetition of the same local environment. In a sense, this makes contours to be similar to trajectories, with the only difference that trajectories are drawn by some repetitious image element in time, whereas contours are drawn by some repetitious image element in space (Palmer 1983, p. 302). Some examples of contour recognition can be found in (Bolles & Cain 1983; Fischer & Bolles 1983; Lifshitz & Pizer 1990; Zhang & Faugeras 1990).

The approach to recognizing contours by representing them by simple primitives called "codons" is developed by Richards & Hoffmann (1976). These primitives are primarily image-based descriptors which have the power to capture important information about the three-dimensional world.

A similar approach is applied to modeling of a higher level of visual perception, corresponding to object recognition by recognizing their constituent parts (Ulmann 1990a). Similarly to a contour which is supposed to be generated by "codons", an object is supposed to be a combination of some primitives called "geons" like cylinders, boxes, etc. (Marr & Nishihara 1978; Brooks 1981). For example, a face contains the eyes, nose, etc., which can be recognized first in order to recognize a face. The number of primitives is supposed to be small (less than 50) and objects are typically composed of a small number of parts (less than 10) (Ullman 1990a).

The interaction between generic elements can be described by topological means and structural descriptions (Barlow 1972; Barlow Narasimmhan & Rosenfeld 1972; Milner 1974; Minsky & Papert 1988). Such methods use the idea of hierarchy, and this "feature hierarchy" is determined both by the constituent parts and by the way they interact with each other. Unlike other artificial perception techniques, these *object decomposition methods*, dealing with a high level of object description, are based on learning, since the image components and their interaction should satisfy some conditions. The review of recent publications on object description from contours can be found in (Mohan & Nevatia 1992; Ulupinar & Nevatia 1993; Bergevin & Levine 1993; Barrow & Tenenbaum 1993; Kanade 1993).

Among the works on modeling particular properties of visual perception, the recognition of *shape from stereo* should be mentioned. The related models process binocular images, using the geometric laws of perspective in order to recognize shape and three-dimensional motion of objects (Marr & Poggio 1976; Marr 1982; Kanatani 1984; Pridmore Mayhew & Frisby 1990). As in other models of artificial perception, no special knowledge about the objects is needed.

The enumerated models of particular visual functions (recognizing shape from motion, recognizing shape from shading, texture segmentation and recognizing shape from texture, recognizing contours and shape from contours, recognizing shape from stereo) are often used in combinations as modules in integral systems of computer vision (Horn 1986; Aloimonos & Shulman 1989). The complementarity of the perception cues implies a better performance of the integral systems of artificial perception.

It should be said that although most of the enumerated models of visual perception use certain physical constraints, they are based on the the assumption that the perception process is primarily data-driven. Such an assumption goes back to Gibson (1950, 1966, 1979) and even to Gestalt psychologists

(Wertheimer 1923) who have postulated the *preference for simple percepts* as a criterion for the *self-organization of perceptual data.* The simplicity principle was explained in terms of the tendency of self-regulating brain activity towards the minimum energy level consistent with the prevailing stimulation.

Later, the simplicity principle has been formulated in terms of information theory as the tendency towards the most economical description (Hochberger & McAlister 1953; Atteneave 1954; 1982). Leeuwenberg (1971; 1978) developed a coding theory where the the simplicity of a figure was estimated by means of the parametric complexity of the code required to generate it. Recently, the simplicity principle was formalized by using the notion of *data complexity* in the sense of Kolmogorov (1965) which is defined to be the amount of memory storage required for the algorithm of the data generation (Tanguiane 1990; Hoffmann 1992).

The simplicity principle is interconnected with the *hierarchization* in data representation. For example, the description of a dynamical scene in terms of objects and trajectories requires less memory than storing all the information about successive images which constitute the totality of data about the scene. At the same time, constructing such an economical representation of a dynamical scene results in the hierarchization of data description. At the lowest level, one has the initial data (stimuli, pixels). The patterns of the first level are similar configurations of data which are traced in successive images and recognized as moving objects. The relationships between corresponding low-level patterns in successive images determine high-level patterns which are associated with the object trajectories. Thus constructing an economical (simple) representation results in constructing a multi-level hierarchy of patterns, where the patterns of a lower level are carriers of the patterns of the higher level (Tanguiane 1990; Raynaut & Samuel 1992).

A general functional approach to understanding the hierarchical structurization in perception was proposed by Leyton (1986). He considered a multilevel architecture of data representation in terms of generative systems of analyzers and formulated structure postulates with respect to the grouping in the representations obtained. The recognition of structure was based on self-organization of data aimed at its economical description.

Thus the simplicity principle implies several important corollaries.

- First of all, it is a general criterion of quality of data representation.
- Next, the simplicity principle justifies hierarchical data representations as saving memory and provides the cues for finding optimal hierarchies for the data representation, corresponding to perception patterns (Witkin & Tenenbaum 1983a; 1983b).
- Moreover, the simplicity principle can help in overcoming the ambiguity in certain hierarchical representations discussed by Morita Kawashima

& Aoki (1992) and Moses & Ullman (1992).

- Therefore, the recognition problem can be formulated as an optimization problem in constructing data representations (Kass Witkin & Terzopoulos 1988; Friedland & Rosenfeld 1992).

- Finally, since any representation is already a description, finding optimal representations is the first step towards understanding the scene semantics (Rock 1983).

Thus we have briefly reviewed perception models of particular visual functions and models of self-organization of visual data. Perception models in audio data processing are developed mostly for the needs of recognition of speech in a noisy background and for music recognition (Bregman 1990; Darwin 1984; Handel 1989; Moore 1982; Warren 1982).

As in visual perception models, in audio pattern recognition two stages are considered, pattern segregation and pattern identification. For example, the first stage corresponds to distinguishing independent acoustical processes, like sounds from different sources or voices in polyphonic music, and the second stage corresponds to process identification, e.g. source recognition or melody recognition.

Until recently audio objects have been identified directly in audio data flows without any intermediate processing. This approach is quite sufficient for the recognition of speech of a single individual or for recognition of monophonic music.

Invariant property methods are used in speech recognition where phonemes are identified by their invariant characteristics (formants, i.e. typical spectral envelopes of vowels, the presence of high frequencies in certain consonants, etc.). In music recognition, tone patterns are recognized in chord spectra by their invariant harmonic structure (Moorer 1975, 1977; Chafe Jaffe Kashima Mont-Reynaud & Smith 1985; Chafe & Jaffe 1986; Katayose Kato Imai & Inokuchi 1989; Katayose & Inokuchi 1989a–b, 1990).

A kind of alignment approach is frequent in speech recognition where input phonemes and words are confronted to memory patterns. The alignment is also used in rhythm recognition under variable tempo (Chafe et al. 1982). As in vision, an input pattern is linked to all memory patterns by means of admissible transformations, and then the memory pattern which is the closest to the input pattern is chosen.

The need for the segregation of concurrent sounds (McAdams 1989, 1991a; Bregman 1990) poses a problem of modeling perceptual grouping in audio. In particular, the cited authors have considered the "common fate" principle for tracking simultaneous acoustical processes. The idea proposed is similar to that in visual scene analysis: The acoustical data are represented as a sequence of short-time spectral cuts, analogous to cinema frames in vision, and then one

has to find spectral patterns whose partials synchronously develop in time, being united by a "common fate".

The related task in speech recognition is known as a "cocktail-party" problem where the overlapping phrases of different participants must be separated according to invariant features inherent in each voice. The voices are recognized by tracing the continuous development of certain spectral constituents. The corresponding models for voice recognition in polyphonic music are discussed by Vercoe & Cumming (1988); Ellis & Vercoe (1991); Mont-Reynaud & Mellinger (1989); Mellinger & Mont-Reynaud (1991), and with direct reference to visual analogy by Mont-Reynaud & Gresset (1990).

The recognition of audio structure by identity and by similarity is based on the same principles as that in visual scene analysis. Thus the recognition of audio structure by identity and similarity is proposed in studies on modeling rhythm perception (Clarke & Krumhansl 1990; Povel & Essens 1985; Rosenthal 1988, 1989, 1992). The rhythmic structure is recognized by finding similar rhythmic phrases and constructing hierarchical representations of time data based on repetitious segments.

Linking notes into melodies with respect to similarity of spectral patterns of voices is considered by Bregman (1990) and Bregman & McAdams (1979). In particular, in the cited works it is shown that notes are linked into a melody only if these notes are performed by the same voice, otherwise the perception of the melody becomes difficult, since a sequence of tones with different timbres is perceived rather as a timbral rhythm (cf. with timbral melody, *Klangfarbenmelodie,* imagined by A.Schoenberg).

The methods of recognition of audio structure from intensity, loudness, and spectral density are reviewed by McAdams (1993). The related models are based on analysis of audio gradients which are similar to textural gradients proposed by Gibson. Similar models are used in visual processing in recognizing shape from texture.

The approach to audio localization and structurization from stereo is developed by Kendall & Martens (1984); Kendall Martens Freed Ludwig & Karstens (1986); Martens (1987) and Wightman & Kistler (1989). These authors developed models where sound localization and source segregation is performed with respect to interaural delay and spectral shaping introduced by head, pinna (outer ear), shoulders, and upper torso. The idea of recognizing an audio scene from comparing data from two distant sensors is similar to the idea of recognition of visual shape from stereo.

Even from our brief remarks one can conclude that audio and visual perception modeling have many common features. Both visual and audio scene analysis make use of "common fate" principle for modeling perceptual grouping. Scenes are recognized from motion, from local characteristics like intensity, textural, or spectral density, from finding identity and similarity, and

from stereo. However, the computer approach to hearing is developed much less than the computer approach to vision. This point of view is shared by several authors, e.g. see McAdams (1991b).

Summing up what has been said in this section, we conclude the following:

1. The difference between artificial perception and artificial intelligence in pattern recognition is understood as follows. Artificial perception is used for discovering structure in visual and audio images by self-organization of data and segregation of patterns. Artificial intelligence is used for pattern identification by their matching to known concepts. Usually, the identification of already segregated patterns is much simpler than their recognition in data flows; thus artificial perception and artificial intelligence are complementary.

2. The artificial perception models of visual functions are classified into

   (a) recognizing shape from motion,

   (b) recognizing shape from shading,

   (c) texture segmentation and recognizing shape from texture,

   (d) recognizing contours and recognizing shape from contours,

   (e) recognizing shape from stereo.

   The artificial perception models of hearing functions are classified into

   (a) recognizing audio processes from voice motion,

   (b) recognizing audio structure from intensity and spectral density,

   (c) recognizing structure from similarity of audio events,

   (d) recognizing audio scene from stereo.

   These models do not use any particular knowledge on the patterns processed.

3. The "common fate" principle is considered as predominant in perceptual grouping. This principle is used directly in recognition of shape from object motion and in recognition of audio processes from voice motion. The segmentation of visual or audio image with respect to the "common fate" principle reveals the structure in the scenes analyzed. This means that this principle, being data-driven, contributes to the recognition of causality in visual and audio data.

4. The simplicity principle of perceptual grouping used in visual perception models is considered as a general criterion of data self-organization aimed at data reduction. The importance of this criterion is caused by its

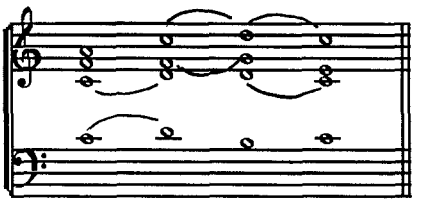

Figure 1.1: Parallel primes, fifths, and octaves prohibited in counterpoint

generality and applicability to any data, independently of their type. In particular, such a criterion justifies the hierarchization principle of data organization, provides a means for chosing an optimal hierarchical representation and for overcoming the ambiguity in data grouping, and enables formulating the pattern recognition problem as an optimization task.

1.4 Development of Correlativity Principle

The development of artificial perception approach to music recognition was stimulated by music studies rather than by computer modeling. The first paper related to the subject was written when the author has studied orchestration with E. Denisov at the Moscow State Conservatory (Tanguiane 1977).

The starting point was a contradiction between some statements of music theory and musical practice. Namely, parallel voices (primes, octaves, and fifths shown in Fig. 1.1) are prohibited in the theory of counterpoint (Aldwell & Schachter 1978), while being widely used in orchestration and pipe organ mixture registers. Recall that a mixture register enables activating several pipes by a single key. Since these pipes are tuned according to a certain chord, playing a melody results in parallel leading of voices of the related chords.

In the available literature on orchestration and musical instruments this inconsistency of theory and practice is not explained. One can find some comments in H. Berlioz' *"Treatise on Orchestration"* (1855), where using mixture registers in pipe organs is severely criticized as "incompatible with the rules of counterpoint and unacceptable for musical ear."

However, according to the author's own experience in playing organ, mixture registers synthesize new timbres rather than provide a polyphonic effect. Moreover, every instrumental voice is always complex, being composed of a series of sinusoidal partial tones. Consequently, playing any musical instrument

results in parallel leading of these partial tones, which is never considered as an undesirable harmonic effect.

Similarly, doubling bass or melody parts at unison or octave in orchestral arrangements makes the given voice brighter, not adding any harmonic quality. For the same purpose, a part can be multiplied at fifths, thirds, and other intervals, as in *Bolero* by M. Ravel (Fig. 1.2).

According to Tanguiane (1977), the prohibition against parallel voices in counterpoint and their use in orchestral arrangements is explained by the fact that voices in counterpoint and orchestration are not the same; to be precise, they have different musical meaning. A part in polyphony is more than simply a physical voice; it is a kind of melodic or harmonic *function.* This implies that a part in polyphony should be independent of other parts and well distinguishable. In orchestration, on the contrary, several instruments can be used to make an effect of a single line, contributing to the same compositional function. Therefore, the rules of counterpoint should be applied not to instrumental voices, but rather to the functional lines, simple or complex, corresponding to single instruments or groups of instruments, respectively, which depends on the context.

Thus there are two different cases in considering parallel voices. The first case arises while using several parallel voices as a single polyphonic part. Then these voices should fuse into one, synthesizing a new timbre, and for this purpose the use of parallel voices is acceptable. The second case arises when different parts are associated with different compositional functions. Then parallelisms should be avoided, since they result in the ambiguity in distinguishing these functions. One can conclude that after the voices have been grouped into functional lines, the rules of counterpoint should be applied to the entire groups (globally), but not inside the groups of voices (locally).

Having tested this conclusion in orchestral arrangements, the author proved that the results have corresponded to the expectations. Consequently, Berlioz' criticism against pipe organ mixture registers can be explained as caused by judging local effects from a global standpoint.

The observation that the perception tends to unite parallel voices into one and distinguishes well non-parallel voices is fundamental for the present study. It is an application of the principle of perceptual grouping of elements with a "common fate" to audio data (Bregman 1990).

At the beginning of the author's research in chord recognition the starting point was formulated as the following conjecture (see Tanguiane 1987; 1988a–b; 1989a–b).

Conjecture 1 (Unseparability of Parallel Voices) *Voices are not separable if they move in parallel with respect to the* $\log_2$*-scaled frequency axis.*

Thus if some partials move in the same direction at the same rate, these

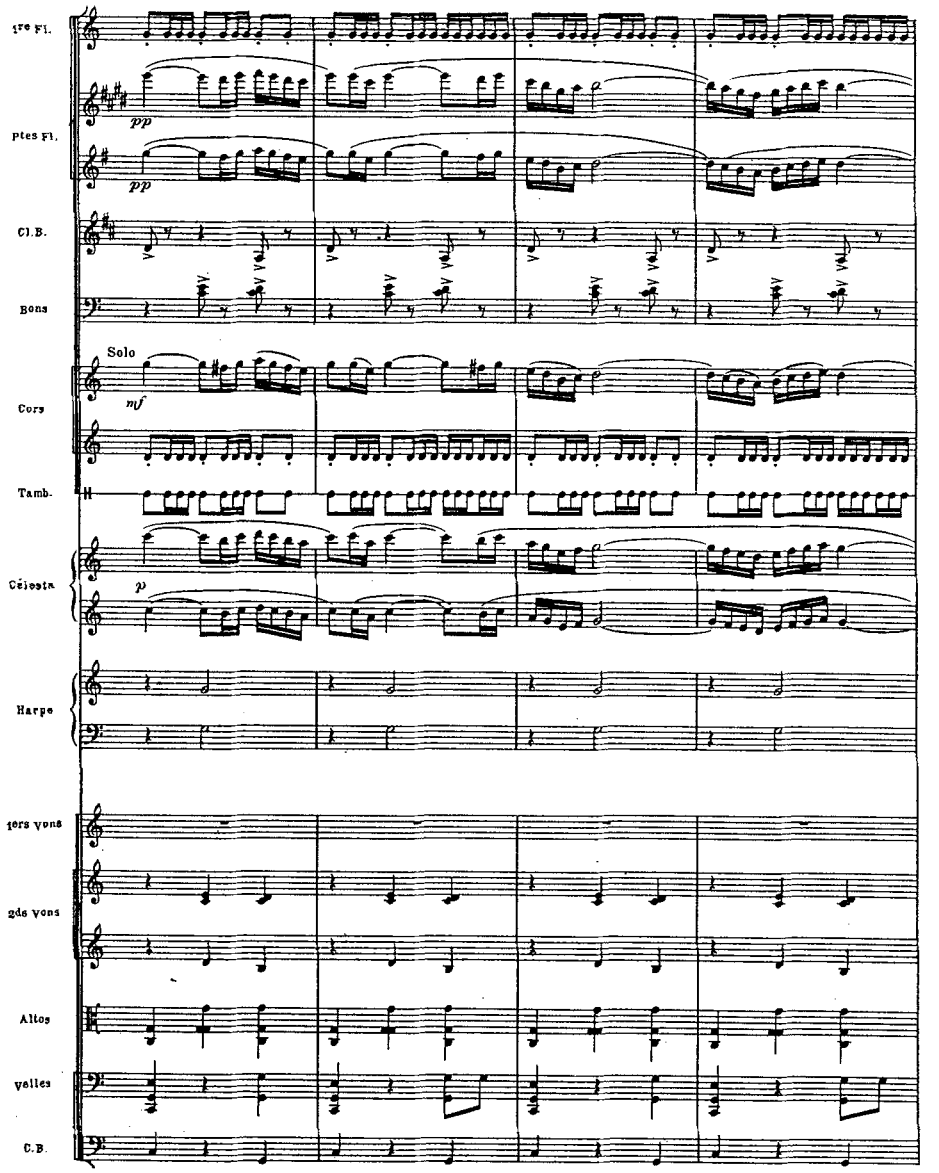

Figure 1.2: The use of parallel voices in *Bolero* by M.Ravel

partials are united by a common law of motion. Conversely, if several groups of partials move differently, these groups can be distinguished by different laws of motion inherent in each group. Thus we obtain the following definition (Tanguiane 1987; 1988a–b).

Conjecture 2 (Part as an Acoustical Trajectory) *A part in polyphony, or a melodic line, is defined to be a dynamical acoustical trajectory drawn by a group of partials which move in parallel along the* $\log_2$*-scaled frequency axis in time. Such a stable group of partials is associated with the voice spectral pattern, or a note pattern.*

Thus recognizing chords is supposed to be realizable in dynamics, by voice tracking with respect to parallel motion of partials. The above definition of a polyphonic part prompts computational means for its recognition, similarly to that for the recognition of objects from motion (Ullman 1979; Marr 1982).

Conjecture 3 (Recognition of Parts by Correlation Analysis of Successive Spectra) *A polyphonic part which is drawn by translations of a voice spectral pattern along the* $\log_2$*-scaled frequency axis in time can be recognized by finding a stable subspectrum in successive short-time spectra of the musical signal. For this purpose one should perform correlation analysis of the successive short-time spectra with* $\log_2$*-scaled frequency axis.*

Next, it was noticed that if the same chord is repeated twice and all its voices have the same spectral pattern, the parts can be recognized not only as sustained, i.e. as determined by repeated notes, but also as linking any pair of notes, as if the voices were crossed. The above observation means that besides correlations between notes of adjacent chords, there should be correlations between notes of the same chord.

Thus note patterns correlate not only in dynamics, but also in statics. Since a melodic line is defined as an acoustical trajectory, a chord corresponds to a statical acoustical contour drawn by a note pattern. Such a similarity of melodic lines and chords is clearly seen when a chord is arpeggiated: Then the chord contour is spread out to a trajectory (Fig. 1.3). Conversely, a trajectory can be compressed into a contour, corresponding to a transformation of an arpeggio (melodic line) into a chord.

From the above reasons it follows that a chord can be defined and recognized in the same way as a melodic line. Thus we obtain the following conjectures (Tanguiane 1987; 1988a–b).

Conjecture 4 (Chord as an Acoustical Contour) *A chord is defined to be a statical acoustical contour drawn by a group of partials which is translated in parallel along the* $\log_2$*-scaled frequency axis. Such a stable group of partials is associated with a note pattern.*

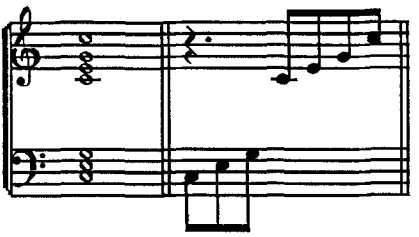

Figure 1.3: Duality of chord contours and melody trajectories

Conjecture 5 (Recognition of Chords by Autocorrelation Analysis of Their Spectra) *A chord which results from translations of a note spectral pattern along* $\log_2$*-scaled frequency axis can be recognized by finding a stable subspectrum in its spectrum. For this purpose one should perform autocorrelation analysis of the short-time spectrum of the chord with* $\log_2$*-scaled frequency axis.*

One can easily see that the duality of contours and trajectories in audio is the same as that in vision (Palmer 1983). On the other hand, the approach to recognizing the structure of chords by finding similar constituents in their spectra meets Witkin & Tenenbaum's (1983b) idea that the visual structure should be based on identity of some elements.

Note that both a polyphonic part and a chord are definied in terms of translations of a voice spectral pattern with no reference to the concept of pitch. Therefore, these definitions are applicable to voices with no pitch salience, like bell-like sounds. Moreover, the separation of voices becomes possible without previous learning, since the task of voice separation is formulated not as recognizing some known subspectra but simply as finding repetitive subspectra.

In fact, we deal with a data representation based on finding *generative elements and their transformations*, or relationships between correlated patterns. Similar representations in vision were considered by Leyton (1986). Such a representation can be justified by the criterion of least complexity by Kolmogorov. Obviously, storing repetitive data as generated by a few elements is more efficient (requires less memory) than storing their totality.

The idea of representing data in terms of generative elements and their transformations was also applied to the problem of rhythm recognition and tempo tracking (Tanguiane 1991b; 1992a–b). The difficulty of the problem is caused by the fact that the tempo is perceived with respect to repeating rhythmic patterns, whereas the rhythmic patterns are recognized as repeated with respect to a certain tempo. It implies the ambiguity in interpreting each duration, since a duration can be identified either as given or as another value

distorted by a tempo change.

In our model the interdependence between rhythm and tempo is overcome by application of the criterion of least complex representation of data. In case of representation of time events, some memory is used for storing generative rhythmic patterns, and some memory is used for describing their transformations, i.e. for storing their elaboration and tempo curve. The total complexity of the representation, i.e. the total amount of memory required, is the sum of these two amounts of memory. Therefore, we look for the representation of time data with minimal total complexity, which is shared between generative rhythmic patterns and a pattern of their transformations associated with the tempo curve.

Our approach to rhythm recognition can be formulated as the following conjecture.

Conjecture 6 (Rhythm Recognition by Optimal Representation of Time Data) *The task of rhythm recognition is formulated as finding the optimal (least complex) representation of time data in terms of generative rhythmic patterns and the pattern of their transformations associated with the tempo curve.*

Note that we use the same approach to three different problems, chord recognition, part recognition, and rhythm recognition. Its principal feature is that instead of finding some known patterns we construct representations of data based on recognizing repetitive messages and optimizing such a representation.

We think that the joint use of the two fundamental principles of perception, "common fate" principle and simplicity principle, is more than simply their union. The interaction of the two grouping mechanisms results in a new quality of perceptual grouping, making it much less ambiguous. At the same time, unambiguous grouping implies stable relationships between blocks of data, preparing ground for the hierarchization in grouping. Let us formulate this as the following conjecture (Tanguiane 1990).

Conjecture 7 (Principle of Correlativity of Perception) *By correlativity of perception we understand its capability to discover similar configurations of stimuli and to form high-level configurations from them. It is equivalent to describing the information in terms of generative elements and their transformations. Among all such representations of the data, the least complex representation must be chosen.*

Correlated blocks of data, said to be low-level patterns, determine high-level patterns which are formed by the relationships between the low-level patterns. This results in a natural hierarchy of patterns where similar patterns of a given level are carriers of the patterns of the next level.

The correlativity in its etymological meaning, "co-relativity," appears in the way patterns are chosen: Firstly, similar to each other (that is correlativity in the lower level); and, secondly, with respect to their interaction in the high-level patterns (that is correlativity between the two levels).

The criterion of least complexity provides the hierarchical arrangement of data with a feedback, guiding the process of data representation in the least complex way. This reduces the ambiguity in data representation, since usually there are alternative representations of the same data with different generative elements and different interactions between them. This way the idea of correlativity of perception is completed by the principle of optimality.

At present we are not capable to prove the hypothesis about the efficiency (optimality) of data representation based on generative elements and their transformations for a general case. Recently, the principle of correlativity was proved in application to perception of chords. Under certain assumptions it was shown that the optimal representation of a chord spectrum is the representation based on generative tones and their translations. Moreover, such a representation is unique, meaning the unique correspondence between optimal representation, chord generation, and perception of the chord (Tanguiane 1992c).

Thus assuming that the perception performs self-organization of data with respect to the criterion of their least complex representation, we explain why chords are perceived not as single complex sounds and not as collections of sinusoidal tones, but rather as sounds composed of sound complexes (notes). The unevidence of this perception phenomenon stems from the fact that the sensation of a chord is not caused by physical matters. Indeed, a sensation of a chord can arise from a sound produced by a single physical body, e.g. a loudspeaker membrane, or a piano board. If physical matters were predominant in sound perception, these sounds would be perceived either as entireties, or as collections of sinusoidal tones, which is not true.

It implies that the problem of chord recognition is a problem of *data representation* rather than that of recognition in a proper sense. Thus we come to the same conclusion as in case of rhythm recognition. In order to solve this recognition/representation problem, we characterize the related data representations and formulate the rules for their construction and quality criteria.

The following conjecture generalizes our approach to music recognition.

Conjecture 8 (Pattern Recognition as Optimal Representation of Data) *The problem of audio pattern recognition can be formulated as the representation of data which is based on their self-organization. The self-organization of data can be aimed at the data reduction and performed by constructing a hierarchy of generative patterns and their transformations. In a sense, analysis of patterns is replaced by synthesis of data representations.*

The above conjecture postulates the primacy of "pure" perception without any special intention to recognize something (cf. the quote from Witkin & Tenenbaum (1983) cited in the previous section). As a result, one obtains a hierarchical representation of data whose elements are treated as patterns (no matter meaningful or meaningless). Only then the patterns are identified (analyzed and matched to memory patterns), learned (stored in the memory for future matching), and labeled.

It is remarkable that representing the data in an optimal way, one can reveal the causality in data generation, implying its semantical interpretation. For example, a chord sound is originally generated by several sources of excitation, which can be recognized even if the chord sound is reproduced through a loudspeaker. We think that the optimality of physical interactions in nature should correspond to the optimality of their description.

Thus we formulate the last conjecture.

Conjecture 9 (Recognizing Physical Causality by Optimal Representation of Data) *Optimal representation of audio data reveals physical causality in the data generation. Thus optimal representation of data is a first step towards understanding their semantics.*

Similar reasons concerning recognizing causality in vision by simple representations are adduced in (Rock 1983, pp. 135, 335). He argues in favor of recognition of a common cause for co-occurring changes than the acceptance of coincidence.

To end this section, we recapitulate its main items.

1. The "common fate" principle is applied to the recognition of polyphonic voices. A polyphonic voice is understood to be a high-level pattern of acoustical trajectory which is drawn by translations of a generative voice spectral pattern in the frequency domain versus time. The recognition of voices is based on the same ideas as the recognition of objects from motion.

2. The "common fate" principle is applied to the recognition of chords. A chord is understood to be a high-level pattern of acoustical contour which is drawn by translations of a generative voice spectral pattern in the frequency domain. The recognition of chords is based on the same principles as the recognition of contours from texture.

3. The "common fate" principle together with the simplicity principle are applied to the recognition of rhythm and tempo. The task is understood as finding an optimal representation of time events in terms of generative rhythmic patterns and their transformations associated with the tempo curve. Using the Kolmogorov criterion of least complex representation

of data enables overcoming the interdependence between rhythm and tempo.

4. The interaction of the "common fate" and simplicity principles is formulated as the principle of correlativity of perception. The correlativity of perception is understood as its capability to discover similar configurations of stimuli (that is the correlativity between patterns of the same level) and to form high-level configurations from them. The ambiguity in data grouping is overcome by the use of Kolmogorov's criterion of least complex representation (this implies the correlativity between the levels of perception).

5. It is supposed that the audio pattern recognition problem can be formulated as the problem of constructing optimal representations of data. Such a representation contributes to revealing the underlying causality in the data analyzed.

1.5 Contribution to Music Recognition

The perception governed by stimulus relationships has been studied in psychology for a long time; for the review see, e.g., Chapter 8 in (Rock 1983). For example, stimulus relationships are predominant in perception of motion and in perception of orientation of objects in the environment.

We argue that stimulus relationships are also predominant in music perception. Indeed, one can see that musical information is transmitted not by sounds but rather by their relationships. In fact, a rhythm cannot be recognized by recognizing time events separately from each other. Another example is the recognizability of melodies in different keys implying the pitch to be less important than interval relationships between the tones.

Hence, we pose two fundamental questions:

- Which relationships and between which events are significant for music perception?

- How and why are they selected from all possible relationships?

Attempting to answer these two questions is our main contribution to music recognition. We suppose that the input data are represented hierarchically in the least complex way, and that such a representation itself reveals certain significant events and significant relationships between them.

Such an approach meets the Gestalt idea that objects are primary psychological representations (Posner 1978, Chapter 7; Posner & Henik 1983, p. 407). In the previous section we have mentioned that the problem of chord recognition as well as that of rhythm and tempo is the problem of data representation

rather than that of recognition in a proper sense. (E.g. a chord sound reproduced through a loudspeaker cannot be identified with several sources other than by its special representation.)

Thus the problem of simulating music perception is formulated in terms of optimal data representation. It turns out that the relationships which are recognized by perception as significant coincide with the correlations of data which are used in constructing optimal data representations.

Generally speaking, subordinating music perception to the simplicity principle seems quite likely. It meets the idea of saving memory and of simplifying further data processing. In our study this *a priori* reason is theoretically and experimentally proved by establishing a correspondence between music perception and optimal representation of musical data.

The grouping with respect to two fundamental psychological principles, "common fate" principle and simplicity principle, constitutes a grouping mechanism which is said to be the correlativity of perception. We show that the two grouping principles control each other, reducing the ambiguity and resulting in a new quality of grouping. The similarity between the properties of music perception and that of the model based on the correlativity principle can be hardly regarded as simply a coincidence. It makes an impression that some general mechanism of music perception is discovered.

1.6 Summary of the Book

The presentation of the material is reversed with regard to the chronology of its development. The general concepts obtained recently are introduced first, whereas initial observations are given as applications. This is done for the same methodological reasons as in mathematics, where certain fundamental statements are formulated as axioms which result actually from successive generalizations.

In Chapter 2, "Correlativity of Perception", we formulate the principle of correlativity of perception and its mathematical model. The problem of recognizing generative elements and their transformations is formulated as finding correlated messages under various distortions of the scale. In order to perform a directional search for the scale distortions which provide high correlation of messages, a method of variable resolution is proposed.

In Chapter 3, "Substantiating the Model", we prove a series of mathematical propositions on representability of chord spectra as generated by spectral patterns of tones. It is shown that the representation of a discrete power spectrum of a chord as translations of a tone spectrum, which corresponds to the causality in the chord spectrum generation, is the least complex representation of spectral data. The demonstration is based on the deconvolution of chord spectra into the convolution product of irreducible spectra, similarly to the

factorization of integers into primes.

In Chapter 4, "Implementing the Model", we show that instead of discrete power spectra it is reasonable to consider discrete Boolean spectra of chords (strings of 0 and 1). Such a simplification has two practical advantages: Computer processing is much more rapid, and Boolean spectra are much more stable with respect to distortions of spectral data, making the model better suitable for the analysis of real acoustical signal. In this chapter we formulate a theorem on the necessary condition for generative patterns in Boolean spectra of chords, and develop an algorithm for their finding.

In Chapter 5, "Experiments on Chord Recognition", the results of computer modeling are outlined. The algorithm proposed is tested and investigated on a series of experiments with synthesized spectra of chords. Possible recognition mistakes and performance of the algorithm are analyzed. The model shows not worse than 98%-reliability in recognizing four-part J.S. Bach polyphony for both harmonic and inharmonic synthesized voices. Besides, it is shown that the limits of the model's recognition capability are similar to that of trained musicians. If chords and voice spectral patterns are simple, they are recognizable by computer and by man. While chords and voices are getting more complex, man and machine fail to correctly recognize them almost simultaneously.

In Chapter 6, "Applications to Rhythm Recognition", the problem of tempo tracking and rhythm recognition is regarded from the standpoint of the principle of correlativity of perception. Repetitious rhythmic patterns (low-level patterns) are considered as carriers of time relationships which determine the perception of tempo (high-level pattern). In other words, rhythmic patterns are understood to be recognizable reference units for tempo tracking. The complexity of data representation is shared between the rhythmic patterns and tempo curve. The problem of rhythm/tempo recognition is formulated as finding the optimal (least complex) total representation of time data. In addition, the definitions of tempo, rhythm, and time are refined.

In Chapter 7, "Applications to Music Theory", we suggest an explanation of some properties of audio perception. In particular, we show that the logarithmic scale in pitch perception and the insensitivity of the ear to the phase of signal result in perceiving musical tones as entire sound objects but not as sound complexes. Besides, we propose a strict definition of interval with no reference to the pitch of tones and explain the function of interval hearing. We show that the properties of the auditory system mentioned, logarithmic pitch scale, insensitivity to the phase of signal, and interval hearing, provide the capability to track simultaneous acoustical processes and to recognize the causality in sound generation, which is necessary for the orientation in acoustical environment. In addition, some rules of music theory are justified as providing the conditions for adequate perception of music.

In Chapter 8, "General Discussion", some remarks on the further develop-

ment of the artificial perception approach are made. In particular, we adduce reasons in favor of applying the artificial perception approach to arranging artificial intelligence data bases for knowledge representation. We suppose that optimal multi-level representation of knowledge, similar to optimal self-organization of sensory data which is studied in this essay, can find applications in computer vision and even in simulation of abstract thinking.

In "Conclusions" we recapitulate the main statements of the book.