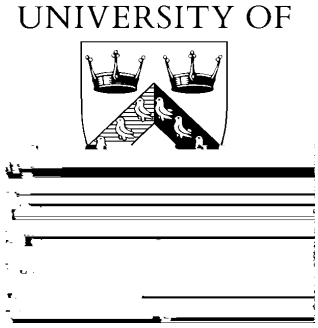
A Connectionist Approach in Music Perception

Otávio Augusto Salgado Carpinteiro

CSRP 426

July 1996

ISSN 1350-3162



Cognitive Science
Research Papers

Contents

1 Introduction **1**

1.1 Introduction to thematic recognition in polyphony 1

1.2 Introduction to musical segmentation and thematic reinforcement 1

1.3 Aim of the research 2

1.4 Structure of the dissertation 3

2 Cognitive aspects of music understanding **4**

2.1 Introduction 4

2.2 Contour and interval representations 4

2.2.1 White’8(e)5.64422(a)5 5e8-1187.67(W)1.53396(h)-4.11026(i)8C08. .e i. .es. .4. . . . . 2.-5 8.244.4.23289(n)-4.187.67(W)1.53396(h)-4.1

6.3 Further work 101

Bibliography **103**

List of Figures

1.1 The connectionist model 2

2.1 Proximity (visual analogy) 11
2.2 Three cases of rhythmic segmentation 12
2.3 Similarity (visual analogy) 12
2.4 First bar of the first two-part invention in C major 16
2.5 First bar of the fifth fugue in D major 17
2.6 Second bar of the fifth fugue in D major 17
2.7 A stretto from the eighth fugue in D sharp minor 19
2.8 A stretto from the eighth fugue in D sharp minor (visual analogy) 19

3.1 Gjerdingen's proposed model 24
3.2 Sejnowski and Rosenberg's model 25
3.3 Backpropagation in time 26
3.4 Backpropagation in time unfolded 27
3.5 Mozer's model 28
3.6 Elman's model 28
3.7 Kangas' model 29
3.8 Chappell and Taylor's model 30
3.9 James and Miikkulainen's model 31

4.1 A musical sequence lasting nine TICs 34
4.2 The model 34
4.3 Representation for the musical sequence in figure 4.1 34
4.4 Training on the first set (first experiment) 38
4.5 Testing on the second set (first experiment) 38
4.6 Testing on the third set (first experiment) 39
4.7 Two first PCs for the patterns in the fourth set (first experiment) 40
4.8 Two first PCs for the negative patterns correctly classified (first experiment) . . . 41
4.9 Two first PCs for the positive patterns correctly classified (first experiment) . . . 41
4.10 Hinton's diagram for three patterns (first experiment) 42
4.11 Training on the first set (second experiment) 45
4.12 Testing on the second set (second experiment) 46
4.13 Testing on the third set (second experiment) 46
4.14 Two first PCs for the patterns in the fourth set (second ex

5.1 An univoiced musical sequence 59
5.2 A multivoiced musical sequence 60
5.3 The model 61

List of Tables

4.1 Generative templates of the pattern sets (first experiment) . . . 36
4.2 Number of negative and positive patterns produced after the number of free slots (first experiment) . . . 37
4.3 Results of the first experiment . . . 43
4.4 Generative templates of the pattern sets (second experiment) . . . 44
4.5 Number of negative and positive patterns produced after the number of free slots (second experiment) . . . 45
4.6 Results of the second experiment . . . 49
4.7 Generative templates of the pattern sets (third experiment) . . . 50
4.8 Number of negative and positive patterns produced after the number of free slots (third experiment) . . . 51
4.9 Results of the third experiment . . . 56

5.1 Representation for a binary sequence . . . 58
5.2 Representation for the musical sequence in figure 5.1 (case I) . . . 59
5.3 Representation for the musical sequence in figure 5.1 (case II) . . . 59
5.4 Representation for the musical sequence in figure 5.2 . . . 60
5.5 Results for model I (first experiment) . . . 64
5.6 Results for model II (first experiment) . . . 64
5.7 Context representation for two binary sequences . . . 64
5.8 Classifications of model I and II (second experiment) . . . 79
5.9 Misclassifications of model I and II (second experiment) . . . 79
5.10 Parameter values of the studies . . . 82
5.11 Classifications of model II (third experiment) . . . 92
5.12 Misclassifications of model II (third experiment) . . . 92

Acknowledgements

Abstract

Little research has been carried out in order to understand the mechanisms underlying the perception of polyphonic music. Perception of polyphonic music involves thematic recognition, that is, recognition of instances of theme through polyphonic voices, whether they appear unaccompanied, transposed, altered or not. There are many questions still open to debate concerning thematic recognition in the polyphonic domain. One of them, in particular, is the question of whether or not cognitive mechanisms of segmentation and thematic reinforcement facilitate thematic recognition in polyphonic music.

This dissertation proposes a connectionist model to investigate the role of segmentation and thematic reinforcement in thematic recognition in polyphonic music. The model comprises two stages. The first stage consists of a supervised artificial neural model to segment musical pieces in accordance with three cases of rhythmic segmentation. The supervised model is trained and tested on sets of contrived patterns, and successfully applied to six musical pieces from J. S. Bach. The second stage consists of an original unsupervised artificial neural model to perform thematic recognition. The unsupervised model is trained and assessed on a four-part fugue from J. S. Bach.

Chapter 1

Introduction

1.1 Introduction to thematic recognition in polyphony

Let us consider the following paragraph:

> The beaker is blue. However, the books are on the table. Despite that, France is in Europe. Thus, he dropped his pen.

Although the words are correctly arranged in the phrases above to convey meaning, one can immediately perceive a complete lack of coherence between each sentence and the whole paragraph. Lower-order structures which govern words placed in phrases are well-constructed. However, higher-order structures which respond to relations between phrases are not.

Current connectionist models which aim to compose music[1] display similar failings (Todd, 1991; Lewis, 1989, 1991; Mozer, 1991; Mozer & Soukup, 1991). Sometimes, small musical fragments, or, even small musical phrases are well constructed. However, musical periods and music as a whole are not. Thus, the compositions produced do not hold any coherence, form, or meaning.

The primary attribute responsible for coherence is musical form. *Musical form*, as Cole (1970, p. 1) briefly defines, is "the structural plan of a musical composition". It describes, in a lesser

segmentation is the limited capacity of human memory. By segmenting the musical piece into small parts, listeners are able to increase the amount of information which can be retained in their memories.

Thematic reinforcement may be performed by listeners as well, through memory mechanisms in the brain. By memorizing themes of musical pieces, listeners would be able to identify their instances whenever they occur throughout the pieces. Alternatively, reinforcement may be performed by performers. In this case, performers would play notes corresponding to instances of theme louder than other notes.

1.3 Aim of the research

The major aim of the current research is to develop a connectionist model to investigate, along with other related issues, the role of cognitive mechanisms of segmentation and thematic reinforcement in thematic recognition in polyphonic music. The connectionist model, which is displayed in figure 1.1, comprises two stages.

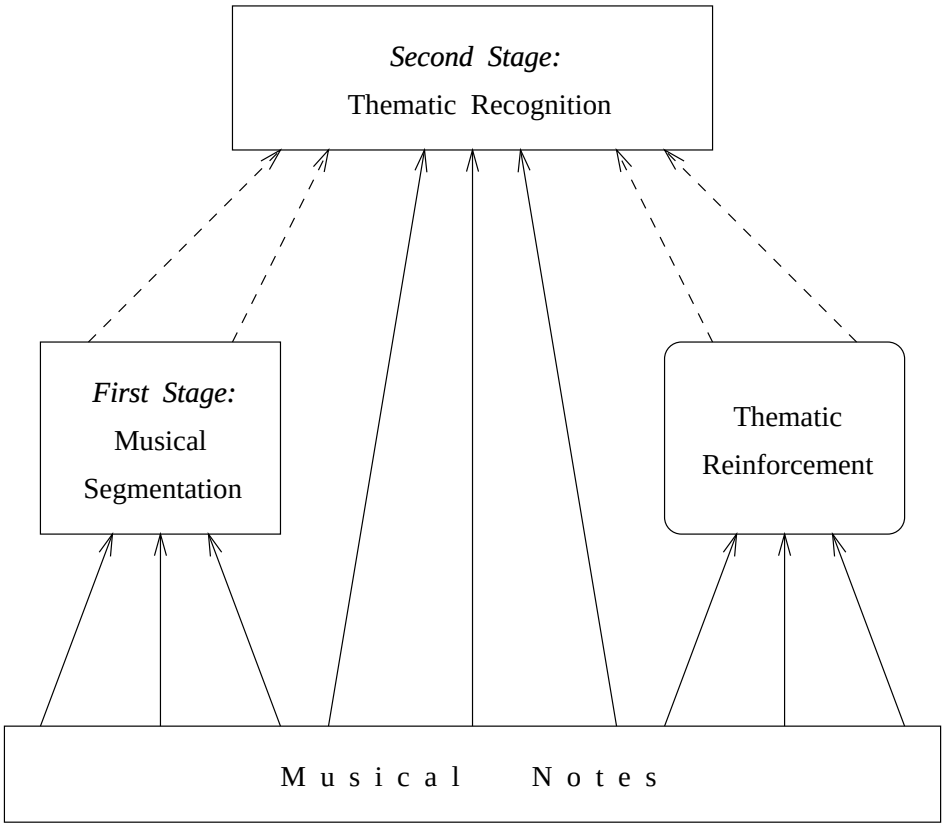


Figure 1.1. The connectionist model

The first stage consists of a supervised artificial neural model to segment musical pieces in accordance with three cases of rhythmic segmentation. The second stage, in its turn, consists of an unsupervised artificial neural model to perform thematic recognition. Thematic reinforcement is not implemented as a stage consisting of an artificial neural model. Instead, it is supplied directly to the input layer of the second stage.

The connections from both the first stage and reinforcement to the second stage, which are represented as dashed arrows in figure 1.1, may be switched on and off. Thus, we may verify how, and to what extent, the mechanisms of segmentation and reinforcement may affect the recognition of instances of themes in polyphony.

1.4 Structure of the dissertation

The dissertation comprises six chapters. The first chapter is this introduction.

The second chapter is concerned with cognitive aspects of music understanding. It offers a review of the literature in contour and interval representations, in musical segmentation, and

Chapter 2

Cognitive aspects of music understanding

2.1 Introduction

This chapter reviews cognitive aspects of music understanding. It is divided into five sections, the first of which is this introduction. The second section is concerned with psychological and neuropsychological studies supporting both contour and interval representations. The third section provides a review of segmentation. In it, along with studies supporting segmentation, we present a brief introduction to Lerdahl and Jackendoff's theory as well as a description of three grouping rules. The fourth section deals with issues relating to thematic recognition in polyphony. Unfortunately, little research has been carried out in this field. Hence, we present existing issues which are open to investigation as well as a succinct introduction to polyphony. Finally, the fifth section provides a summary of the main ideas discussed in the chapter.

2.2.4 Dowling's experiments

Dowling (1971) performed a set of experiments on subjects to

lost in recognition of inversions was replicated and extended to retrogrades and retrograde inversions". Second, "as retrograde inversions were the most difficult to recognize, the pitch-vector characterization of the process by which subjects handle the task seems the more plausible psychological model than does the interval-vector characterization". Nevertheless, as retrograde and inversion transformations were not equally well recognized, Dowling kept reservations about this last conclusion.

Dowling (1978) is an extension of Dowling and Fujitani (1971). While in the latter the comparison melodies in the experiment were all atonal, in the former all but one were tonal. By introducing tonality, Dowling was able to include one more type of comparison melody — the tonal answer. Tonal answer keeps contour and tonal key of the standard melody, yet changes the interval sizes. As the subjects found it extremely difficult to distinguish between exact transpositions and tonal answers, Dowling claims that "even with tonal melodies, contour (in the sense of ups and downs measured in diatonic intervals) and interval sizes (measured in semitones at the level of tonal material) are stored independently [in memory]".

2.2.5 Dowling and Bartlett's experiments

Another important result in tonality was obtained by Dowling and Bartlett (1981). They performed a set of four experiments to investigate the role of contour and interval information in memory for melodies. In the first two experiments, subjects heard a list of excerpts from Beethoven string quartets. Their task was then to discriminate between targets and lures, and between related items

second pause, a comparison melody which possessed one altered note was heard. The subjects' task was then to detect the altered note. In the contour task, subjects were asked to observe the contour of a standard melody. Also, after a five second pause, a comparison melody which possessed a contour alteration was heard. The subjects' task was to detect the contour alteration.

In both tasks, the comparison melody could be heard either transposed or not. Moreover, all melodies were tonal and varied from 5 to 15 notes in length. Thus, Edworthy could check the resistance of contour and interval information to decay in short-term and long-term memory.

Based on the results, she concludes that "contour information is immediately available regardless of novelty, familiarity, transposition, or non-transposition". More, "accurate encoding of contour does not depend upon the listener's ability to establish a key". However, contour information is easily lost when melody increases in length, suggesting that it is related to short-term memory. On the other hand, "when the inherent difficulty of establishing a key is great, when melodies are both novel and transposed, interval information is initially imprecise". Yet, it is encoded immediately and precisely when key becomes established. Furthermore, familiarity with melodies makes interval information very resistant to decay, suggesting thus, that it is related to long-term memory.

2.2.7 Neuropsychological research on contour and interval information

Finally, neuropsychological research has lent support to the dichotomy between contour and interval information as well. Commenting on the subjects' performances in former studies reported in Peretz and Morais (1987) and Peretz and Morais (1988), Peretz and Babaï (1992) claim that the cerebral right hemisphere is involved in tasks which require contour information. On the other hand, the left hemisphere is involved in tasks which require interval information. Moreover, the study of brain-damaged patients has provided consistent evidence to the fact that "interval-based and contour-based approaches of melodies are not only functionally but anatomically distinct". Indeed, Peretz (1990) has shown that a vascular lesion in the left hemisphere affects the ability of representing melodies in terms of their intervals, but not in terms of their contour, whereas a vascular lesion in the right hemisphere affects both abilities. Based on that, Peretz and Babaï (1992) conclude that "melodic contour serves as a necessary anchorage frame for encoding interval information", and also that "a lesion in the right hemisphere is detrimental because it disrupts both the processing subsystem required for representing the melody contour and deprives the intact left hemisphere with Mftqube ms ubng t0(t)-248.029(e)27.5467(v)]TJ260.64 0 Td[(i)-6.931814(f)-324.529(m)10.876

variations in intensity or in frequency enabled subjects to

features are recognized, and then the combination of features is recognized to identify the pattern". Similar processes seem to be involved in recognition of musical patterns (Drake & Palmer, 1993).

Segmentation is necessary for several reasons. First, the human processing system is limited. Peretz and Babaï (1992) claim that "when faced with a stream of rapidly changing ephemeral events, a well-established and useful propensity of the perceiving human organism is to group these events in small chunks in order to increase the amount of information that can be retained by the limited capacity of our processing system". Drake and Palmer (1993) are more specific, and attribute the limitations of the human processing system to the limited capacity of human memory.

A second reason is that segmentation produces grouping. Groups produced by segmentation become musical units. Attneave and Olson (1971) affirmed that "people rarely treat individual tones as auditory units or tie particular behaviors to particular pitches". "A very simple case of an auditory unit with realistic object properties is that of a brief sequence of tones, as in a melodic phrase".

The third reason is to build up mental structures where grouping stands as a basis. Lerdahl and Jackendoff (1983a, p. 13) assert that "when a listener has construed a grouping structure for a piece, he has gone a long way toward 'making sense' of the piece: he knows what the units are, and which units belong together and which do not. This knowledge in turn becomes an important input for his constructing other, more complicated kinds of musical structure. Thus grouping can be viewed as the most basic component of musical understanding".

2.3.4 Lerdahl and Jackendoff's theory

2.3.4.1 Description of the theory

Lerdahl and Jackendoff (Jackendoff & Lerdahl, 1981; Lerdahl & Jackendoff, 1983b, 1983a) have proposed a theory to describe, by means of formal rules, the principles which listeners follow to build up mental structures of heard western tonal music in order to understand it. The theory is reductionist. Therefore, the formal rules describe how listeners parse the musical surface to represent it internally in a hierarchical form.

Reduction of the musical surface is necessary for musical understanding, indeed. Jackendoff (1991) provides a complete explanation in an example. He considers the case where a listener hears a "piece that is not altogether identical to a remembered piece — say a variation on a known theme or a new arrangement of a popular song. Here the musical surfaces may differ considerably; even the metre and mode of the variation may differ from those of the theme. In order to recognize the relation between the heard piece and the remembered one, then, the processor must be comparing not just musical surfaces but the abstract structures of the two pieces, in particular the reductions".

2.3.4.2 Experiments supporting the theory

Lerdahl and Jackendoff's theory has received experimental support. Deliege (1987) performed two experiments to check the validity of the grouping rules of the theory. Musicians and non-musicians were asked to segment musical sequences in accordance with their preferences. Each musical sequence contained an instance of a single rule in the first experiment. In the second, each sequence contained an instance of two conflicting rules.

The results confirmed the validity of the grouping rules. Both musicians and non-musicians segmented properly the sequences according to the rules in most cases, although the performance of the former have been much superior.

Serafine, Glassman, and Overbeeke (1989) carried out a set of experiments to evaluate the importance of hierarchical structure in music. In one of the experiments, subjects heard sets of three musical sequences — a short melody, its reduction, and a foil reduction. Their task was to identify which reduction was the most similar to the melody. As the subjects matched the correct reductions to the melodies, Serafine, Glassman, and Overbeeke conclude that not only did the subjects hear the musical sequences, but also recognized their hierarchical structures.

In the third experiment, the background melody was unfamiliar whereas the foreground was familiar. In addition, subjects were told beforehand which foreground would be employed. The background and foreground melodies interleaved in the same pitch range, and the degree of overlapping was kept constant throughout the experiment. The subjects' task was to say whether they identified the foreground melody. During the trials, Dowling substituted the original foreground by another one. Subjects were not told about this misleading operation. Thus, it was possible to verify whether they were, in fact, recognizing the foreground melody.

In the fourth experiment, the background melody was familiar whilst the foreground was not. Half of the subjects was informed beforehand which background would be employed. The other half was not. The degree of overlapping between the tones of the melodies was varied throughout the experiment. The paradigm used in the second experiment was used here as well. Therefore, the subjects' main task was to say whether the comparisons were identical to the standards or not.

The results of the first, second, and fourth experiments led Dowling to the same conclusion, that is, overlapping interferes with recognition of melodies. Recognition is easier when background and foreground melodies do not overlap. However, as the results from the third experiment indicates, "listeners can overcome the interference effect [of overlapping] and recognize a familiar target melody if the target is prespecified, thereby permitting them to search actively for it".

The results of the third experiments have also implications for perception of polyphonic music. They suggest that "active search for a well-known melody can lead to discerning it in a confusing context when it would go unnoticed by the passive listener". "The listener who knows the typical pattern of recurrences of a theme in a fugue . . . is able to perceive that theme more easily than the listener who does not know what to expect".

contour-judgement task, but to some extent in the pitch-judgement task, diatonic pitches[3] outside the expected range[4] of the melody were difficult to perceive". "A typical [listeners'] observation was that on some trials the critical note seemed simply to disappear. The listeners were sure that something had happened at that point in time, and that whatever pitch had occurred was not within the region they had been attending to".

In the third and fourth experiments, listeners carried out the same tasks as those in the second experiment, yet comparison melodies were not interleaved with distractor notes. There is a sharp contrast between the results from the second experiment and those from the third and fourth. Remote pitches — critical notes outside the expectancy window — were difficult to perceive in the second experiment. Nevertheless, by removing the distractor notes from the comparison melodies

dards. Variations were formed by changing one of the pitches of the standards. The change could occur at any voice, at any serial position, and could keep or not former harmonic relations.

Listeners were firstly instructed to recognize a standard composition. Then, they were presented with comparisons related to that standard. The listeners' task was to indicate whether or not each comparison was identical to the standard they had learnt.

Standards and comparisons were played on a piano during the first experiment. In the second experiment, they were sounded by a pure sine tone generator t

"If two melodies which are to be played or sung together are so written as to be capable of inversion, that is, if either of them may be above or below the other, and the harmony still be correct, we have ***double counterpoint***, a term which simply means invertible counterpoint" (Prout, 1969, p. 1). Double counterpoint is the basis of polyphonic imitation, canon, and fugue, for it establishes the rules which let a theme shift from one to another part.

2.4.5.4 Imitation and canon forms

Imitation is the repetition of a ***melodic figure*** or a ***theme***, whether at the same or at a different pitch, in a different part (Prout, 1969). "Imitation which is maintained continuously, either throughout a whole piece, or at least through an entire phrase, is said to be ***canonic***; and if a composition is so written that the various parts imitate one another throughout, such a piece is called a ***canon***" (Prout, 1969, p. 145).

Many of the two-part and three-part inventions of Bach (Bach, 1970) were composed under imitation and canon forms (Adams, 1982a, 1982b). Inventions, and also fugues of The Well-Tempered Clavier of Bach (Bach, 1989), will be focused here because they took part as a domain in our experiments.

2.4.5.5 Bach's inventions

Bach composed two sets of ***inventions*** (Bach, 1970). The first set contains 15 ***two-part inventions***, and the second, 15 ***three-part inventions***. They are considered to be a pedagogic work, for Bach composed them for his pupils, having in mind their needs in the harpsichord technique (Beard, 1985; Flindell, 1983).

The inventions of Bach are contrapuntal works where a theme is elaborated by its imitation shs

the musical surface to represent it internally in a hierarchical form. Three of the grouping rules — segmentation by rests, by longer durations, and by breaks of similarity — were detailed for they came into the domain of our experiments.

Little research has been carried out in order to understand the mechanisms underlying the perception of polyphonic music. Perception of polyphonic music involves thematic recognition, that is, recognition of instances of theme through the voices, whether they appear unaccompanied, transposed, altered or not. Although they are loosely related, studies in perception of interleaved melodies have provided a few clues for research in perception of themes in polyphonic music.

Important issues concerning thematic recognition in polyphony are still open to investigation. For instan(i)58(i9293404(o93T)-301.192(r)4.23170691(p)-0(n)-4.11026((i)58(i)-6.9307(t)-6.ld[(F)29.958(o)181(

Chapter 3

Review of connectionism

3.1 Introduction

This chapter provides a review of connectionism. Connectionist models, as Kohonen (1988) points out, "are massively parallel interconnected networks of simple (usually adaptive) elements and their hierarchical organizations which are intended to interact with the objects of the real world in the same way as biological nervous systems do".

Connectionist models are very suitable for cognitive domains. As Anderson (1990) writes, they "are regarded in cognitive psychology as displaying considerable promise in finally bridging the gap that has existed between the brain and higher-level cognition". They have also been chosen extensively by researchers as models in cognitive domains because they hold significant properties which parallel those held by humans. Among those properties, we may find pattern completion, approximate matching, good generalization, graceful degradation, robustness[1], context sensitivity, inductive learning, soft constraint satisfaction[2], and good scaling up.

Connectionist models have been widely employed in the musical domain. Among them, we may cite models for pitch perception (Sano & Jenkins, 1991; Taylor & Greenhough, 1994), chord perception (Laden & Keefe, 1991), tonal perception (Scarborough, Miller, & Jones, 1991; Leman, 1991; Bharucha, 1987, 1991; Bharucha & Todd, 1991), musical time perception (Desain & Honing, 1991), perception of musical sequences (Page, 1994), musical pattern categorizations (Gjerdingen, 1990, 1991), and musical composition (Todd, 1991; Lewis, 1989, 1991; Mozer, 1991; Mozer & Soukup, 1991). Three of these models — Laden and Keefe (1991), Leman (1991), and Gjerdingen (1990, 1991) — had particular relevance to my work, and consequently, were reviewed in the second section of the chapter.

The chapter is divided into five sections. The first section is this introduction. The second section reviews the three connectionist models mentioned above. The third and fourth sections review, respectively, four supervised and three unsupervised models of sequence classification in time. The main virtues and limitations of the models are object of special consideration in these sections. Finally, the fifth section summarizes the chapter.

3.2 Connectionist models in musical perception

3.2.1 Laden and Keefe's model

Laden and Keefe (1991) assessed supervised neural nets as models of pitch and chord perception. Three-feedforward-layered neural nets, which were trained with error backpropagation (Rumel4(e)5.6 i tlKer

tone, whereas the output units represented the existing 36 pitch classes in three octaves. Training set consisted of 36 complex tone patterns — one for each of the 36 pitches over three octaves. Training patterns were taken randomly from a pool of patterns. The pool was based on spectra of harmonics of four tones from four musical instruments — clarinet, violin, trumpet, and pipe organ. Each spectrum was used as a template to build 36 tone patterns which spanned a three-octave range[5]. Thus, the pool held a total of 144 patterns.

The net was able to identify 35 out of 36 tone patterns in the training set. It was tested on novel input as well. Novel input consisted of the remaining 108 patterns in the pool, 36 sine tone patterns, 144 missing fundamental patterns, and 19 patterns which reproduced real spectra of 19 tones of the four musical instruments mentioned above. The net generalized very well on novel input. Performance on identifying missing fundamental patterns was better than that on identifying sine tone patterns, thus suggesting that a combination of upper harmonics contributes more than the fundamental to pitch perception.

The overall level of performance of the nets in the experiments, in particular of those which used harmonic complex representation, was high. Thus, Laden and Keefe claimed that by employing harmonic complex representation as input, three-feedforward-layered neural nets trained with error backpropagation can be used to model perceptual phenomena, such as chord and pitch perception.

3.2.2 Leman's model

Leman (1991) performed a set of experiments in tonality. He employed a self-organizing map model (Kohonen, 1989) for studying relations between tones in a tonal context.

As Laden and Keefe (section 3.2.1), Leman also made use of subharmonic complex representation (Terhardt, 1974) in one of his experiments. Twelve input units, which were assigned to the twelve pitch classes in a scale, represented subharmonics of tones of chords. Activation of each input unit was dependent on the number of subharmonics falling in the pitch class assigned to that unit. The map contained 20×20 units, and the training set consisted of 115 different chords, including triads and seventh chords.

Leman could observe that the map self-organized in terms of the circle of fifths. Chords which were tonally related in terms of the circle of fifths were close in the map, whereas tonally non-related chords lay far apart from each other. Moreover, the response regions of tonally related chords overlapped in the map, whereas response regions of tonally non-related chords did not.

Overlapping of response regions could model the phenomenon of perceptual facilitation of chords. A sequence of chords which are tonally related is more easily perceived because units which lie in the intersection of response regions of those chords are frequently activated. The interaction of activations of neurons in overlapping response regions could also explain "how a tonal context can be set and how tones get their particular tonal function with respect to this context".

The model employed in the experiment could explain some important perceptual phenomena in tonality. Leman concludes thus that "aspects of tonality can in principle be accounted for by internal representations that develop through self-organization from invariant features in the musical environment".

3.2.3 Gjerdingen's model

Gjerdingen (1990, 1991) used a self-organizing ART2 net (Carpenter & Grossberg, 1987) to classify musical patterns in six of Mozart's earliest compositions. The training set consisted of 793 separate musical patterns. The input layer held 34 units, whereas the output layer held 25 units. The input units represented 34 musical features, such as pitch classes of the bass, inner and melodic

[5]It is worth mentioning that spectra produced by tones of a musical instrument are not identical, but rather, they vary over the instrument frequency range.

voices, and the contours of bass and melodic voices. Activations of the input units decayed in time to display the order of input. Activations also varied according to the metric, that is, strong and weak beats in measures produced different levels of activation in those units.

By analyzing the 25 weight vectors after training, Gjerdingen verified that the net could acquire memories for small groups of notes present in the compositio

into groups which would be classified and related to each other in a hierarchical way (see section 2.3.3). Gjerdingen acknowledged that the model should be able to classify subsequences within a long sequence as well as to sharply distinguish between sequences which held the same events, yet in different sequential orders.

Later on, nevertheless, Gjerdingen (1992) abandoned the idea, and adopted a masking field (Cohen & Grossberg, 1987) embedded in an ART3 architecture (Carpenter & Grossberg, 1990) to classify temporal patterns of chords. We think that Gjerdingen's idea of employing two self-organizing nets in a hierarchical architecture was promising. We regret that he has not explored it fully.

3.3 Supervised models to classify sequences in time

Several researchers have extended the three-feedforward-layered architecture under backpropaga-

qee ce

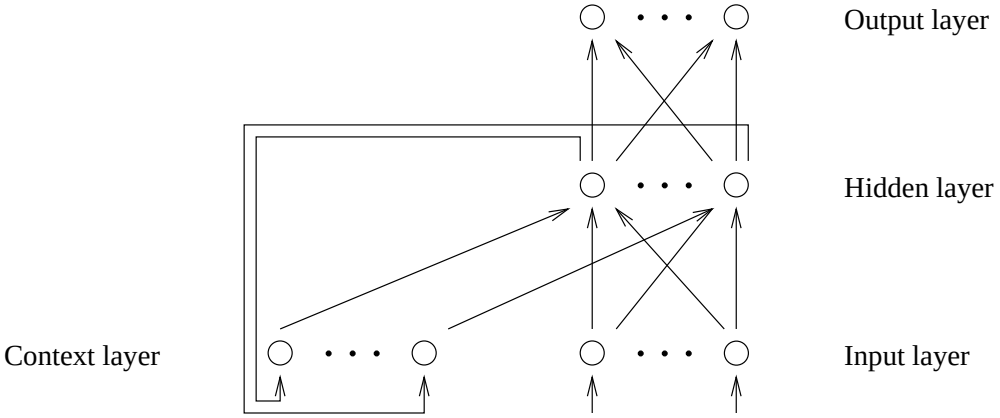


Figure 3.5. Mozer's model

This led Mozer to foresee a plausible drawback, that is, the model might be unable to learn in domains where a rather large memory of past events were required. Besides, as another possible drawback, Mozer reported that activations of context units might become very large. This would be particularly true in generalization, in case of inputting novel sequences which were longer than training sequences[6].

3.3.4 Elman's model

Elman (1990) proposed a model which is another simplification of backpropagation in time (section 3.3.2). It is presented in figure 3.6.

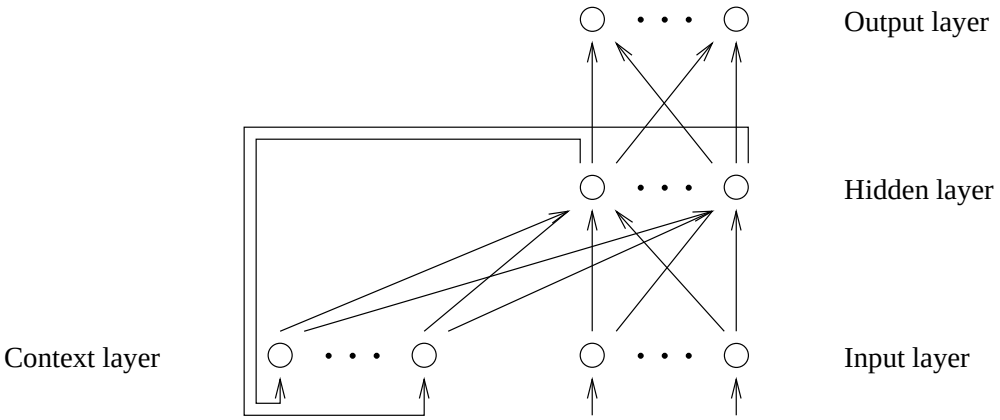


Figure 3.6. Elman's model

The model has an architecture which is identical to that of backpropagation in time. Thus, it consists of a recurrent net in which each hidden unit receives activation from all context units, and sends activation back to its corresponding context unit.

Elman's model simplifies backpropagation in time algorithm by calculating current error signals only, disregarding error signals of previous steps in time. Hence, the values of activations of context units are now considered as independent variables, no longer dependent on weights

[6]Both drawbacks in Mozer's model were confirmed in experiments performed by us. These experiments, however, are not reported in the dissertation.

being adjusted (Fahlman, 1991). Moreover, the model does not require the storage of previous net activations, but rather, of current ones only. Thus, its memory requirements are much lower than those of backpropagation in time.

Elman carried out a set of experiments. In one of them, he trained the model to predict the next pattern in a training sequence. The training sequence consisted of consonants and vowels in a structure where each consonant precisely determined type and number of following vowels. A distributed representation representing features of consonants and vowels was employed to input and output units.

The model had 6 input units, 20 context units, 20 hidden units, and 6 output units, and was trained in 200 epochs. It was tested on the training sequence, and could predict correctly the type and number of vowels which followed the consonants.

Elman's model has a serious drawback. By truncating the backpropagation of error signals at the context layer, the model loses its ability to adjust properly its weights, because it leaves out of account for that adjustment, the information which comes from inputs occurring on previous steps in time. As a result, the model keeps a very short trace of past inputs, and thus, is unable to learn in domains where a rather long memory of past events is demanded[7].

3.4 Unsupervised models to classify sequences in time

The self-organizing feature map model (Kohonen, 1989, 1990) has been extended to classify sequential information. The problem involves either classifying a set of sequences of vectors in time or recognizing sub-sequences inside a large and unique sequence. The most recent approaches are

(section 3.3.1). Thus, its most serious deficiency is that it becomes computationally expensive as wider windows are required. Its main quality is that it is capable of holding a precise memory of past events.

3.4.2 Chappell and Taylor's model

Chappell and Taylor's model (1993) follows the time integral approach[8]. In this type of approach, the activation of a unit is a combination of its current input and its former outputs decayed in time. In their model, the time integrator is applied to the units in the map, as displayed by the recurrent connections in figure 3.8.

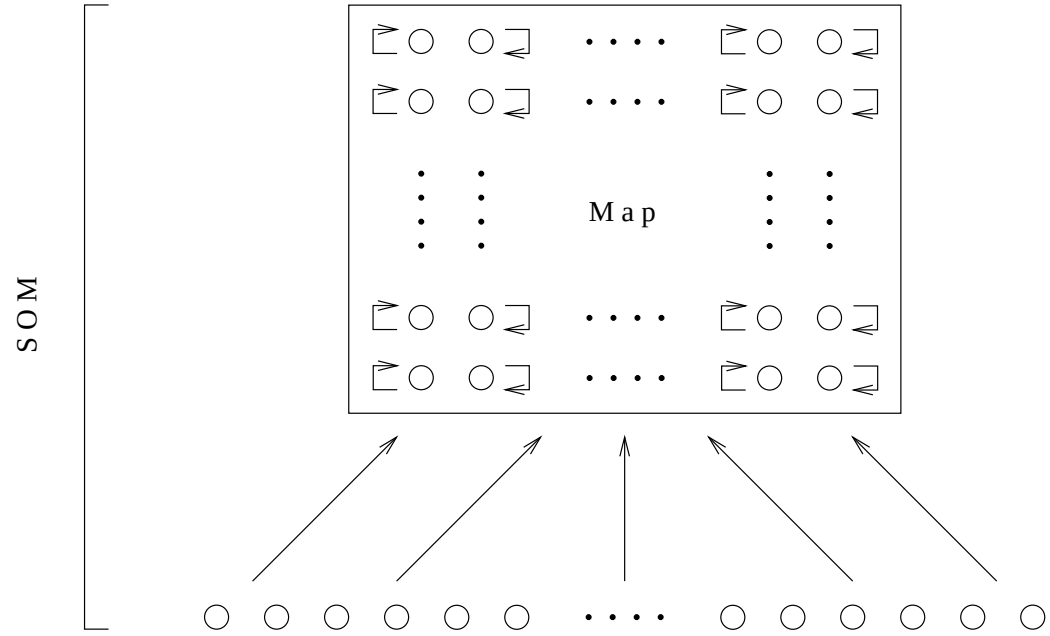


Figure 3.8. Chappell and Taylor's model

Chappell and Taylor performed two experiments. In the first, they used a net configuration with two input units and 4×4 units in the map. The input units represented the numbers 0, 1, 2, and 3 in a binary representation. The training set was made up by 16 sequences of length 2, which consisted of the 16 possible combinations of those four numbers. After 1000 epochs of training, the net was able to distinguish each sequence in the training set, for each unit in the map responded to an unique input sequence.

In the second experiment, Chappell and Taylor employed longer training sequences to verify whether the model had sensitivity to earlier elements in these sequences as well. The training set consisted then of 3 sentences with 5 words each. These words were selected from a pool containing 9 words. The word 'dry' appeared in fourth position in each sentence. A net configuration with 4 input units and 8×8 units in the map was employed. After training, the net was able to classify

3.5 Summary

Ten connectionist models were reviewed in this chapter. The first three — one supervised and two unsupervised — were employed as models of pitch perception, chord perception, tonal perception, and musical pattern categorizations. Although the models have performed well, and have been able to learn important human cognitive tasks, it is worth noticing the fact that they performed on simple cognitive tasks, which are hierarchically situated in cognitive levels lower than those of segmentation and thematic recognition.

Next, we reviewed four supervised models to classify sequences in time. All these employ extensions of the three-feedforward-layered architecture under backpropagation learning algorithm.
f cle twata

musicalsegmentation, which deman

ashigh compatih-6.930914(n)-4..0484(a)5.64422(

Chapter 4

A neural model for musical segmentation

4.1 Introduction

Owing to memory constraints, it is believed that listeners do not grasp a musical piece in its entirety, but on the contrary, they segment it into parts which can be analysed, and then later related to each other (see section 2.3.3). Studies of segmentation of non-musical sound sequences as well as of musical sequences suggest that the Gestalt principles of proximity and similarity may be the basis on which listeners segment music (see section 2.3.1). Based on such principles, several researchers have proposed three cases of rhythmic segmentation.

The three cases of rhythmic segmentation — rests, longer durations, and breaks of similarity — are described by three Lerdahl and Jackendoff's grouping rules (Jackendoff & Lerdahl, 1981; Lerdahl & Jackendoff, 1983b, 1983a) presented in the second chapter (sections 2.3.4.4 and 2.3.4.5). These cases of segmentation are also acknowledged by Drake and Palmer (1993), and Kirkpatrick (1984), and are supported by experiments performed by Deliege (1987). In our experiments presented in this chapter therefore, we assume the validity of the three Lerdahl and Jackendoff's rules. We assume that listeners do have the ability to recognize the cases of segmentation and perform segmentation according to the rules when listening to music.

This chapter proposes a novel representation for rhythmic sequences, and a neural model to segment musical pieces in accordance with the three cases of segmentation. It comprises seven

Activation a_i of each hidden unit i is given by the sigmoid function

$$a_i = \frac{1}{1 + e^{-net_i}} \tag{4.1}$$

net_i is given by

$$net_i = \sum_j w_{ij} a_j + bias_i \tag{4.2}$$

where w_{ij} is the weight from input unit j to hidden unit i, a_j is the activation of input unit j, and $bias_i$ is a special weight which adjusts values of net_i to make an efficient use of threshold of the sigmoid.

The output layer holds linear units to avoid *flat spots*[1] (Fahlman, 1988). Activation a_i of each output unit i is thus given by

$$a_i = net_i = \sum_j w_{ij} a_j + bias_i \tag{4.3}$$

where w_{ij} is the weight from hidden unit j to output unit i, a_j is the activation of hidden unit j, and $bias_i$ is again a special weight[2].

Weights are updated according to generalized delta rule (Rumelhart et al., 1986a),

$$\Delta w_{ij}(p) = \alpha \delta_i a_j + \beta \Delta w_{ij}(p-1) \tag{4.4}$$

where $\alpha, \beta \in (0,1)$ are the learning rate and momentum respectively. Subscript p indexes pattern number, and learning takes place on a pattern-by-pattern basis.

At the end of each epoch, both learning rate and momentum are m

Table 4.1. Generative templates of the pattern sets (first experiment) — R: rest; NS: note sustained; NO: note onset; ***FS***: free slot;

Pattern Templates
[NS] [NS] [NS] *(7×FS)*
[NS] [R] [R] *(7×FS)*
[NS] [R] [NO] *(7×FS)*
[NS] [NO] [NS] *(7×FS)*
[NS] [NO] [R] *(7×FS)*
[NS] [NO] [NO] *(7×FS)*
... ...
[NS] [NO 7×NS] [NS]
[NS] [NO 7×NS] [R]
[NS] [NO 7×NS] [NO]
[R] [R] [R] *(7×FS)*
[R] [R] [NO] *(7×FS)*
[R] [NO] [NS] *(7×FS)*
[R] [NO] [R] *(7×FS)*

Table 4.2. Number of negative and positive patterns produced after the number of free slots (first experiment)

Free Slots	No. Neg. Pats.	No. Pos. Pats.
0	1	1

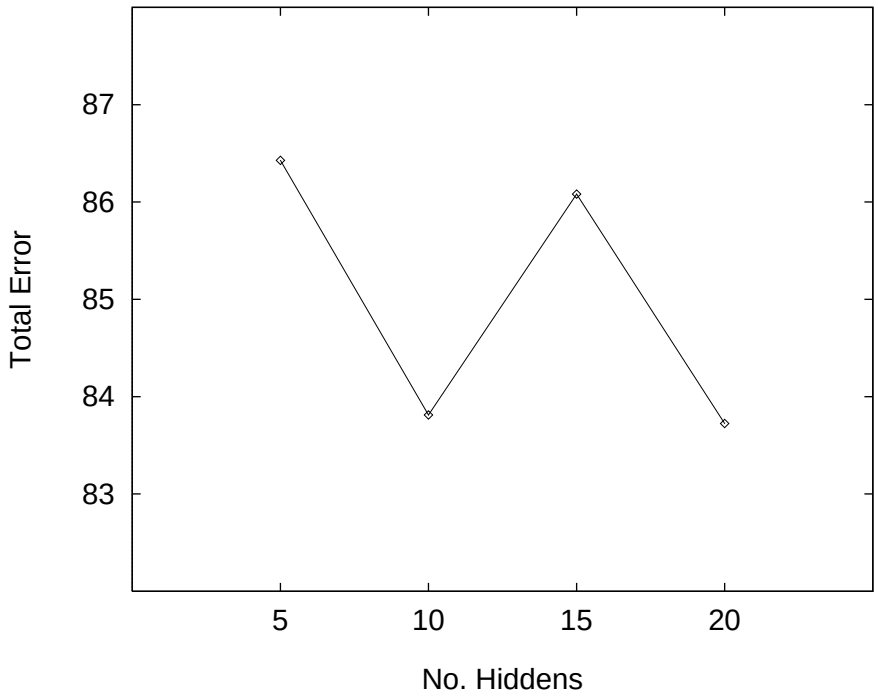


Figure 4.6. Testing on the third set (first experiment)

As the first three sets, a fourth set of patterns was generated from all templates in table 4.1 as well. Free slots were also filled up randomly with the three events, and each template generated just one pattern. The fourth set contained thus, 8 positive and 73 negative patterns.

Principal component analysis[4] (Everitt & Dunn, 1991; Everitt, 1993) was performed on the

Table 4.3. Results of the first experiment — #NP: number of negative patterns; #PP: number of positive patterns; %NM: percentage of misclassified negative patterns; %PM: percentage of misclassified positive patterns;

Pieces	#NP	#PP	%NM	%PM
1	814	2	0	0
2	784	16	0	0
3	1188	12	0	0
4	2293	10	0	60
5	4556	44	0	14
6	2204	36	0	0

1500 negative and 1500 positive patterns. The sets were generated through the pattern templates described in table 4.4. The number of patterns produced through a template was also dependent on the number of free slots available in the template. The numbers of negative and positive patterns produced after the numbers of free slots are displayed in table 4.5.

The structure of the templates was based on the rule of segmentation given by longer durations. It is made up by four parts, which are represented by the four groups of square brackets occurring in each template displayed in table 4.4. The first and third parts contain either one rest event, or one note onset event, or one note onset event followed by any number, between one and three, of note sustained events. The second part contains either one rest, or one note onset, or one note onset followed by any number, between one and nine, of notes sustained. The last part contains either one rest event, or one note onset event.

The templates consist of all possible combinations of the events in each part. The positive templates, that is, the templates which generate positive patterns, are all those which satisfy three conditions. First, the first, third, and last parts of the template contain one note onset event. Second, the second part contains one note onset event followed by a number of note sustained events. Third, when either the first, or third parts, or both contain note sustained events, then the number o5.64311(h)-8

Table 4.4.

Table 4.5. Number of negative and positive patterns produced after the number of free slots (second experiment)

Free Slots	No. Neg. Pats.	No. Pos. Pats.
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	3
10	1	9
11	1	28
12	1	85
13	5	256
14	16	712
15	36/37	—

Following the same process of the first experiment (section 4.4), a fourth set containing 110 positive and 440 negative patterns was generated from the templates in table 4.4. Principal component analysis (Everitt & Dunn, 1991; Everitt, 1993) was carried out on the activations of the hidden units given by each pattern in the set. Figure 4.14 plots the two first principal components (PCs) for the patterns in the set. Figures 4.15 and 4.16 plot, respectively, the two first principal components (PCs) for the negative and positive patterns in the set which were correctly classified by the neural model. As in the first experiment, it can be noticed that there is no correlation between negative and positive patterns, and thus, the internal representations stored in the hidden units are able to distinguish between the two types of patterns.

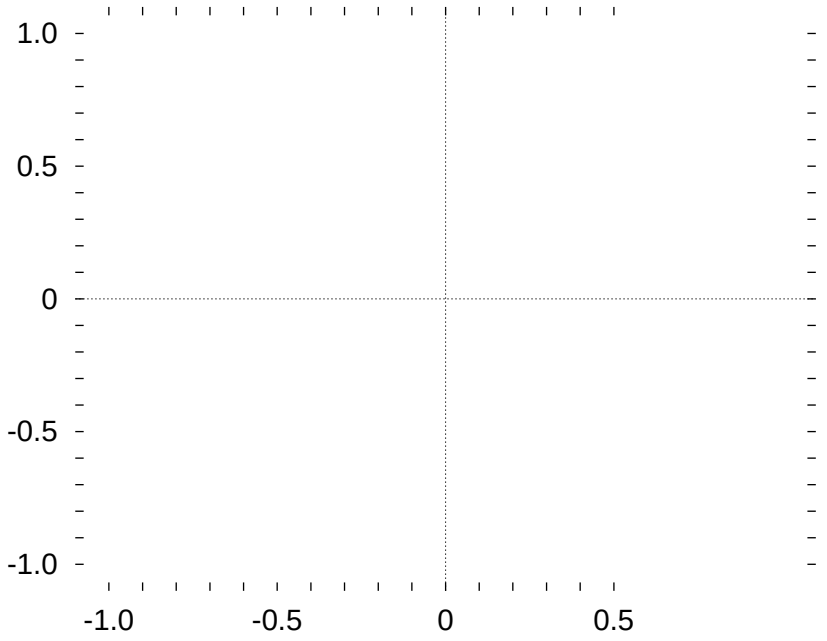


Table 4.6. Results of the second experiment — #NP: number of negative patterns; #PP: number of positive patterns; %NM: percentage of misclassified negative patterns; %PM: percentage of misclassified positive patterns;

Pieces	#NP	#PP	%NM	%PM
1	764	52	2	0
2	790	10	0	0
3	1171	25	10	0
4				

Table 4.7. Generative templates of the pattern sets (third experiment) — R: rest; NS: note sustained;

Table 4.8. Number of negative and positive patterns produced after the number of free slots (third experiment)

Free Slots	No. Neg. Pats.	No. Pos. Pats.
0	1	—
1	1	—
2	1	—
3	1	—
4	1	750
5	3	—
6	11	—
7	34	—
8	122	—

The structure of the templates was based on the rule of segmentation given by breaks of similarity. It is made up by eight parts, which are represented by the eight groups of square brackets occurring in each template displayed in table 4.7.

The templates were divided into two sets to reduce the number of possible combinations of events in them. The first set contains templates which include any number, between one and eight, of rest events. Thus, each of the eight parts of the templates contains either one rest event, or one note onset event, or one note onset event followed by one note sustained event. The 33 templates in this set were created randomly, and are described by the first 33 lines of table 4.7.

The second set contains templates which do not include rest events. Each of the eight parts of the templates contains thus either one note onset event, or one note onset event followed by one note sustained event. The templates in this set consist of all possible combinations of the events in each part. They are described by the last five lines of table 4.7.

The positive templates, that is, the templates which generate positive patterns, are all those which satisfy four conditions. First, none of the parts of the template contains rest events. Second, the first, second, third, and fourth parts contain identical events. Third, the fifth, sixth, seventh, and eighth parts contain identical events as well. Fourth, the events in the first four parts are different from those in the last four parts. The negative templates are all those otherwise.

As in previous experiments, we might illustrate the generation of a pattern in an example. Let us consider the positive template in which each one of the first four parts contains one single note onset, and each one of the last four parts contains one note onset followed by one note sustained. There are four free slots at the beginning of the template, and thus, according to table 4.8, 750 positive patterns are generated from the template. The generated patterns consist of four events which are determined randomly, followed by the events in the eight parts of the template.

The training method was identical to that followed in previous experiments (sections 4.4 and 4.5). Figures 4.18, 4.19, and 4.20 show the training and testing on the sets for four configurations. For each configuration, the window in the input layer held 16 pairs of units, and initial weights were set randomly. Similar performances were reached by neural models with different configurations. Yet, configurations with 5 and 20 hidden units performed better in terms of generalization. The configuration chosen was that with 5 hidden units, traine

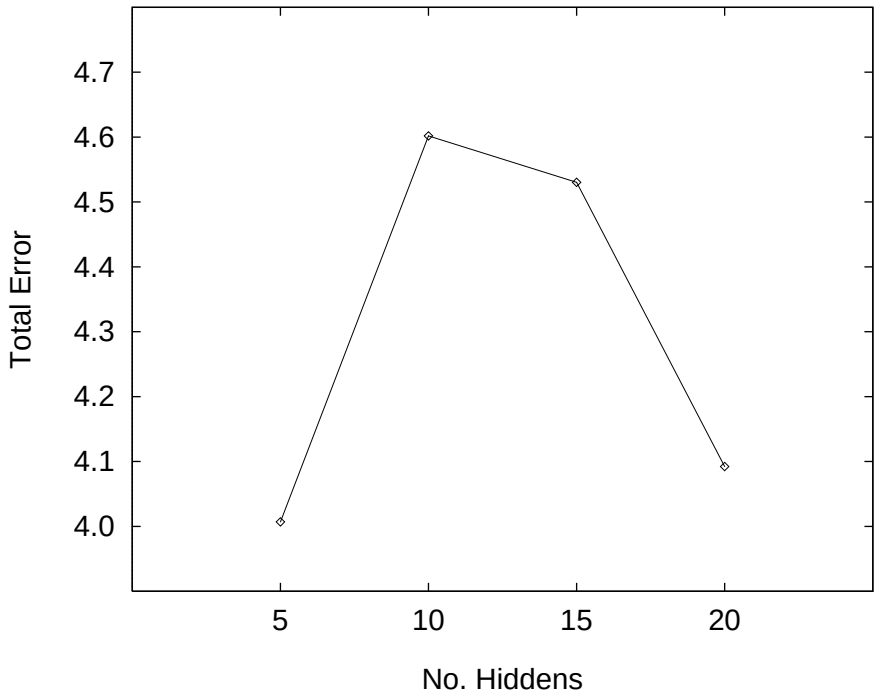


Figure 4.20. Testing on the third set (third experiment)

A fourth set containing 2 positive and 287 negative patterns was generated from the templates in table 4.7, according to the process employed in previous experiments (sections 4.4 and 4.5). Principal component analysis (Everitt & Dunn, 1991; Everitt, 1993) was performed on the activations of the hidden units given by each pattern in the set. Figure 4.21 plots the two first principal components (PCs) for the patterns in the set. Figures 4.22 and 4.23 plot, respectively, the two first principal components (PCs) for the negative and positive patterns in the set which were correctly classified by the neural model. Unlike previous experiments, positive and negative patterns are not separated into two different clusters.

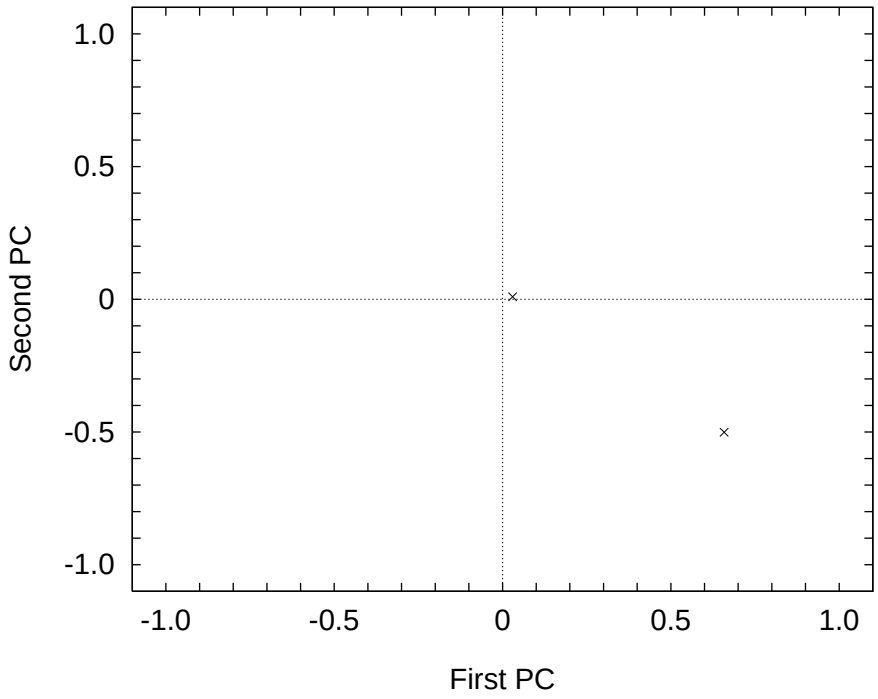


Figure 4.23. Two first PCs for the positive patterns correctly classified (third experiment)

One negative pattern from the cluster of the plot in figure 4.22, and one positive pattern — the farthest from origin — from the plot in figure 4.23 were selected. The internal representations stored in the hidden units for both patterns are shown in Hinton's diagram of figure 4.24. It can be observed, unlike previous experiments (sections 4.4 and 4.5), that the representations are

Table 4.9. Results of the third experiment — #NP: number of negative patterns; #PP: number of positive patterns; %NM: percentage of misclassified negative patterns; %PM: percentage of misclassified positive patterns;

Chapter 5

A neural model for thematic recognition

5.1 Introduction

Studies in perception relating to interleaved melodies and multivoiced music presented in the second chapter are loosely related with thematic recognition. However, some of them, such as Dowling's (1973, 1987) studies (sections 2.4.1 and 2.4.2) conce

Table 5.1. Representation for a binary sequence

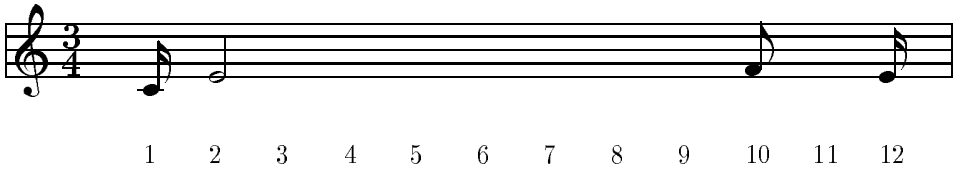


Figure 5.1. An univoiced musical sequence

Table 5.2.

although it has particularly been used to represent the input data in the third experiment (section 5.8).

The input data in the third experiment consists of a sequence of musical intervals (see section 2.2), which corresponds to a fugue (see section 2.4.5.6). Data is input one TIC (see section 4.2) at a time.

As in section 5.3, fifteen neural units are used in the representation. Each unit represents one musical interval ranging from an octave down to an octave up. We assume here, therefore, that

5.5 The model

As those reviewed in section 3.4, the model introduced here is an extension of the self-organizing map. As Chappell and Taylor's model (section 3.4.2), it follows a time integral approach, and

smallest distance $\Psi(i^*,t)$. For each output unit *i*, the distance $\Psi(i,t)$ between the input vector $\mathbf{X}(t)$ and the unit's weight vector $\mathbf{W}_i$ is given by

$$\Psi(i,t) = \sum_{j=1}^{m} [x_j(t) - w_{ij}(t)]^2 \tag{5.2}$$

Each output unit *i* in the neighbourhood N^* of the winning unit i^* has its weight $\mathbf{W}_i$ updated by

$$w_{ij}(t+1) = w_{ij}(t) + \alpha\Upsilon(i)[x_j(t) - w_{ij}(t)] \tag{5.3}$$

where $\alpha \in (0,1)$ is the learning rate. $\Upsilon(i)$ is the ***neighbourhood interaction function*** (Lo & Bavarian, 1991), a gaussian type function, and is given by

$$\Upsilon(i) = \kappa_1 + \kappa_2 e^{-\frac{\kappa_3[\Phi(i,i^*)]^2}{2\sigma^2}} \tag{5.4}$$

where κ_1, κ_2, and κ_3 are constants which confer the shape to the function. We have set κ_1, κ_2, and κ_3 to be 0.1, 0.7, and 10 in our experiments. σ is the radius of the neighbourhood N^*, and $\Phi(i,i^*)$ is the distance in the map between the unit *i* and the winning unit i^*. The distance $\Phi(i',i'')$ between any two units i' and i'' in the map is calculated according to the maximum norm,

$$\Phi(i',i'') = max\left\{\left|l'-l''\right|, \left|c'-c''\right|\right\} \tag{5.5}$$

where (l',c') and (l'',c'') are the coordinates of the units i' and i'' respectively in the map.

The neighbourhood interaction function has proved to be useful, indeed. It provokes two main effects. First, it speeds up the training of the network by reducing the number of epochs required. Second, it improves the quality of the map by enforcing its topological order (Lo et al., 1991). In rough terms, the neighbourhood interaction function avoids the existence of *local winning units*. The values of the distances $\Psi(i,t)$ increase as the values of the distances $\Phi(i,i^*)$ increase.

The input to the top SOM is determined by the distances $\Psi(i,t)$ of the *n* units in the map of the bottom SOM. The input is thus a sequence in time of *n*-dimensional vectors, $\mathbf{S_2} = \Lambda(\Psi(i,1))$, $\Lambda(\Psi(i,2)), \ldots, \Lambda(\Psi(i,t)), \ldots, \Lambda(\Psi(i,z))$, where $\Lambda(\Psi(i,t))$ is a *n*-dimensional ***transfer function*** on a *n*-dimensional space domain. We have used two different kinds of transfer function in our experiments. Λ can be defined as a gaussian type function as

$$\Lambda(\Psi(i,t)) = e^{-\frac{\kappa[\Psi(i,t)]^2}{\rho^2}} \tag{5.6}$$

where κ is a constant, and ρ is the radius of the gaussian. The advantage of using such a function is that the contributions to the input of the top SOM depend entirely upon the distances $\Psi\varepsilon$ /P19 6 .93404(o)–4.11137

The sequence $\mathbf{S_2}$ is then presented to the input layer of the top SOM, one vector at a time. The input layer has n units, one for each component of the input vector $\Lambda(\Psi(i,t))$, and a time integrator. The activation $\mathbf{X}(t)$ of the units in the input layer is given by

$$\mathbf{X}(t) = \Lambda(\Psi(i,t)) + \delta_2 \mathbf{X}(t-1) \tag{5.8}$$

where $\delta_2 \in (0,1)$ is the decay rate.

The dynamics of the top SOM is identical to that of the bottom SOM. In all experiments (sections 5.6, 5.7, and 5.8), we have first trained the bottom SOM, and then, the top SOM, for the sake of efficiency. As the model presented has two SOMs, it will be referenced as *model II* in the rest of the chapter. To be better evaluated, the performance of the model II is compared to that of the *model I*. Model I has only one SOM. As the bottom SOM of the model II, model I also has m input units, a time integrator applied on them, and the same dynamics. Model I can be classified as a single SOM which follows the time integral approach. So, as pointed out in section 3.4.2, model I suffers from loss of context.

5.6 First experiment

The first experiment was on mapping a set of sequences. In it, the model was applied to a small scale problem in order to analyse its behaviour.

The input data consisted of a set of sixty six-bit binary sequences (e.g., 011101). The sequences were generated randomly. The sequence 001011 was chosen arbitrarily as a reference. It is named *referential sequence* ($\mathbf{S_r}$). The referential sequence has the largest number of similar sequences in the set, that means, sequences which differ slightly from the referential sequence in the order and values of the bits.

The experiment aimed at verifying how accurate the classification of the referential sequence yielded by models I and II was. In other words, we verified the number of sequences in the set which were misclassified by models I and II as the referential sequence.

The two SOMs of model II and the SOM of model I were trained in two phases — coarse-mapping and fine-tuning. The initial learning rate was set to 0.5, and the size of the neighbourhood was set to the size of the map in the coarse-mapping phase. Both the learning rate and the radius of the neighbourhood were reduced linearly to the values 0.01 and 1 respectively. In the fine-tuning phase, the learning rate was kept constant in 0.01, and the radius in 1. The coarse-mapping phase took 20%, and the fine-tuning phase took 80% of the total number of epochs. The initial weights were given randomly, in the range between 0 and 0.1, to all SOMs.

Different decay rates were tried. In the bottom SOM of model II, they ranged from 0.4 to 0.7, and in the top SOM, from 0.7 to 0.95. In the model I, the decay rate ranged from 0.7 to 0.95. The input layer of the model I and of the bottom SOM of model II held two units. The representation employed in these units is fully described in section 5.2.

Model I was tested with three different map sizes, 9×9, 15×15, and 21×21, trained in 400, 700, and 1000 epochs respectively. In model II, the map sizes were set to 6×6 (trained in 250 epochs) and 9×9 (trained in 400 epochs) to the bottom and top SOM respectively. The transfer function Λ was given by equation 5.7, with $N^* = \{i^*\}$.

The best results of models I and II are displayed in the tables 5.5 and 5.6 respectively. A sequence $\mathbf{S_a}$ is said to have the same classification as that of the referential sequence $\mathbf{S_r}$ if the distance $\Phi(i_a^*, i_r^*) < 2$, where i_a^* and i_r^* are the (last) winning units of $\mathbf{S_a}$ and $\mathbf{S_r}$.

As expected, model I suffers from loss of context and misclassifies several sequences. It is difficult for the model to distinguish variations in the first bits of a sequence because the contribution of these first bits to the classification of the sequence is very low. For instance, let $\mathbf{S_a}$ = 100000 and $\mathbf{S_b}$ = 010000 be two sequences. Considering a decay rate of 0.8, the activations of the two input units would be 3.362 and 0.328 after the entrance of the last bit of $\mathbf{S_a}$. The activations would

Table 5.5. Results for model I (first experiment)

Map Size	Decay Rate	No. Miscl.
9×9	0.7	9
15×15	0.7/0.9	5
21×21	0.9	2

Table 5.6. Results for model II (first experiment)

Decay Rate Bottom SOM	Decay Rate Top SOM	No. Miscl.
0.4	0.7	1
0.5	0.75/0.8	1
0.6	0.8	1

be 3.280 and 0.410 for $\mathbf{S_b}$. The differences in the activations between $\mathbf{S_a}$ and $\mathbf{S_b}$ are not relevant, and probably the sequences would be classified as identical by model I.

The problem with model I is that the SOM sees just bits in its input. Yet, its performance would be much improved if the input not only represented bits, but also the context where they appeared. Different input units would then be activated depending upon the order that the bits were input. For example, considering a representation that includes three bits at most, $\mathbf{S_a}$ and $\mathbf{S_b}$ would be represented by table 5.7. As the representation makes a clear distinction between the beginnings of $\mathbf{S_a}$ and $\mathbf{S_b}$, it helps model I to distinguish between the two sequences as well.

Table 5.7. Context representation for two binary sequences

Seq.	time: 1	time: 2	time: 3	time: 4	time: 5	time: 6
$\mathbf{S_a}$	(1)	(10)	(100)	(000)	(000)	(000)
$\mathbf{S_b}$	(0)	(01)	(010)	(100)	(000)	(000)

The idea of encoding context in the representation to distinguish variations ininhe

a contextless input of single bits in its map.

We might illustrate these ideas in an example. The sequence $\mathbf{S_a}$ = 001111 is presented to the

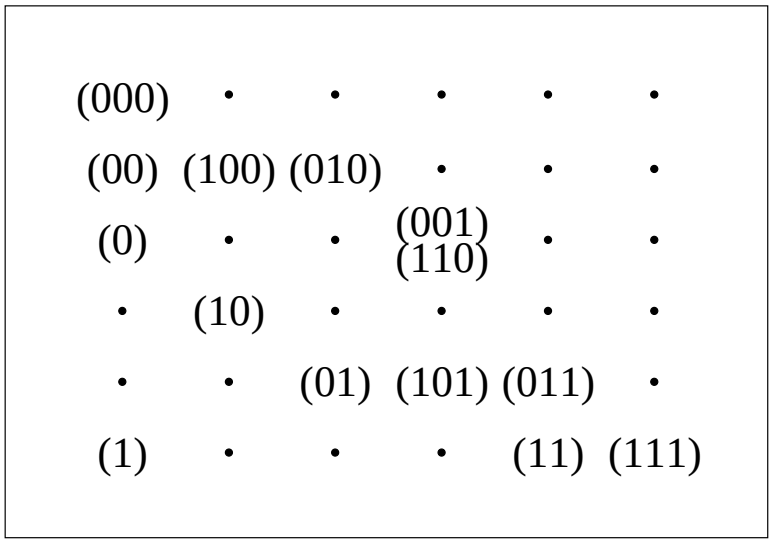


Figure 5.5. The map for three-bit binary sequences

5.7 Second experiment

The second experiment was on thematic recognition on an univoiced musical sequence. As introduced in the second chapter (section 2.4.5.1), an univoiced musical sequence is a sequence which contains just one single voice.

The input data consisted of two sets, hereafter referred to, in this section, as ***input set I*** and ***input set II***. Input set I consisted of a large and unique sequence of musical intervals, which corresponded to the third voice of the sixteenth four-part fugue in G minor of the first volume of The Well-Tempered Clavier of Bach (see section 2.4.5.6). Input set II, in its turn, contained many sequences, which corresponded to groupings produced by segmentation given by rests (section 2.3.4.4) applied to the third voice of the fugue.

Model I was trained and tested on input set I only. We realized that it would not be worth applying it to input set II as well, since it performed poorly on set I. Model II was trained and tested in both sets I and II. Therefore, when set I was used, models I and II were trained and assessed on the recognition of sub-sequences within the third voice of the fugue. Otherwise, when set II was used, model II was trained and assessed on the recognition of sequences produced by segmentation.

The fugue in G minor has 544 TICs, and TI is a sixteenth note. The ***theme*** of the fugue (figure 5.6), a ***referential sequence (or sub-sequence)***, was divided into two parts — ***theme I*** and ***theme II***.

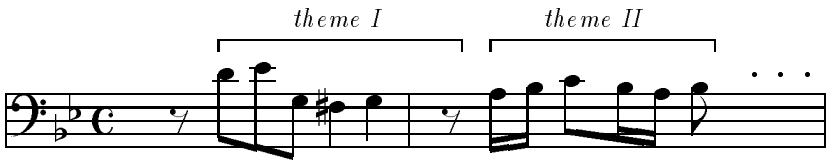


Figure 5.6. Theme of the sixteenth fugue in G minor

The fugue in G minor was chosen for several reasons. First, as

phase took 20%, and the fine-tuning phase took 80% of the total number of epochs. The initial weights were given randomly, in the range between 0 and 0.1, to all SOMs.

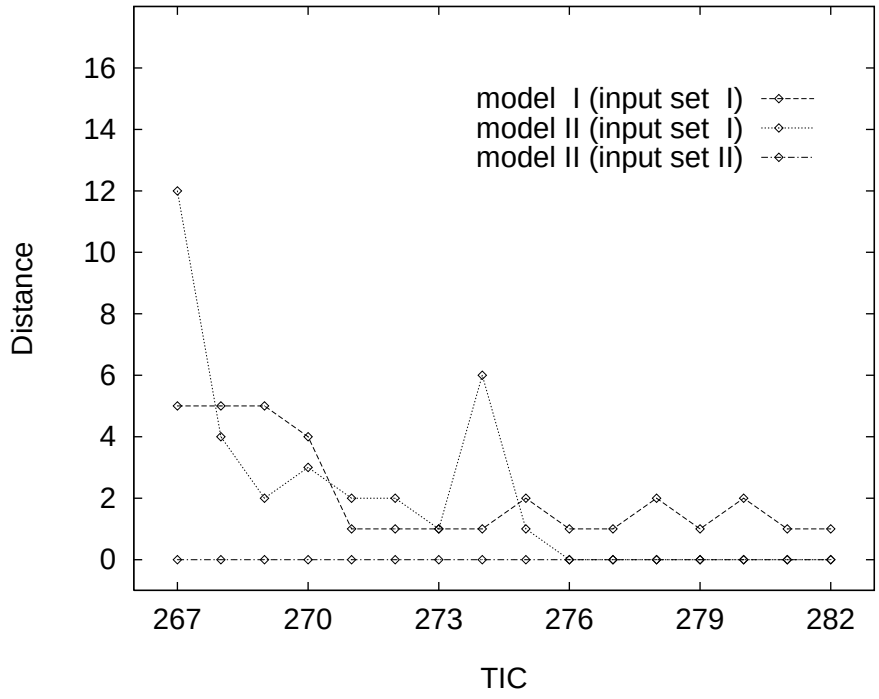
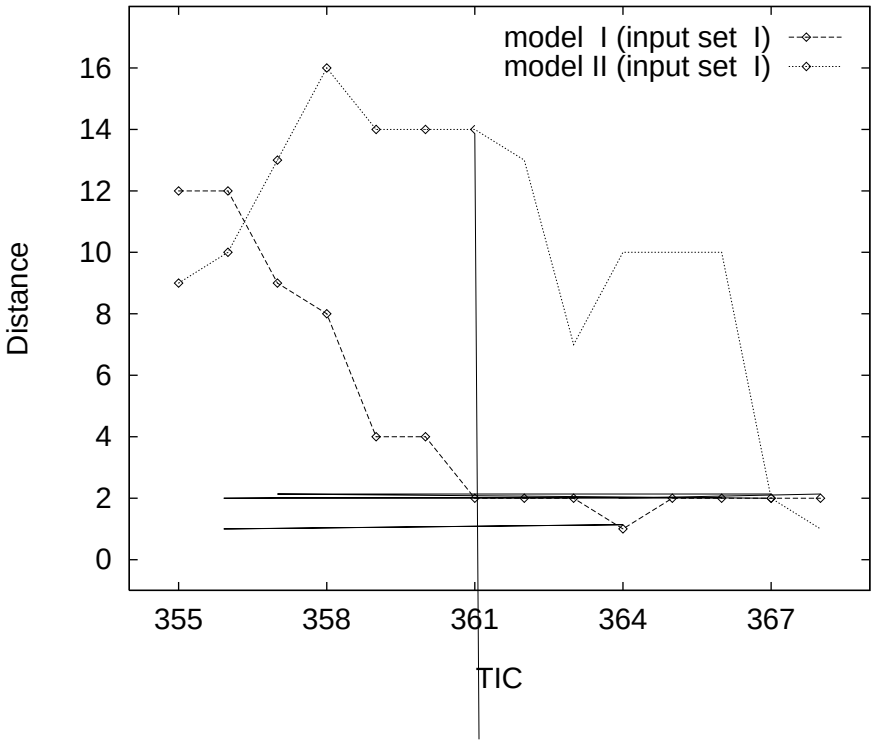

model I (input set I)
model II (input set I)
model II (input set II)
Distance
TIC
0
2
4
6
8
10
12
14
16
267
270
273
276
279
282

Figure 5.8.


model I (input set I)
model II (input set I)
Distance
16
14
12
10
8
6
4
2
0
355
358
361
364
367
TIC

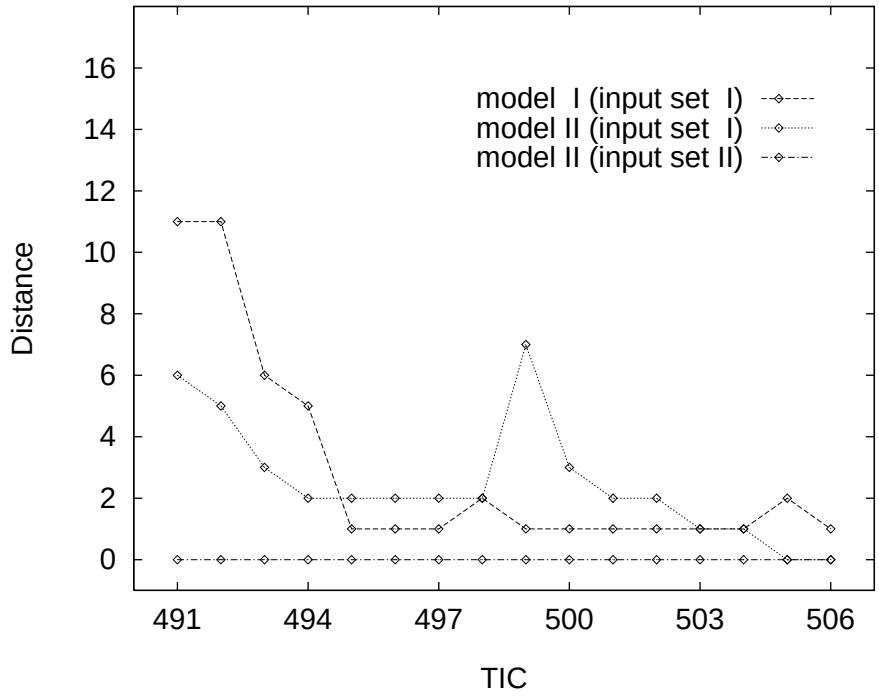
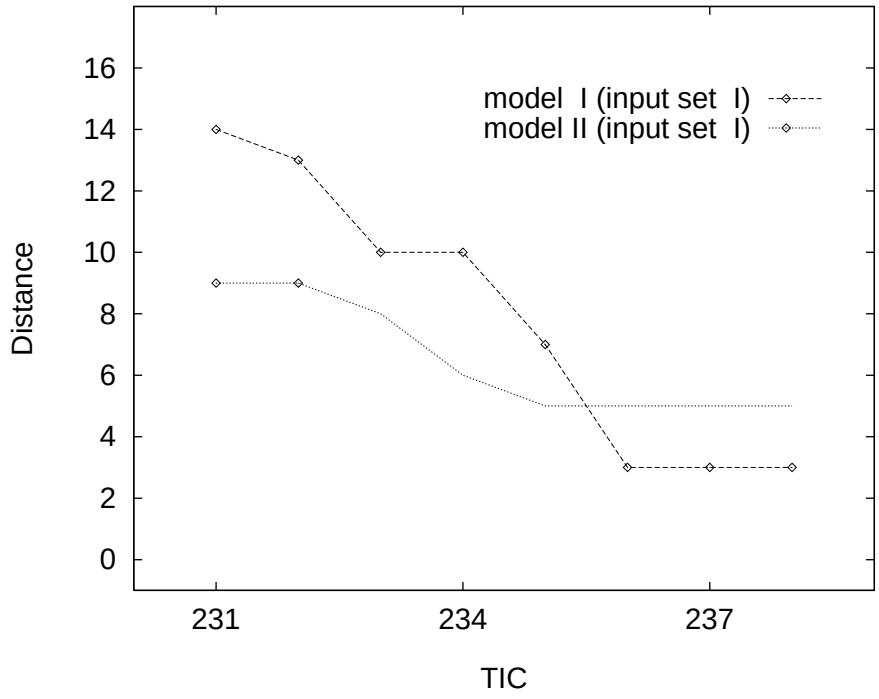


Figure 5.10. Classifications of the fourth instance of theme I (TICs 491 – 506) relative to theme I. The instance is a perfect copy of the theme I. There is a rest before the entrance of the fourth instance, and therefore, the activation of the input units immediately before the entrance of the instance is high when using input set I, and non-existent when using input set II. The classifications of the instance produced by model I and II converge to that of the theme I.


16
14
12
10
8
6
4
2
0
Distance
231
234
237
TIC
model I (input set I)
model II (input set I)

- -

0 - -

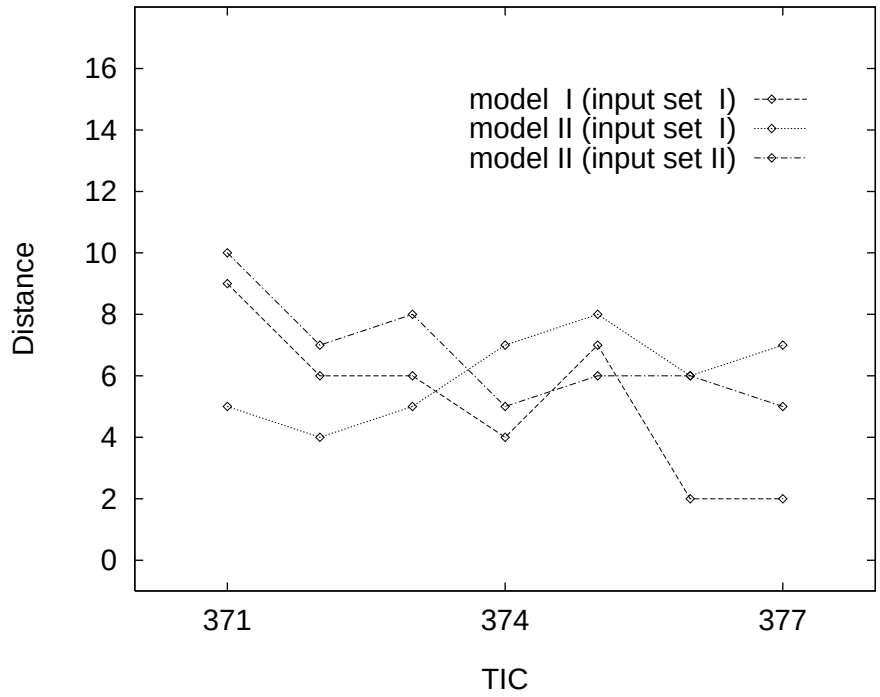


Figure 5.14. Classifications of the fourth instance of theme II (TICs 371 – 377) relative to theme II. The instance is a perfect copy of the theme II. The instance of theme II is preceded by an instance of theme I, which occurs altered in its first and last two TICs. Consequently, the activation of the input units immediately before the entrance of the instance is not similar to that immediately before the entrance of theme II. The classifications of the instance produced by model I and II do not converge to that of the theme II.

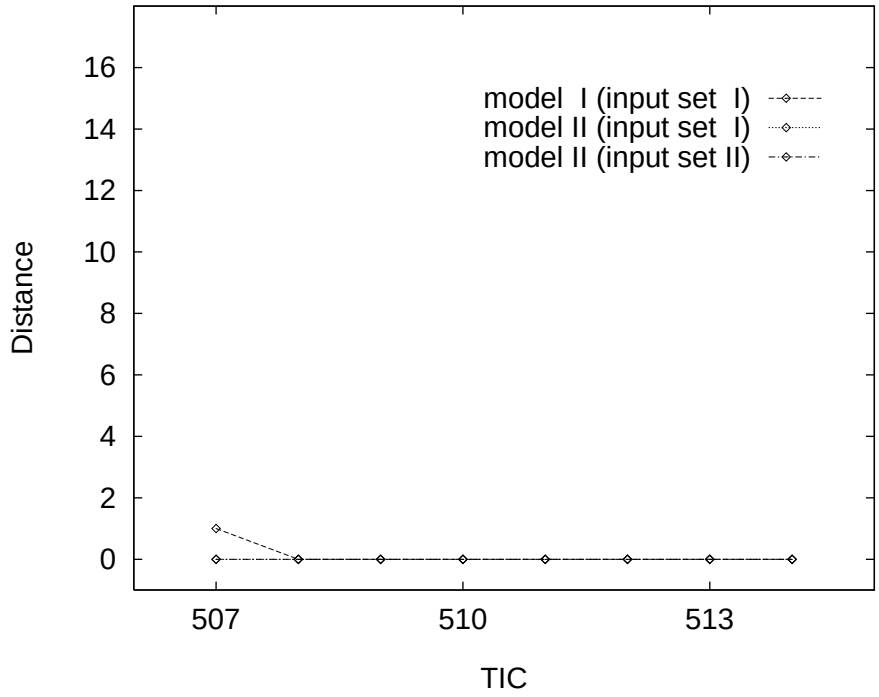


Figure 5.16. Classifications of the sixth instance of theme II (TICs 507 – 514) relative to theme II. The instance is a perfect copy of the theme II. The sixth instance of theme II is preceded by a perfect instance of theme I. Moreover, theme I ends with a rest. Two cases exist, therefore. When input set I is used, the context preceding the instance is similar to that preceding theme II, and consequently, the activation of the input units immediately before the entrance of the instance is similar to that immediately before the entrance of theme II. In its turtu4419.85 300i(e)12(r6971(f)-2.9988n)-4.00ut set II is empl10(i)-10(o)-4.00y-(4412(e)]TJ31

Table 5.8.

Considering input set I only, one may verify in the figures 5.7 to 5.16, and in the table 5.8, that the mean errors of model I are lower than those of model II. The reason is that, as expected, model II takes into a better account the past context. As the previous context varies from instance to instance, model II takes more time to discard the previous context of the instance to converge to the theme. One may also verify that both models fail in recognizing three instances of theme II. In all of these cases however, theme II either was preceded by a modified theme I or was not preceded by theme I at all. Once more, the different previous context was responsible for that failure.

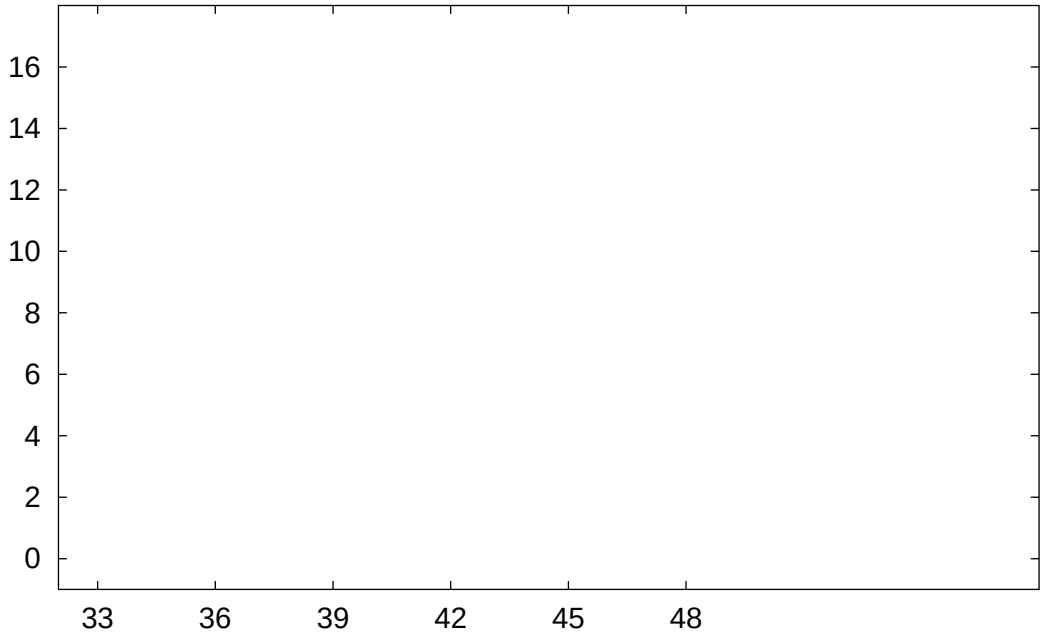
The performance of model II is better appreciated in the results displayed in table 5.9. The fugue is made up mostly by contiguous intervals (e.g., seconds and thirds up and down) in different orders and rhythms. It is worthwhile to observe the fact that

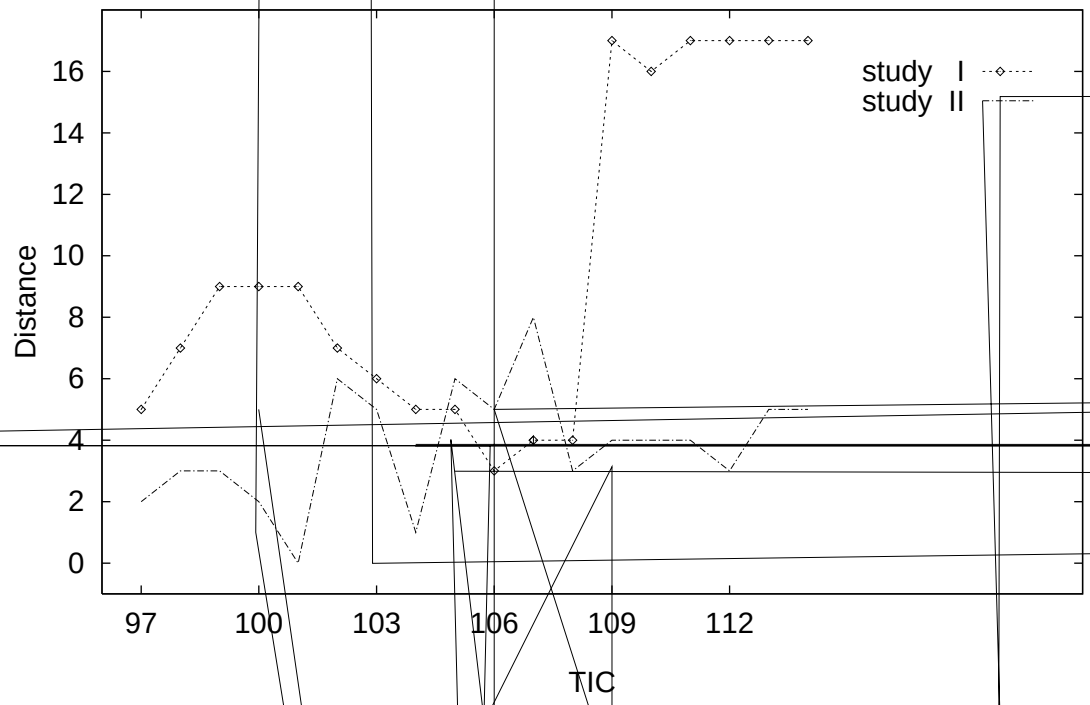
activation values of notes onset and sustained, whether reinforced or not, as well as the input set employed in each study.

Table 5.10. Parameter values of the studies

Study	Input Set	Reinforcement Value	Note Onset	Note Sustained	N. Onset (Reinforced)	N. Sustained (Reinforced)
I	I	1	0.1	0.07	0.1	0.07
II	I	5	0.1	0.07	0.5	0.35
III	I	10	0.1	0.07	1.0	0.7
IV	I	100	0.1	0.07	10.0	7.0
V	II	100	0.1	0.07	10.0	7.0

Reinforcement was provided from the seventh common TIC between the theme and any of its instances. For example, let us consider the sequence of theme $\mathbf{S_t}$, and a sequence $\mathbf{S_a}$, which is an instance of theme. Let us suppose that, on a determined TIC, $\mathbf{S_a}$ holds an interval onset or sustained which matches one of the intervals in $\mathbf{S_t}$. Let us suppose now that, on the next five TICs, the next five intervals in $\mathbf{S_a}$ also match the next corresml e ict fv 310.957(n) 354.809(i) 6.93404(a) 1.1
st corranl(a) 5.04534(n) 5.04534(ie) 5.04534(


16
14
12
10
8
6
4
2
0
33
36
39
42
45
48


study I
study II
Distance
TIC
0
2
4
6
8
10
12
14
16
97
100
103
106
109
112

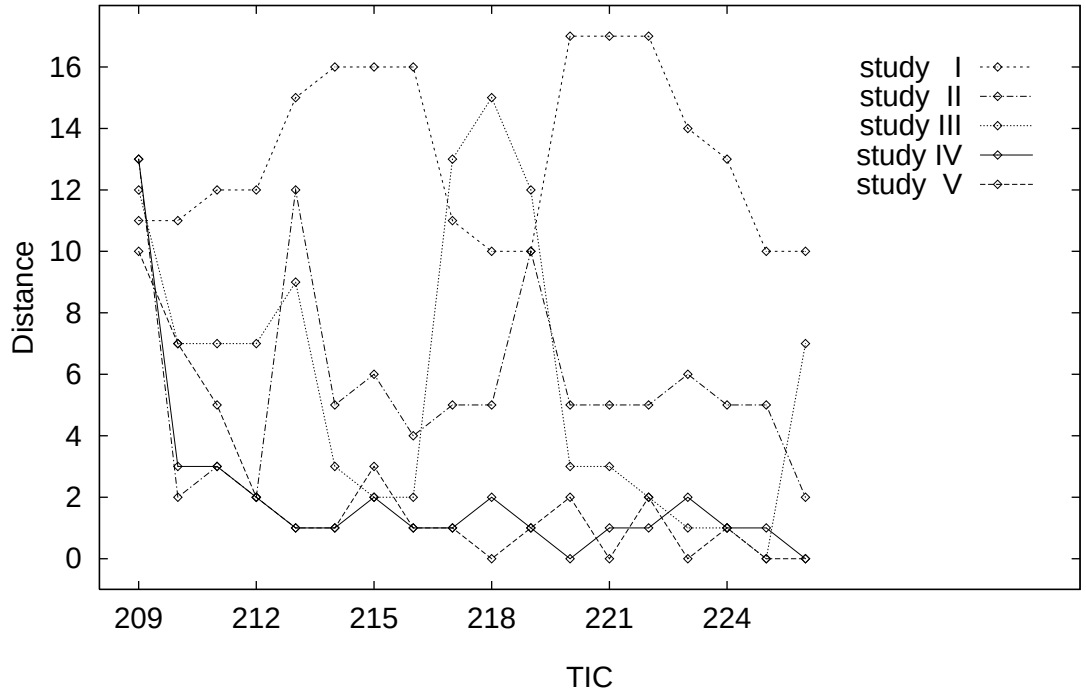
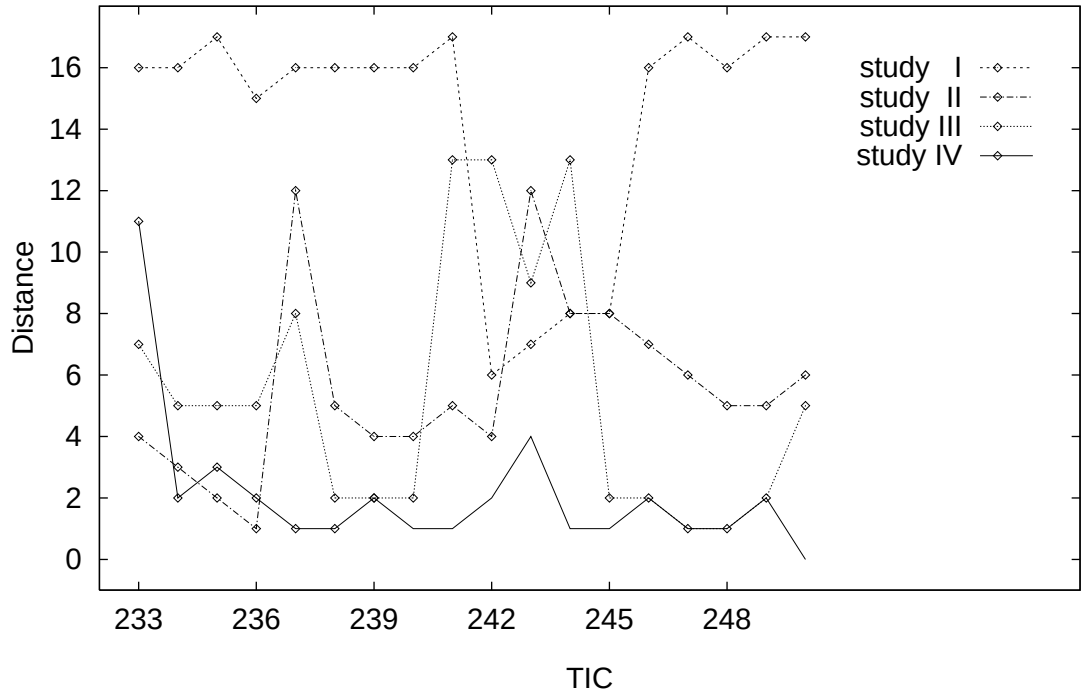


Figure 5.21. Classifications of the fifth instance of theme (TICs 209 – 226) relative to theme. The instance occurs in the first voice, concurrently with two other voices. The instance differs from the theme in its first two TICs. The classifications of the instance yielded by model II only converge to that of the theme in the fourth and fifth studies.



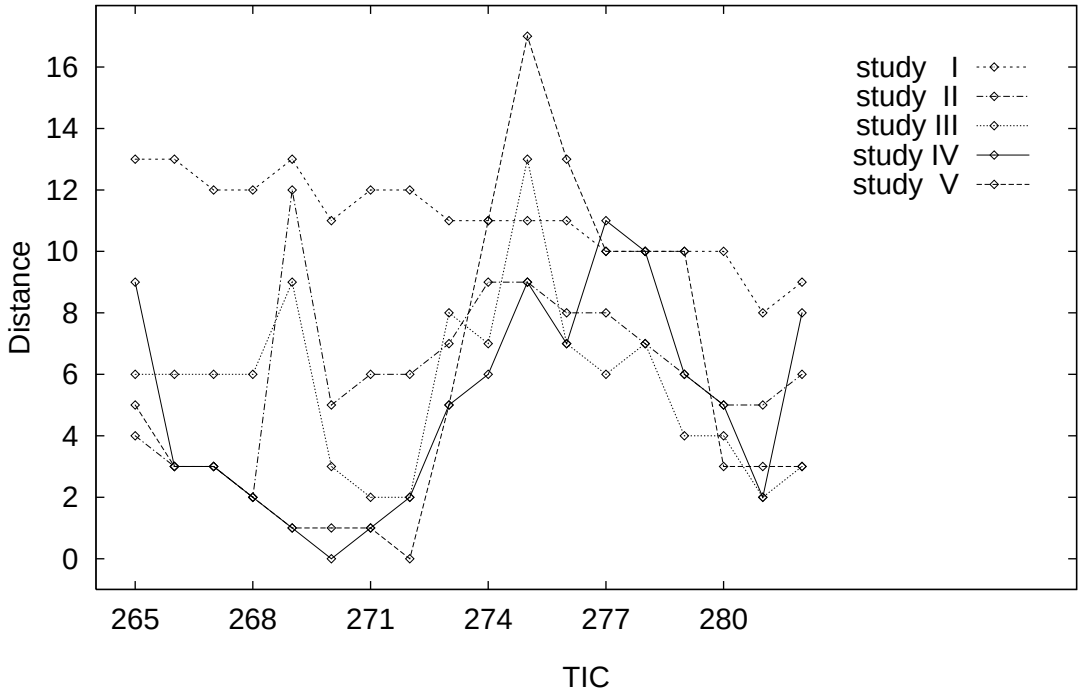
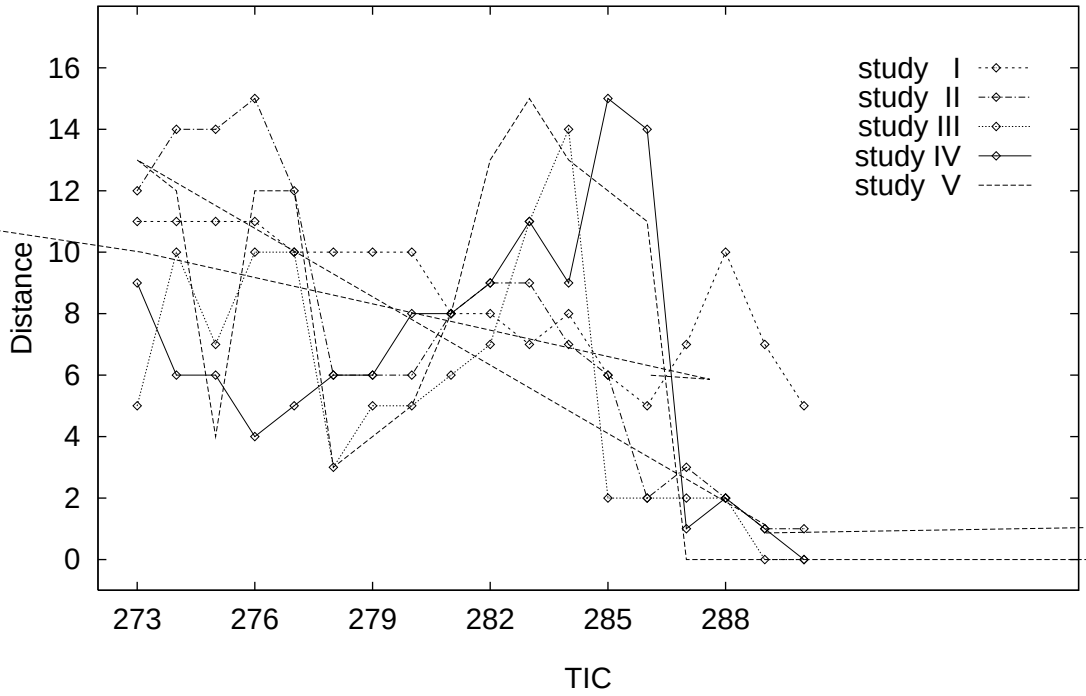


Figure 5.23. Classifications of the seventh instance of theme (TICs 265 – 282) relative to theme. The instance occurs in the first voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance yielded by model II do not converge to that of the theme in any of the studies.



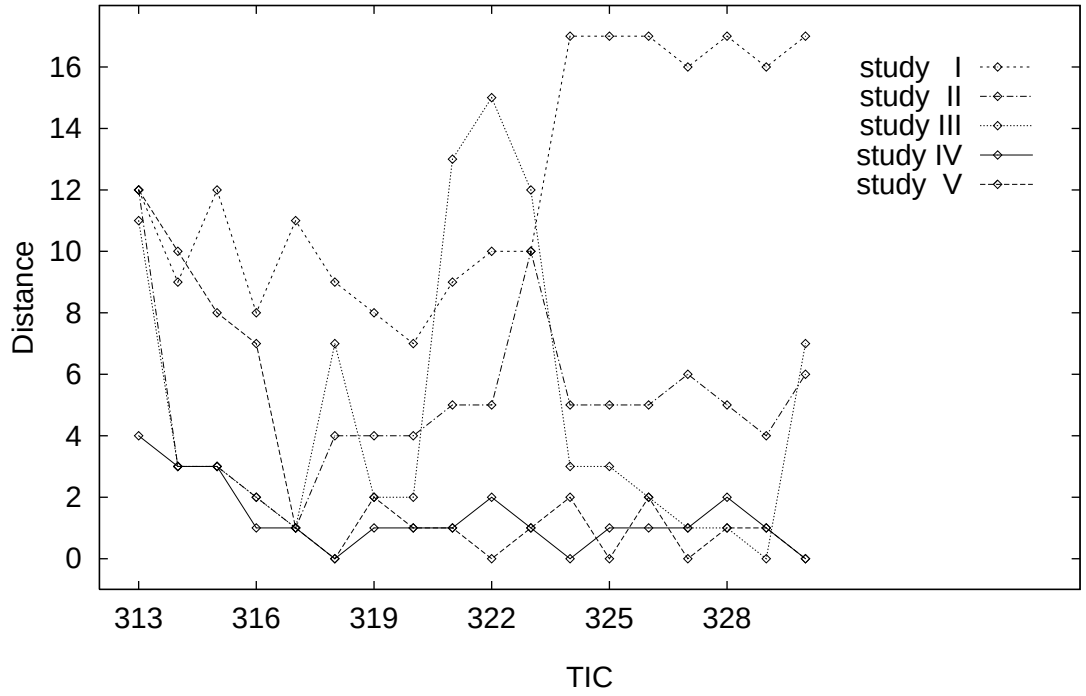
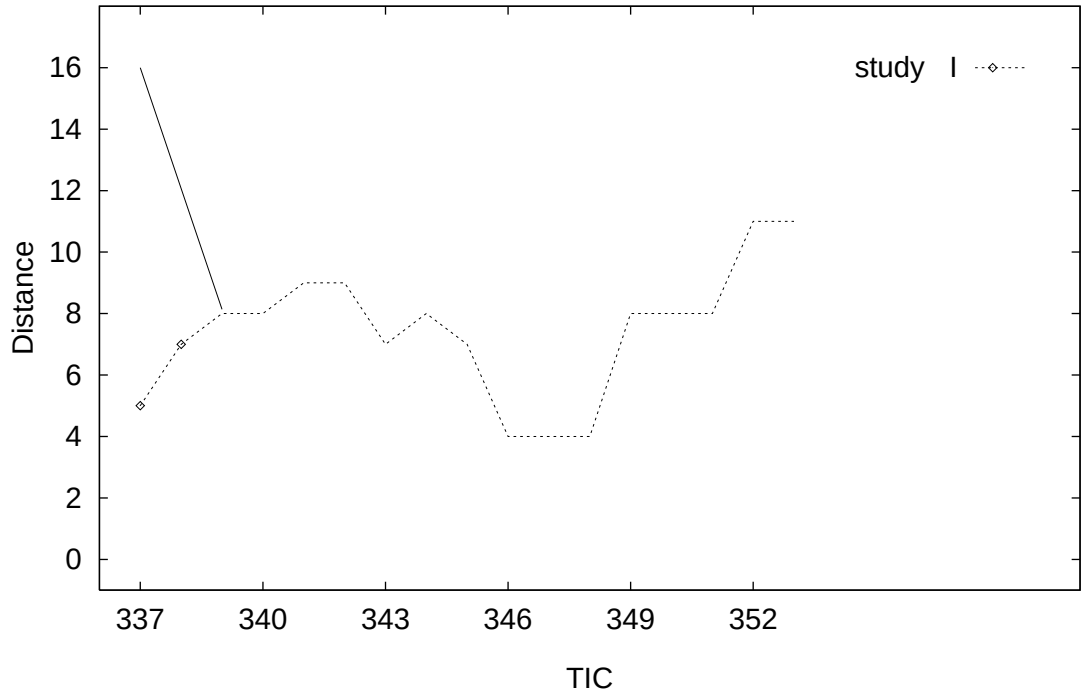


Figure 5.25. Classifications of the ninth instance of theme (TICs 313 – 330) relative to theme. The instance occurs in the first voice, concurrently with two other voices. The instance is a perfect copy of the theme. The classifications of the instance yielded by model II only converge to that of the theme in the fourth and fifth studies.




8
6
4
2
0

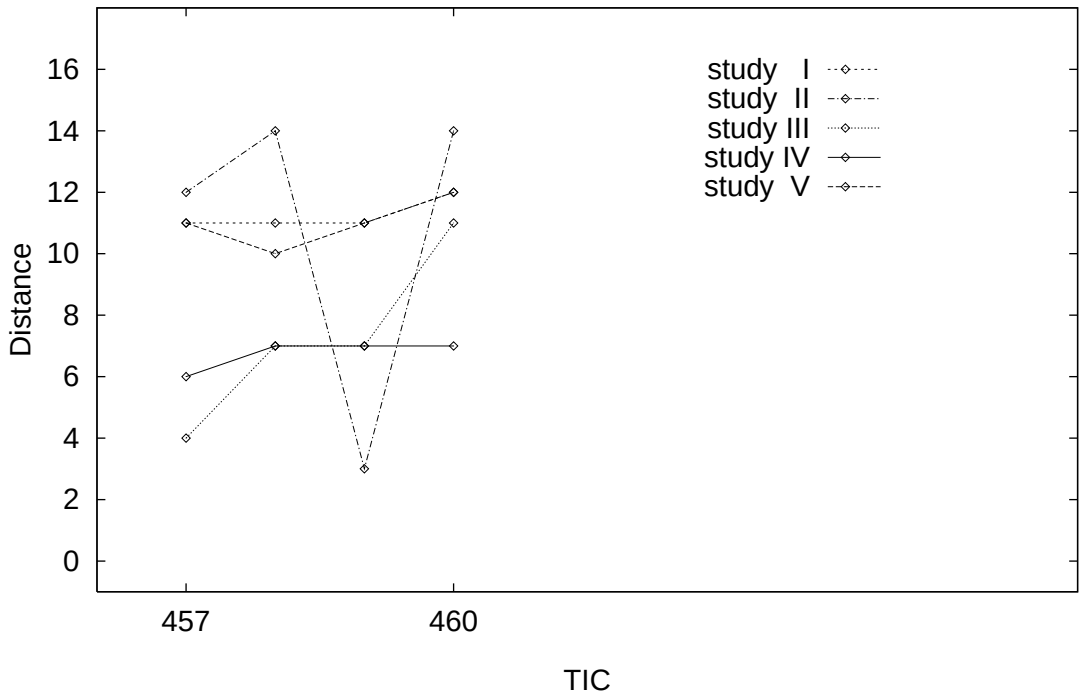
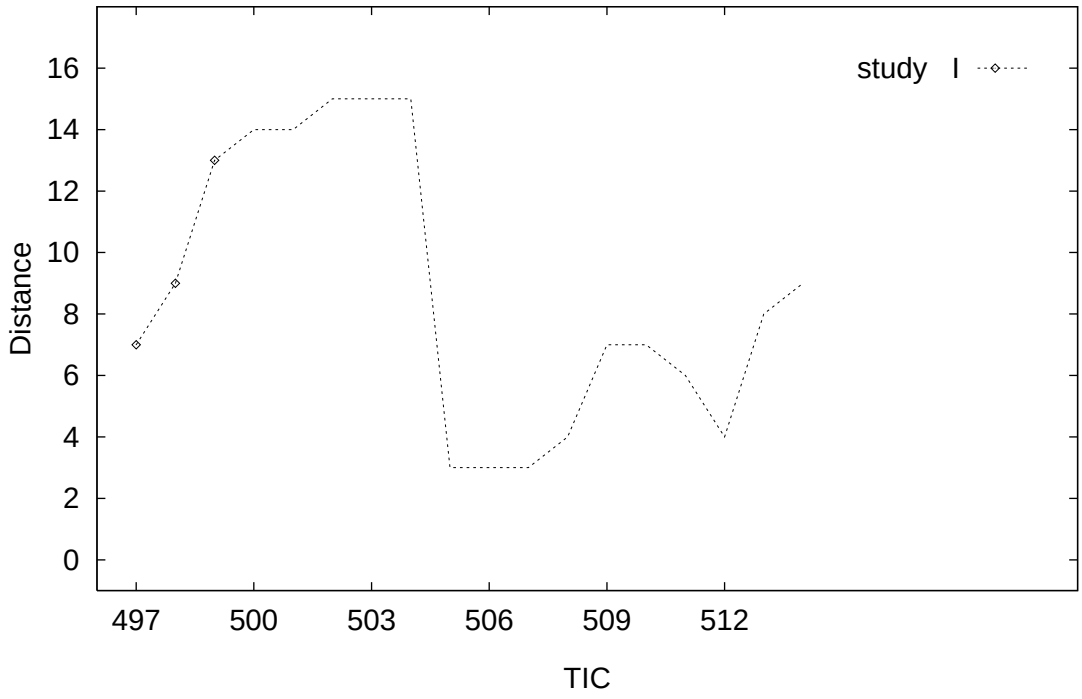


Figure 5.30. Classifications of the fourteenth instance of theme (TICs 457 – 460) relative to theme. The instance occurs in the first voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II do not converge to that of the theme in any of the studies.



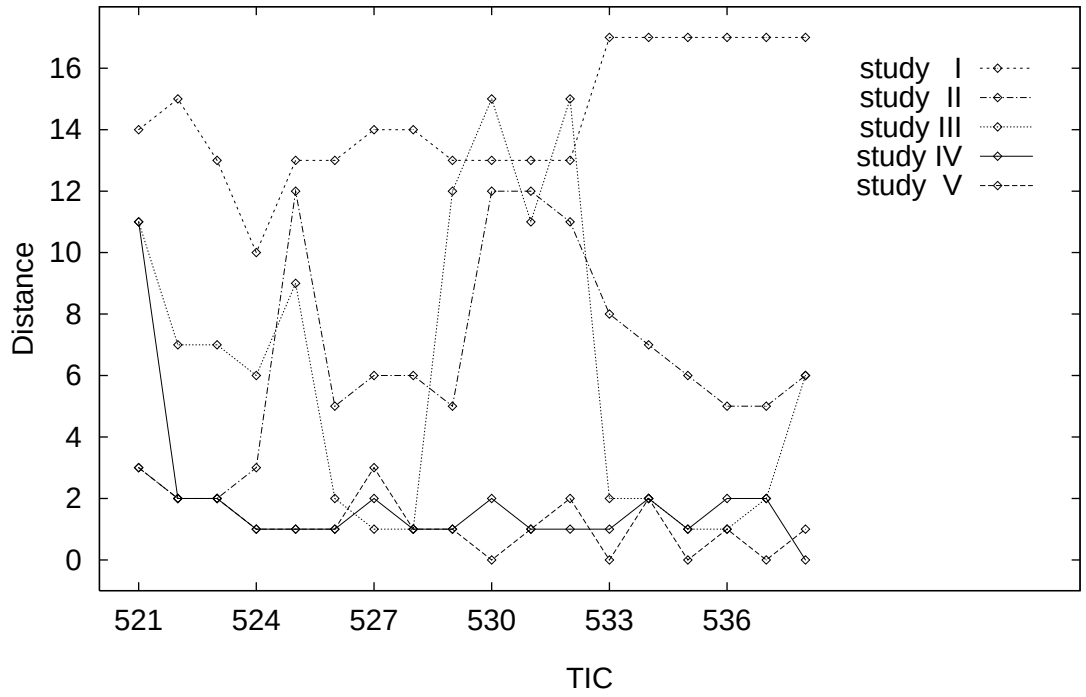


Figure 5.32. Classifications of the sixteenth instance of theme (TICs 521 – 538) relative to theme. The instance occurs in the second voice, concurrently with thre

Table 5.11. Classifications of model II (third experiment)

Study	No. Hits	No. Failures	Mean Error
I	0	16	170.50
II	3	13	85.50
III	4	12	73.75
IV	11	5	42.50
V	13	3	49.00

Table 5.12. Misclassifications of model II (third experiment)

Study	No. Minor Miscl.	No. Major Miscl.
I	2	5
II	0	6
III	0	14
IV	11	0
V	6	0

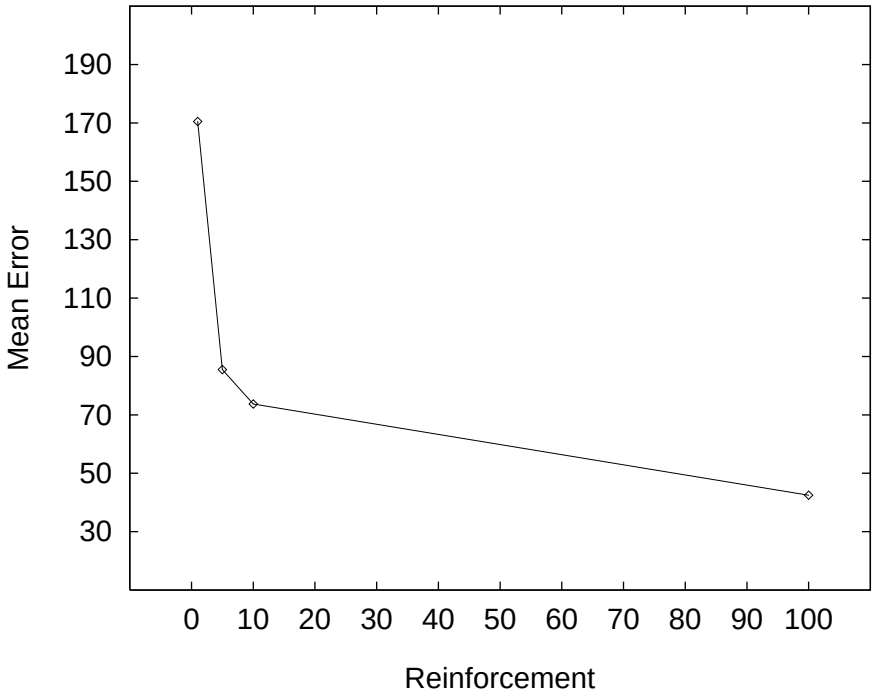


Figure 5.33. Mean error of classifications

Some conclusions may be drawn from the results. First, as displayed in table 5.12, the model held a high number of misclassifications in the third study. Such a high number was due to the fact that the model classified an intermediate part of theme as its final part, and consequently, kept on misclassifying intermediate parts of instances of theme as their final parts as well.

Second, by analysing the results displayed in the figures 5.17 to 5.33, and in the table 5.11, one may observe that the model was fault tolerant to errors. It classified properly several instances which differed slightly from the theme, whether in the pitch or in the duration of one single note. The model performed classification efficiently in the presence of noise as well. When instances of theme occurred concurrently with other polyphonic voices, the degree of noise was so high that it caused the model not to classify instances correctly. However, when thematic reinforcement was given to instances, the remaining polyphonic voices started playing roles of noisy backgrounds, and then, the model started classifying rightly instances of theme.

Third, as it may be observed in figures 5.23, 5.28, and 5.30, the model failed, in all studies, in recognizing three instances of theme. It succeeded, however, as shown in figures 5.24 and 5.29, in recognizing two other instances of theme in the last four studies. These instances, which occur between TICs 265 and 282, TICs 273 and 290, TICs 441 and 456, TICs 449 and 466, and TICs 457 and 460, overlapped through the voices, making up the two cases of *stretto* present in the fugue. One may conclude, therefore, that the recognition of *strettos* was not performed reasonably by the model.

Fourth, by comparing studies IV and V in tables 5.11 and 5.12, one may verify that there is not a significant difference between their results. The straightforward conclusion which may be drawn is thus, that thematic recognition in polyphony is not dependent upon segmentation, and consequently, the latter does not facilitate the former.

Finally, by observing the results in tables 5.11 and 5.12, one may conclude that reinforcement does facilitate thematic recognition in polyphony, and thus, listeners might rely heavily upon it in order to recognize properly instances of theme. In a real situation, reinforcement might be provided under two forms. It might be provided by performers. Indeed, as mentioned in section 2.4.6, Kirkpatrick suggests that pianists reinforce notes of the theme in polyphonic music by playing them louder. Alternatively, reinforcement might be provided by memory mechanisms in the brain. As described in section 2.4.1 on experiments with interleaved melodies, listeners are able to recognize a familiar target melody amid overlapping voices if the target is prespecified. Therefore, by memorizing the theme of a fugue, which occurs unaccompanied at the beginning of the fugue, listeners would be able to identify its instances whenever they occur throughout the fugue.

5.9 Summary

The results obtained in this chapter hold implications into two distinct fields. Firstly, the chapter introduces original representations for musical sequences, and an original artificial neural model for thematic recognition. The model has a topology made up of two self-organizing map networks, one on top of the other. It encodes and manipulates context information effectively, and that enables it to perform sequence classification and discrimination efficiently (Carpinteiro & Barrow, 1996). The model has application in domains which demand classifying either a set of sequences of vectors in time or sub-sequences into a unique and large sequence of vectors in time.

Secondly, by assuming the artificial neural model as a plausible model for the field of music perception, the chapter then presents results which are very relevant to that field. Experiments with the model when applied to thematic recognition in musical domains led us to important conclusions. First, segmentation facilitates thematic recognition when carried out on an univoiced musical domain. Second, thematic recognition is particularly difficult when performed on passages containing *strettos*. Third, thematic recognition in polyphony is not dependent upon segmentation. Last, listeners might rely heavily upon reinforcement in order to carry out thematic recognition.

Such reinforcement might thus be provided either by performers or by memory mechanisms in the brain.

Chapter 6

Conclusion

6.1 Summary

The aim of the current research was to develop a connectionist model to investigate, along with other related issues, the role of segmentation and thematic reinforcement in thematic recognition in

The model holds a hidden layer and an output layer, which consists of two output units. The model was trained to display activation values (10) in these units when the window in the input layer is representing a negative pattern, that means, a rhythmic pattern which is not a case of segmentation. It was also trained to display values (01) when the window is representing a positive pattern, a rhythmic pattern which is a case of segmentation.

6.1.1.3 Experiments

Three experiments were carried out. In each, the model was trained and tested on sets of contrived patterns, and applied to six musical pieces from J. S. Bach.

The training method followed the same procedure in all experiments. For each experiment, four sets of contrived patterns were generated using pattern templates. The first of the three sets was training set. Periodically, training was halted, and the model was tested on second set. When total error stopped decreasing, training was ended, and the model was tested on third set. We could thus evaluate different net configurations to find the optimum number of hidden units. Principal component analysis was performed on the activations of the hidden units given by each pattern in the fourth set.

The first experiment was on recognizing cases of segmentatio

6.1.2 Thematic recognition stage

6.1.2.1 Background

Little research has been carried out in order to understand t

their former activation values decayed in time. Each input unit represents locally one musical interval ranging from an octave down to an octave up. When there is a rest, none of the input units receives activation. Otherwise, when a note is onset or sustained, the unit corresponding to the interval receives activation.

6.1.2.2 Model

The original unsupervised model is an extension of Kohonen's (1989) self-organizing map (SOM). It holds a hierarchical topology made up of two SOMs — one on top of the other. Two time integrators — one for each SOM — are applied to units in the input layers of the SOMs.

The hierarchical topology united with time integrators enables the model to encode and manipulate context information efficiently, which is manifested through a high computational power in terms of sequence classification and discrimination. The representations in the input layer of the top SOM include context information. Such representations are not handmade beforehand, but instead, they are built up by the bottom SOM. The advantage of this approach is twofold. First, one does not need to worry about encoding context once the bottom SOM is in charge of making an internal representation of context in its map. Second, only the representations required by the application will be built up by the bottom SOM reducing thus, the necessary number of units in the input layer of the top SOM.

As the original unsupervised model has two SOMs, it is referenced as model II in the dissertation. To be better evaluated, the performance of the model II was compared to that of the model I. Model I is a neural model which holds a topology made up of one SOM, and a time integrator applied to units in the input layer of the SOM.

6.1.2.3 Experiments

Three experiments were carried out. In all of them, the training method followed the same procedure. The two SOMs of model II and the SOM of model I were trained in two phases — coarse-mapping and fine-tuning. The learning rate was set to an initial value, and the size of the neighbourhood was set to the size of the map in the coarse-mapping phase. Both the learning rate and the radius of the neighbourhood were reduced linearly. In the fine-tuning phase, the learning rate and the radius were kept constant. The coarse-mapping phase took 20%, and the fine-tuning phase took 80% of the total number of epochs.

The first experiment was on mapping a set of sequences. In this, the model was applied to a small scale problem in order to analyse its behaviour. The second and third experiments were on thematic recognition on an univoiced musical sequence and on a polyphonic musical sequence respectively.

Two input sets were used in the second experiment. The first consisted of a large and unique sequence of musical intervals, which corresponded to the third voice of the sixteenth four-part fugue in G minor of the first volume of The Well-Tempered Clavier of Bach (Bach, 1989). The second, in its turn, contained many sequences, which corresponded to groupings produced by segmentation given by rests applied to the third voice of the fugue. The models were trained and evaluated on these input sets.

Two input sets were used in the third experiment as well. The first consisted of a large and unique sequence of musical intervals, which corresponded to the sixteenth four-part fugue in G minor of the first volume of The Well-Tempered Clavier of Bach (Bach, 1989). The second, in its turn, contained many sequences, which corresponded to groupings produced by segmentation given by rests applied to the fugue. Model II was trained and evaluated on these input sets. Model I was not used in the third experiment for reasons of its poor performance on the previous experiment.

The fugue in G minor was chosen for several reasons. First, as many of fugues of Bach, it has four voices. Second, it possesses several perfect and modified instances of theme. Third, it includes two cases of *stretto* (see section 2.4.6). One case occurs between its seventeenth and eighteenth bars, in which two instances of theme overlap, and the other case occurs between its

the role of segmentation and reinforcement mechanisms in thematic recognition, so that results of the connectionist model could be better scrutinized and criticised in the light of those experiments.

Fourth, segmentation mechanisms were modelled for three cases of rhythmic segmentation only. Thus, the supervised model described in the fourth chapter could be extended to incorporate other cases of rhythmic segmentation, as well as cases of metric and melodic segmentation (see section 2.3.2).

Fifth, segmentation was performed separately on each voice of six Bach's polyphonic pieces, as described in the fourth chapter. However, segmentation could be performed concurrently in all polyphonic voices by having a dedicated supervised neural model for each voice. The output of these models could be manipulated by another neural model on the top, which would decide whether or not to segment. This proposed connectionist model consisting of a top neural model and a number of supervised neural models would be able to vary the degree of contribution of cases of segmentation occurring in each single voice to the final decision on carrying out segmentation, and as a consequence, would be a more complete model for segmentation in polyphonic music.

Sixth, as described in the fifth chapter, thematic reinforcement was performed by means of providing reinforcement in activation for units in the input layer of the unsupervised model. Reinforcement, nevertheless, could be performed by another artificial neural model in charge of providing extra activation for those units in cases of instances of theme. The artificial neural model would thus be modelling memory traces of fugal themes in the brain.

Finally, the unsupervised model described in the fifth chapter could be better explored. In spite of the good results, it is still open to further research. In principle, the model could have any number of self-organizing map nets — the more nets, the more similar and longer the sequences of vectors in time which could be recognized.

Bibliography

Adams, C. S. (1982a). Organization in the two-part inventions of John Sebastian Bach (part I). *Bach*, *13*(2), 6–16.

Adams, C. S. (1982b). Organization in the two-part inventions of John Sebastian Bach (part II). *Bach*, *13*(3), 12–19.

Anderson, J. R. (1990). *Cognitive Psychology and Its Implications* (Third edition). W. H. Freeman, New York.

Attneave, F., & Olson, R. K. (1971). Pitch as a medium: a new approach to psychophysical scaling. *American Journal of Psychology*, *84*, 147–166.

Bach, J. S. (1970). *Inventionen und Sinfonien*. BWV 772–801. Bärenreiter Kassel, Basel, Germany.

Bach, J. S. (1989). *Das Wohltemperierte Klavier*. Vol. 1. BWV 846–869. Bärenreiter Kassel, Basel, Germany.

Beard, K. (1985). Exploring the two-part inventions. *Clavier*, *24*(3), 18–21.

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: a connectionist framework. *Music Perception*

Fahlman, S. E. (1991). The recurrent cascade-correlation architecture. Tech. rep. CMU-CS-91-100, School of Computer Science — Carnegie Mellon University, Pittsburgh, PA.

Flindell, E. F. (1983). Apropos Bach's inventions (part I). *Bach*, *14*(4), 3–14.

Flindell, E. F. (1984). Apropos Bach's inventions (part II). *Bach*, *15*(1), 3–16.

Francès, R. (1988). *The Perception of Music*. LEA. Translated by W. J. Dowling.

Fux, J. J. (1971). *The Study of Counterpoint from Gradus ad Parnassum*. W. W. Norton, New York. Translated and edited by A. Mann.

Gabrielsson, A. (1973). Similarity ratings and dimension a

Mozer, M. C., & Soukup, T. (1991). Connectionist music composition based on melodic and stylistic constraints. In Lippmann, R. P., Moody, J., & Touretzky, D. S. (Eds.), *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 3, pp. 789–796. Morgan Kaufmann.

Serafine, M. L., Glassman, N., & Overbeeke, C. (1989). The cognitive reality of hierarchic structure in music. *Music Perception*, *6*(4), 397–430.

Taylor, I., & Greenhough, M. (1994). Modelling pitch perception with adaptive resonance theory artificial neural networks. *Connection Science*, *6*(2&3), 135–154.

Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America*, *55*(5), 1061–1069.

Todd, P. M. (1991). A connectionist approach to algorithmic composition. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 173–194. The MIT Press, Cambridge, MA.

Vos, P. G. (1977). Temporal duration factors in the perception of auditory rhythmic patterns. *Scientific Aesthetics*, *1*(3), 183–199.

West, R., Howell, P., & Cross, I. (1991). Musical structure and knowledge representation. In Howell, P., West, R., & Cross, I. (Eds.), *Representing Musical Structure*, chap. 1, pp. 1–30. Academic Press.

White, B. W. (1960). Recognition of distorted melodies. *American Journal of Psychology*, *73*, 100–107.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior.