A Connectionist Approach in Music Perception

Otávio Augusto Salgado Carpinteiro

CSRP 426

July 1996

ISSN 1350-3162

UNIVERSITY OF

SUSSEX

AT BRIGHTON

Cognitive Science

Research Papers

In memory of

my grandma Ana

When you are old and grey and full of sleep,
And nodding by the fire, take down this book,
And slowly read, and dream of the soft look
Your eyes had once, and of their shadows deep;

How many loved your moments of glad grace,
And loved your beauty with love false or true,
But one man loved the pilgrim soul in you,
And loved the sorrows of your changing face;

And bending down beside the glowing bars,
Murmur, a little sadly, how Love fled
And paced upon the mountains overhead
And hid his face amid a crowd of stars.

(When You Are Old, W. B. Yeats)

Mar bravo, que em ondas despontas,
Bramindo em cadência invulgar,
Arrastas contigo o que encontras
E, enfim, as areias a par.

Ó tempo, que em ondas dolentes,
Corróis o futuro ao passar,
Por que tua angústia consentes
Em meu coração assentar?

Agitas minha alma exaurida,
Arrancas, de mim, meu sonhar
E levas, por fim, minha vida,
Tal como às areias, o mar …

(Mar do Tempo, O. A. S. Carpinteiro)

Contents

List of Figures

List of Tables

Acknowledgements

I thank Capes for the partial financial support.

I thank my colleagues in COGS, in special Guillaume Barreau, Hilan Bensusan, Julian Budd, Paulo Costa, Changiz Delara, Stephen Eglen, Vicente Guerrero-Rojo, Philip Jones, Ulysses de Oliveira, Rafael Perez y Perez, Margarita Sordo-Sanchez, and Rosemary Tate.

I thank Jackie Dugard and Linda Thompson from COGS secretarial office, and Catherine Bancroft-Rimmer from the accommodation office.

I thank the members of my committee — Jonathan Cross, and Christopher Thornton.

Several people extended to me a warm friendship, which made my life in England less hard. Among them, Godela and Raul Abreu, Ana and Aluízio Araújo, Fernando Baldi, Orlando Coelho, Alexandre Direne, Valéria Judice, Ana and Márcio Maia, Fernando Nasser, Cristina and Rogério Quintella, Monica Sacret, Mônica Salvagnini, Conceição and Hugo Santos, Jaísa de Souza, and Célia and Marcelo Zambrano. Thank you all.

Many thanks to Pavel Horák, Blanka Horáková, Stuart Paul, Thaís and Maurício Piccinini, Michael Pople, Salete and Oscar Rotava, and Petra Smejkalová, who were caring and supportive friends in difficult times.

I wish to convey my gratitude to Fr. Pedro Barreto, Fr. Sean Finnegan, and Fr. Paul Hayward. I am indebted to Ricardo Martins for his assistance with a few affairs in Brazil.

A special thank to my family — my father, my brother and sister-in-law, my sister, my godson, and my grandma, who foresaw that I would not see her again. I owe my parents the accomplishments which I have achieved in my life.

The last words are directed to my supervisor Harry G. Barrow, whom I learnt to respect as a scientist and human being. His advice and encouragement were fundamental on several occasions, when my work seemed to be in a real deadlock. I owe him much, and wish heartily to thank him.

I dedicate the dissertation to the memory of my grandma Ana.

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.

The research carried out in this dissertation contributes into two distinct fields. Firstly, it contributes to the field of artificial neural networks. The original unsupervised model encodes and manipulates context information effectively, and that enables it to perform sequence classification and discrimination efficiently. It has application in cognitive domains which demand classifying either a set of sequences of vectors in time or sub-sequences within a unique and large sequence of vectors in time. Secondly, the research contributes to the field of music perception. The results obtained by the connectionist model suggest, along with other important conclusions, that thematic recognition in polyphony is not facilitated by segmentation, but otherwise, facilitated by thematic reinforcement.

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 or greater extent, how themes are presented, contrasted, repeated, and developed. In fugal form, for instance, themes[2] are initially presented in each polyphonic voice in turn, and subsequently, contrasted, repeated, and developed throughout the voices. Hence, music acquires coherence, form, and meaning primarily through continuous manipulation of thematic ideas.

Recognition of musical forms in a polyphonic domain thus involves thematic recognition, that is, recognition of instances of theme through polyphonic voices, whether they appear unaccompanied, transposed, altered or not. Little research has been carried out in order to understand the mechanisms underlying the perception of polyphonic music, and consequently, important issues relating to thematic recognition in the polyphonic domain are still open to investigation. One of those issues, in particular, is the question of whether or not cognitive mechanisms of segmentation and thematic reinforcement facilitate thematic recognition in polyphonic music.

1.2 Introduction to musical segmentation and thematic reinforcement

It is believed that listeners do not grasp a musical piece in its entirety, but rather, they segment it into parts, which can be more easily analysed. One important reason for listeners to perform

[1] In the dissertation, music refers to as Western tonal music.

[2] In the fugal domain, *theme* is known as *subject*.

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.

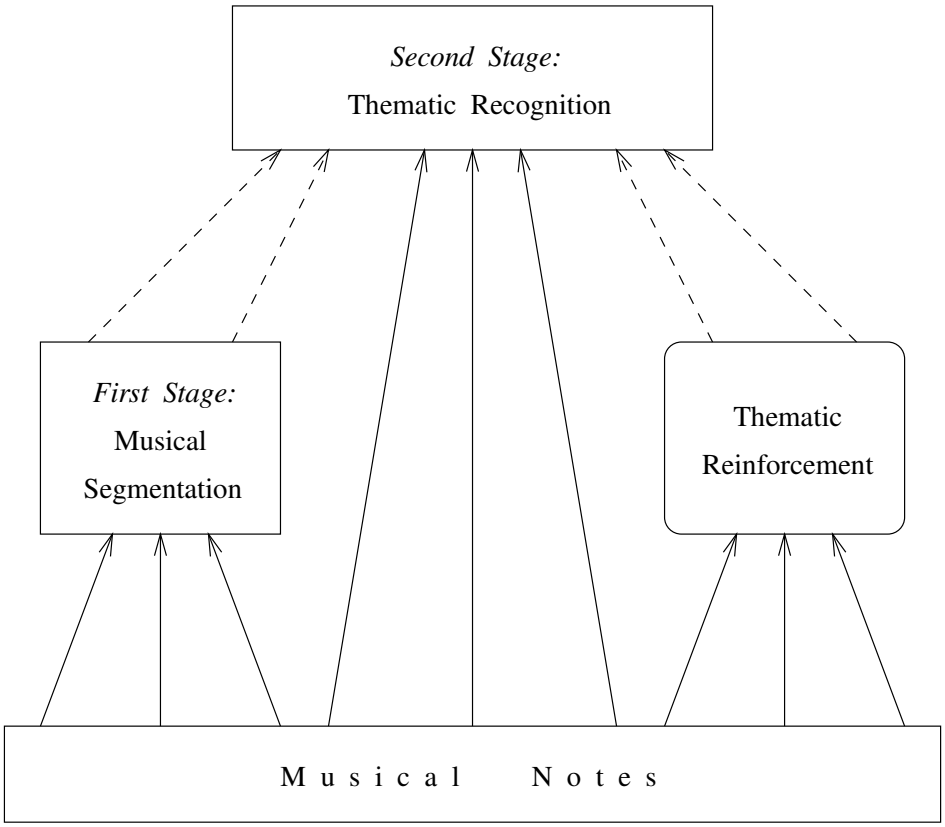


Figure 1.1. The connectionist model

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 in thematic recognition in polyphony. Lerdahl and Jackendoff's theory of music understanding is presented, and cases of rhythmic segmentation are detailed. A little introduction to music is provided as well.

The third chapter provides a review of the literature in connectionist models. Models of chord perception, pitch perception, tonal perception, and of musical pattern classification are presented. Four supervised and three unsupervised artificial neural models of sequence classification in time are also described.

The fourth chapter introduces the first stage, that is, the supervised artificial neural model to segment musical pieces according to three cases of rhythmic segmentation. In it, the training of the model on contrived sequences and assessment on real music are detailed. A novel representation for rhythmic sequences is proposed.

The fifth chapter introduces the second stage, that is, the unsupervised artificial neural model to perform thematic recognition. The training and evaluation of the model on binary sequences and on real music are detailed. A novel representation for musical sequences is proposed as well.

Finally, the sixth chapter summarizes the research, enumerates its contributions, and outlines some directions for further work.

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 Contour and interval representations

Marr (1982, p. 6) declares that "modern representational theories conceive of the mind as having access to systems of internal representations; mental states are characterized by asserting what the internal representations currently specify, and mental processes by how such internal representations are obtained and how they interact". Several representational theories have been proposed to describe the ways a listener constructs an internal representation of a melody.

Individuals with absolute pitch must obviously possess a kind of internal representation which includes absolute pitches. However, this kind of representation does not seem to be relevant in recognition of melodies. Indeed, the ability that both individuals with absolute or relative pitch have to recognize a melody irrespective of the key in which the melody is heard suggests that individuals should represent the melody not as a sequence of absolute pitches, but either as a sequence of intervals or as a contour — a sequence of ups and downs — or both. This section contains a review of studies supporting contour and interval representations.

2.2.1 White's experiment

White (1960) performed an experiment on a set of 10 well-known melodies in order to verify whether musical information is carried in the sequence of intervals between adjacent notes. Each melody was subjected to 12 different transformations. Some of these transformations were linear, as to multiply, add or subtract all intervals from an integer. Some were non-linear, as to set all intervals equal to 1 but maintaining contour, or to set all intervals equal to 0 but keeping rhythmic information intact.

Linear transformations are least disruptive for they preserve the relative size of the intervals as well as rhythm and contour. The 12 variations on each of the 10 melodies together with the

original versions made a total of 130 selections. The selections were then presented to subjects in order to be identified.

The results showed that linear transformations were more easily identified than non-linear transformations. The most difficult transformations to be identified were temporal reversals of the melodies, and those which set all intervals equal to 0 but leaving rhythmic information unchanged. Thus, White concludes that his results support "the supposition that subjects depended heavily on interval-information in identifying melodic patterns".

2.2.2 Dowling and Fujitani's experiments

Dowling and Fujitani (1971) carried out two experiments to explore the role of contour in memory for melodies. The first experiment consisted in presenting pairs of five-note melodies. The experiment involved three distinct tasks. In the first, subjects were asked to decide whether the second melody was an identical or a randomly modified version of the first. In the second task, subjects had to decide whether the second melody was an identical version or a modified but same-contour version of the first melody. In the third, subjects were asked to check whether the second melody was a modified but same-contour version or a randomly modified version of the first melody. The second melody was presented either untransposed or transposed.

Subjects were asked to recognize either undistorted or distorted versions of familiar folktunes in the second experiment. The versions were distorted in three different ways. In the first, both contour and relative sizes of successive intervals were preserved. In the second way, only contour was preserved, and in the third, only the first note of each measure was left unchanged.

The first experiment explored the role of contour in short-term memory because subjects were given contrived melodies. On the other hand, the second experiment involved long-term memory for the subjects were familiar with the melodies.

Dowling and Fujitani drew some conclusions from the results of the first experiment. The first task was the easiest among the three, and all them were generally harder when the second melody was transposed. Performance in the first task was better than in the third task when the second melody was not transposed, and so, they concluded that "subjects were mainly using recognition of pitches in solving the untransposed tasks". On the other hand, with transposition, "contour seems to provide the basis for recognition" because subjects performed both tasks one and three with the same proficiency. Moreover, "recognition seems not to be dependent on recognition of exact interval sizes". This was shown by the fact that performance in the second task, which could be based on the discrimination of interval sizes, was poor.

Dowling and Fujitani drew some conclusions from the results of the second experiment as well. Recognition of undistorted versions was performed much better than recognition of distorted versions which preserved contour and relative sizes of successive intervals. So, "it appears that subjects remember more about tunes they recognize than just the contour and relative interval sizes". Therefore, "subjects appear to have good long-term memory for exact interval sizes in the context of familiar tunes".

2.2.3 Attneave and Olson's experiment

Attneave and Olson (1971) came up with similar conclusions. In one of their experiments, subjects were asked to transpose a well-known musical pattern — the NBC chimes — made up of three tones. Subjects were given an anchor tone which corresponded either to the first, or to the second, or to the third tone of the pattern. Their task was then adjust the other two tones to make the pattern sound like the NBC chimes.

From the results, Attneave and Olson conclude that "the long-term memory trace is encoded in terms based upon relations, or intervals, or distances on a log frequency scale". Pitches and key might provide additional information to memory trace, yet "it is evident that normal individuals do not, in fact, preserve this kind of information with any high degree of precision. If they did, absolute pitch would be commonplace".

2.2.4 Dowling's experiments

Dowling (1971) performed a set of experiments on subjects to verify their performance in recognition of melodic inversions. The experiments consisted in presenting a five-note melody which was followed, after an interval of time, by a comparison melody. The second melody was either a transposition or an inversion of the first, and could also be either an exact interval-size-preserving comparison or a contour-preserving comparison.

The experiments were divided into three tasks. In the first, subjects were asked to distinguish whether the comparison melody preserved either contour or exact interval sizes. The second task required subjects to distinguish between comparisons which either preserved contour or were different. In the third, subjects were asked to distinguish whether the comparison melody either preserved exact interval sizes or was different.

Based on the fact that recognition of inversions was clearly better than chance, Dowling concluded that inversion is actually perceived by listeners. Furthermore, as the performance of subjects when the comparison melody preserved exact interval sizes was similar to that when the comparison preserved just contour, he concluded that recognition of transpositions and inversions "seems to be mainly on the basis of contour".

Dowling (1972) was concerned with recognition of melodic transformations — inversion, retrograde, and retrograde inversion — and was therefore, an extension of Dowling (1971). He had two main motivations in this study. First, to verify "whether transformations of melodic material can function as an actually perceived aspect of musical structure or whether they must be relegated to the category of merely intellectual conceits of the composer". Second, to understand the "kind of process listeners use when they recognize these transforms". The process listeners use might be characterized either as operations on intervals or as operations on pitches.

Dowling shows that inversion and retrograde transformations are performed in one operation on the vector of pitches, while retrograde inversion is performed in two operations on that vector. On the other hand, retrograde transformation is carried out in two operations on the vector of interval sizes, while inversion and retrograde inversion are carried out in only one operation on that vector. Thus, if listeners have better performance when recognizing retrograde transformations, that would suggest that they work on the vector of pitches, for such type of transformation requires only one operation on such vector. Alternatively, for analogous reasons, if they are better in recognizing retrograde inversion transformations, then they should operate on the vector of interval sizes.

The experiment consisted in presenting pairs of melodies — a standard melody followed by a transformed one. Subjects were then asked to recognize which type of transformation was accomplished on the second melody. The second melody could preserve either the exact intervals of the transformation or only contour of the transformation.

The "results clearly demonstrate that inversions, retrogrades, and retrograde inversions of brief melodies can be recognized with better than chance accuracy". Therefore, "such manipulations of melodic material are perceptually accessible to the listener". Moreover, Dowling states that such transformations would be much more accessible to listeners in a real musical context in spite of the simplicity of the contrived melodies used in the experiment. He gives some reasons to support this statement. First, the atonal nature of the melodies he used which is "typically difficult to deal with". Second, the intervals used in the contrived melodies are much smaller than those encountered in real music. Third, in real music, "there is a rhythmic dimension present that was avoided in the experiment". Fourth, the melodic material in real music usually helps in distinguishing between separated melodies, while in the experiment the material was selected to provide maximal homogeneity.

Dowling could draw two more conclusions from the results of the experiment. First, subjects were better in recognizing the type of transformation applied on the second melody when the second melody did not preserve the exact intervals, but preserved only contour of the transformation. Thus, he concludes that "Dowling's (1971) result that exact interval size information becomes

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 and lures. Targets were copies of the excerpts in the list. Related items consisted of excerpts taken from the quartets which preserved contour of the excerpts in the list but differed in interval sizes. Lures were excerpts from the quartets which preserved neither contour nor interval sizes of the excerpts used in the list. The test procedure was done in two steps. First, subjects heard the whole list of excerpts, and then, after a five minute interval of time, comparison melodies — targets, related items, and lures — were presented.

Dowling and Bartlett were surprised by the results. Contrary to their expectations, "recognition of targets was much better than recognition of relateds, and recognition of relateds was barely better than chance". So, they performed another experiment to verify whether the results obtained in the two first experiments, which used real musical excerpts, could also be attained by using contrived tonal material. Thus, results could not be attributed to the complexity of actual music. Notwithstanding, the results of the third experiment confirmed those from previous experiments.

The last experiment used the same contrived tonal material as the third experiment. Standard and comparison melodies were presented in two pairs — one pair within the other. The outside standard was presented first. The inside standard was then presented after a ten second pause. After a five second pause, the inside comparison, and finally after a ten second pause, the outside comparison. Thus, they could verify the role of contour and interval information in short-term and long-term memory situations.

Based on the results of the last experiment, Dowling and Bartlett concluded that "contour information is easily extracted from novel musical stimuli, but contributes to performance only with short ... retention [time] intervals. In contrast, interval ... information is difficult to extract, but contributes more or less equally to performance over a broad range of retention intervals". Also, on the account of the results from the four experiments, they suggest that melodies might be retained in the form of two separate memory traces — a short-term trace representing absolute pitch, contour, and key, and a long-term trace representing precise interval information.

2.2.6 Edworthy's experiments

Edworthy (1985) has shown the dichotomy between contour and interval information as well. She performed a set of six experiments on subjects, who were all musicians.

The experiments were divided into two tasks — an interval task and a contour task. In the interval task, subjects were asked to concentrate on the intervals of a standard melody. After a five

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 the necessary outline for encoding interval information".

2.3 Segmentation and grouping

2.3.1 Experiments supporting segmentation and grouping

Gabrielsson (1973) performed a set of ten experiments to investigate how listeners evaluate rhythmic patterns. The patterns were presented in pairs, and subjects had to rate the degree of similarity or dissimilarity of the rhythms within each pair, and also, to provide verbal descriptions of their evaluations. Among other conclusions drawn from the results, Gabrielsson could verify that variations in intensity at different beats of the measure led subjects to perceive the rhythms as segmented. Moreover, the first events of rhythmic groups were perceived as accented.

Handel (1974) carried out two experiments to study the perception of repeating auditory patterns. In the first experiment, a pattern formed of eight elements, each element being a binary event, was repeated continuously in a time sequence. Frequency and intensity were two of these binary events. Thus, frequency patterns were composed of eight elements where each element could be either a low frequency of 800 Hz or a high frequency of 1300 Hz. In intensity patterns, on the other hand, each element could be either a low intensity of 60 db or a high intensity of 75 db. Subjects listened to a sequence of a repeating pattern for 60 seconds, and were then asked to identify the repeating pattern. By analysing the results, Handel could observe that such binary

variations in intensity or in frequency enabled subjects to segment properly the sequence, and so, to identify the pattern which was repeated in time.

Vos (1977) performed two experiments to investigate the perception of repeating rhythmic patterns. Seven different stimuli were applied in each experiment. Each stimulus consisted of a sequence of 15 repetitions of a fixed pattern. In the first experiment, each pattern was made up of two tones separated by two silent intervals. In the second, each pattern consisted of three tones separated by three silent intervals. The duration of tones and silent intervals, that is, the duration between onsets and offsets, differed from pattern to pattern. Subjects were asked to listen to each sequence, and to identify the corresponding repeating pattern. Based on the results, Vos verified that variations in duration between onsets and offsets led subjects to segment properly the sequences, and consequently, to perceive the repeating patterns.

As evidenced from above, variations in intensity, frequency, or in duration between onsets and offsets cause segmentation in sequences. Not only intensity, frequency, and duration between onsets and offsets, but also variations in interonset timing[1] (Garner & Gottwald, 1968), in sound quality (Deutsch, 1986), and in timbre (Deutsch, 1982) induce segmentation in sequences. In fact, as Drake and Palmer (1993) point out, studies of segmentation in sound sequences suggest that "a change in any sound parameter leads to the perception of a break in the sequence and thus the creation of groups separated by the change". Furthermore, it is worth noticing that all these means of segmentation have the Gestalt principles of proximity and similarity as a basis.

2.3.2 Perceptual accents

Segmentation provokes the emergence of accents. An accent is "an event that stands out and captures a listener's attention" (Drake & Palmer, 1993), or in other words, "anything that is relatively attention-getting" (Jones, 1987). As mentioned above (section 2.3.1), Gabrielsson verified that the first event of a rhythmic group is perceived as accented. Jones (1987) claims, in its turn, that both the first and last events of a rhythmic group are perceived as accented. Apart from accents provoked by rhythmic grouping, there are melodic and metric accents as well. Melodic accents arise from segmentation caused by variations in melodic contour. Metric accents arise from segmentation produced by metre, which divides music into segments of equal duration — the bars.

Drake and Palmer (1993) performed three experiments to investigate whether performers confer additional emphasis to three accent types — rhythmic grouping, melodic, and metric accents — by playing them louder, and longer, for instance. In doing so, performers would help listeners to segment properly musical sequences. In the first experiment, pianists played musical sequences which contained a single accent type whereas in the second experiment, they played more complex musical sequences containing accent types which could either coincide or conflict. In the third experiment, a concert pianist played a Beethoven sonata containing coinciding or conflicting accent types.

Drake and Palmer drew three main conclusions from the results. First, performers do emphasize accents. It is worth mentioning that, as a concert pianist, Kirkpatrick (1984) has indirectly proposed to emphasize rhythmic grouping accents by recommending segmentation of certain kinds of rhythmic sequences in music. Second, "the last event in rhythmic groups was emphasized", that means, "performers played it louder, delayed, and preceded by a pause". Third, rhythmic grouping was the most important accent type, for rhythmic grouping accents dominated other accent types when combined, that is, they remained unaffected by the presence of other accent types.

2.3.3 Reasons for performing segmentation and grouping

A question that arises is why listeners segment music, and consequently, do grouping. In the visual domain, Anderson (1990) has stated that "recognizing a complex pattern involves a feature analysis in which the overall pattern is decomposed into a set of primitive features, the individual

[1] Interonset timing is the time span between two contiguous onsets.

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.

Two of Dibben's (1994) experiments pursued identical objective of that of Serafine, Glassman, and Overbeeke's (1989). In the first, musicians heard a musical extract, and then were given three pairs of reductions to identify, in each pair, which was the reduction which best matched the extract. One of the pairs contained the correct reduction and a foil reduction. The other two pairs contained foil reductions only. In the second experiment, musicians were asked to rate the reductions and foil reductions according to their coherence as examples of tonal chord sequences. Dibben performed this experiment in order to verify if, in the first experiment, the musicians were matching the reductions to the extract based on the degree of similarity between the reductions and the extract in terms of the reductional process itself, or if they were doing so based on the degree of similarity between the reductions and the extract in terms of harmonic coherence.

From the results obtained in both experiments, Dibben concludes that "hierarchic structure is perceived: subjects appear to have access to a hierarchically organized representation of each extract". This conclusion suggests that "tonal works are heard in terms of a hierarchy of events of the sort proposed by Lerdahl and Jackendoff".

2.3.4.3 The grouping structure of the theory

The *grouping structure* of Lerdahl and Jackendoff's theory is one of the four structures on which the theory is built. It expresses a hierarchical segmentation of the piece into motives, phrases, and sections. It is made up of two sets of rules — grouping well-formedness rules, and grouping preference rules.

Grouping well-formedness rules specify the formal structure of grouping patterns. They define which kind of grouping patterns are formally possible, and which are not. *Grouping preference rules* designate, among the formally possible grouping patterns, those which correspond to the listener's actual intuitions.

Three of the grouping preference rules — segmentation by rests, by longer durations, and by breaks of similarity — held a particular interest and will be focused here, for they came into the domain of our experiments. They are cases of rhythmic segmentation, and can be found in Drake and Palmer (1993), Kirkpatrick (1984), and Lerdahl and Jackendoff (1983a). The Gestalt principles of proximity and similarity underlie the three grouping preference rules, and thus, it is convenient to make use of their visual analogies.

2.3.4.4 Segmentation by rests and longer durations

Let us consider figure 2.1. The two left most circles are perceived as making up a group whereas the right circle stands as separate. Relative distance in terms of spatial proximity is the principle behind this kind of grouping.

Figure 2.1. Proximity (visual analogy)

Segmentation by rests exemplified in figure 2.2a and *segmentation by longer durations* exemplified in figure 2.2b are auditory analogues to the visual grouping. The two left most notes group together, whereas the others stand as separate. Relative distance in these examples means the interval of time between two auditory events. It follows the definition of segmentations by rests and longer durations.

> **Rule: segmentation by rests**
> Consider a sequence of four notes $n_1 n_2 n_3 n_4$. Then, the transition n_2–n_3 may be heard as a group boundary if the interval of time from the end of n_2 to the beginning of n_3 is greater than both the interval of time from the end of n_1 to the beginning of n_2, and that from the end of n_3 to the beginning of n_4.

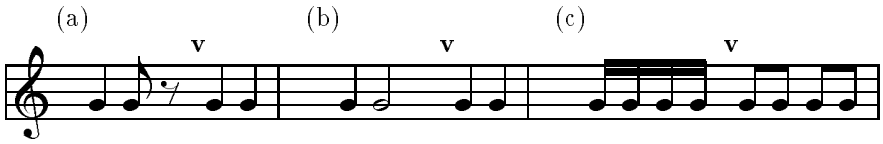


Figure 2.2. Three cases of rhythmic segmentation: (a) rests; (b) longer durations; (c) breaks of similarity;

Rule: segmentation by longer durations
Consider a sequence of four notes $n_1n_2n_3n_4$. Then, the transition n_2–n_3 may be heard as a group boundary if the interval of time between the attack points of n_2 and n_3 is greater than both the interval of time between the attack points of n_1 and n_2, and that between the attack points of n_3 and n_4.

2.3.4.5 Segmentation by breaks of similarity

Let us consider now figure 2.3. The four squares group together whilst the four circles are perceived as separate. Similarity, or change of it, is the principle underlying this kind of grouping.

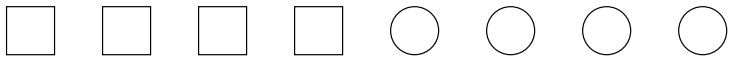


Figure 2.3. Similarity (visual analogy)

Segmentation by breaks of similarity exemplified in figure 2.2c is an auditory analogue to the visual grouping. The four left most notes group together, whilst the others stand as separate. In the example, therefore, change of similarity means change in duration of notes in different groups of notes. It follows the definition of segmentation by breaks of similarity.

Rule: segmentation by breaks of similarity
Consider a sequence of eight notes $n_1n_2n_3n_4n_5n_6n_7n_8$. Then, the transition n_4–n_5 may be heard as a group boundary if the durations of n_1, n_2, n_3, n_4 are identical, the durations of n_5, n_6, n_7, n_8 are also identical, and the durations of the first four notes are different from those of the last four.

2.4 Thematic recognition in polyphony

2.4.1 Dowling's experiments

Dowling (1973) performed four experiments on subjects to investigate the perception of interleaved melodies. In all experiments, the stimuli consisted of two melodies whose tones were interleaved. The melody which was to be perceived was named foreground melody. The other, which interfered with the foreground melody, was named background melody.

In the first experiment, the interleaved melodies were familiar, and their tones overlapped. The degree of overlapping was reduced progressively by transposing upwards one of the melodies. Subjects were asked to identify both the background and foreground melodies.

In the second experiment, as in the first, the tone range overlap was varied, yet the interleaved melodies were unfamiliar. It was used a short-term recognition-memory paradigm. Subjects heard firstly a standard non-interleaved melody. Then, after a two second pause, a comparison interleaved melody followed the standard. The comparison foreground melody was either identical to the standard or different. The subjects' task was, therefore, to recognize those comparisons which were identical to the standards.

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".

2.4.2 Dowling, Lung, and Herrbold's experiments

Dowling, Lung, and Herrbold's (1987) work is an extension of Dowling's, presented above in the section 2.4.1. Besides, it also aims at exploring perceptual effects of listener expectancies. A new metaphor — the expectancy windows — is introduced to this purpose. They argue that, when listening to a familiar melody, listeners focus their attention in regions where they expect the notes of the melody should occur. These regions, which vary in pitch and time around the notes, are named expectancy windows. Knowledge of pitch and time ranges at which notes of the melody are likely to occur enables thus a listener to carry out the task of recognizing a familiar melody when interleaved with other notes.

The first experiment made use of familiar target melodies interleaved with distractor notes. Target melodies were either on or off beat, and distractors either overlapped or not the notes of the melodies. Distractors could either belong or not to the key of the melodies. Subjects were told beforehand which target melody would be employed, and their task was to say whether they identified it. The results reveal that "the task was easier with distractors outside of target range[2]. Off-beat targets were more difficult to identify than on-beat ones, and that was especially true with distractors inside the target range". The results suggest then that "listeners were using their knowledge of the temporal organization of the target melody to aim their expectancies at times when target notes should occur".

Listeners carried out two types of task in the second experiment — a contour-judgement task and a pitch-judgement task. Both tasks consisted in presenting a standard melody followed by a comparison melody. Comparison was the standard melody interleaved with distractor notes. In contour-judgement task, the penultimate note of the comparison — the critical note — could change its pitch, and so, the listeners' task was to say whether the contour of the comparison melody had been changed in relation to the standard. In pitch-judgement task, in its turn, the middle note of the comparison — the critical note — moved to another pitch, and the listeners' task was therefore to identify the new pitch. The results indicate that, "especially in the

[2]Target range is the pitch range of a target melody.

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 in the third and fourth experiments, the remote pitches became perceptually very salient. Dowling, Lung, and Herrbold then conclude that, to discern target melodies interleaved with distractors, it is required "a focusing of attention in pitch-time expectancy windows within which events critical for target identification are likely to occur. As a result of that narrowing of attention, events that would otherwise be highly salient were not distinctly perceived and were difficult to judge accurately".

2.4.3 Gallun and Reisberg's experiments

Gallun and Reisberg (1995) performed four experiments to replicate the results of Dowling, Lung, and Herrbold's second experiment (section 2.4.2). Gallun and Reisberg realized that results and method employed in that experiment deserved a close scrutiny for three reasons. First, the results suggest that expectancies are influencing what is perceived. They are, therefore, important once the question of whether expectation can influence perception itself is still open to debate. Second, the results are strongly counterintuitive. As opposed to what those results are suggesting, large changes in a stimulus are usually more salient and noticeable than small changes. Third, Dowling, Lung, and Herrbold's method might be applied to other realms. Apart from pitch anticipation, it would be worth investigating, for instance, whether or not subjects anticipate loudness or timbre.

In the first of the experiments, Gallun and Reisberg employed stimuli and procedures similar to those employed in Dowling, Lung, and Herrbold's second experiment. In the other three experiments, they tried other alternatives, as to reduce the speed of presentation of the standard melodies, increase the number of their presentations, use subjects with different levels of musical experience, and utilize melodies which were able to be more easily anticipated.

In all four experiments, the results were similar, and did not replicate those obtained by Dowling, Lung, and Herrbold in their second experiment. Gallun and Reisberg's results indicate that "the larger contrasts were easier to detect than the smaller changes", and consequently, suggest that "subjects are failing to generate expectancy windows". Thus, they conclude that it is necessary "to reject the view that anticipation guides perception in the interleaved-melodies paradigm", and also, "to reopen the question of whether (or how) anticipations shape perception".

2.4.4 Palmer and Holleran's experiments

Palmer and Holleran (1994) carried out two experiments in order to investigate the influences of harmonic relations, melody location, and frequency height in perception of multivoiced music. Four three-voice musical pieces were composed for the experiments. Two of the pieces held a single melodic line, and so, were of homophonic compositional structure. The other two were polyphonic holding two melodic lines. Melodic lines in the four pieces appeared either in the highest or in the lowest of the three voices.

Twenty seven comparison compositions were created based on the four standard compositions. A third of the comparisons were identical to the standards. The rest were variations of the stan-

[3]Diatonic pitches are the critical notes in this context.

[4]The expected range was considered as being the pitch range of the overall melody.

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 to avoid upper harmonics inherent in piano timbre.

The results from the second experiment replicated those from the first, and thus, let Palmer and Holleran draw three main conclusions. First, "harmonic relationships among voices affected the detection of pitch changes". Changes which kept the harmonic relations were more difficult to detect than those which did not. Second, "frequency height influenced the detectability of pitch changes. The worst detection of pitch changes ... occurred in the middle-frequency voice, and the best detection ... occurred in the highest-frequency voice". Third, "both melody location and frequency height aided detection of pitch changes". "This finding suggests that listeners may attend more readily to the melodic voice, especially when it occurs in the high-frequency range".

2.4.5 Brief introduction to music

2.4.5.1 Univoiced and multivoiced music

The study of Palmer and Holleran (section 2.4.4) is considerably important because it is situated in the realm of multivoiced music. Despite the fact that most western tonal music is multivoiced, little research has been carried out in order to understand the mechanisms behind perception of such type of music.

Multivoiced music contains multiple *voices*[5] which are sounded simultaneously. The opposite of multivoiced music is univoiced music. *Univoiced music*, which is also known as *monophonic music*, contains just one single voice.

Perception of the musical structures in multivoiced music is usually complex for the voices interact among themselves. Western tonal multivoiced music is divided into two separated fields — homophonic music and polyphonic music.

2.4.5.2 Homophonic and polyphonic music

Homophonic music contains one primary voice — the *melody* — and additional secondary voices which provide the harmonic *accompaniment*. The melodic line seems to hold independence from the other lines. Indeed, as Lashley (1954, p. 425–426) points out, "while listening to a musical composition the listener can follow the melodic line and abstract it from the instrumental accompaniment". Povel and Egmond (1993) provided experimental support to that. Their results show that "melody recognition is unaffected by the variations in accompaniments, supporting the view that melody and accompaniment are processed relatively independently".

In *polyphonic music*, in its turn, there is not the notion of a primary line, for all lines share the same importance. *Polyphony*, as Francès (1988, p. 203) writes, "is a hierarchy of levels of sound from which each in turn detaches itself momentarily without the others being reduced to the function of ground". All polyphonic musical forms from the sixteenth to eighteenth centuries, such as canon and fugue, set their grounds on counterpoint. Hereafter therefore, contrapuntal music will be referred to as polyphonic music, and vice versa.

2.4.5.3 Counterpoint and double counterpoint

Counterpoint, as Prout (1890, p. 15) defines, is "the art of combining two or more parts [or lines] or voices, each of which possesses independent melodic interest and importance". Despite their independence, contrapuntal lines should observe harmonic rules, so that, when sounded together, they may produce correct harmony (Kitson, 1924; Fux, 1971).

[5] *Voices* are also known as *parts* or *lines* in music.

"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 imitations and transpositions through the voices (Beard, 1985; Flindell, 1983). The first two-part invention in C major was composed under imitation form. Figure 2.4 displays its first bar in which the theme is presented in the highest voice, and then imitated in the lowest.

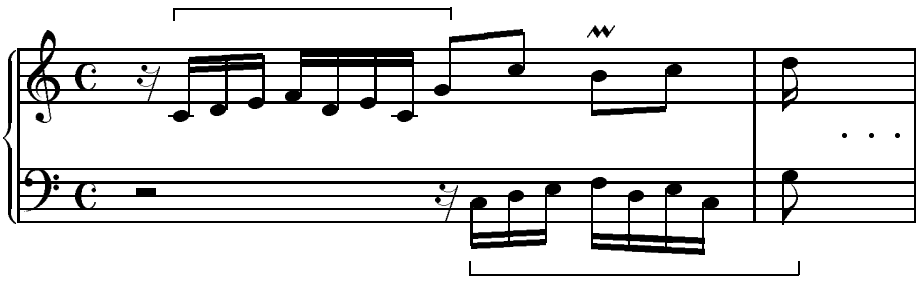


Figure 2.4. First bar of the first two-part invention in C major

2.4.5.6 Fugal form

Musical forms of the baroque period are not completely defined and formalized (Flindell, 1984). As a result, there is no precise framework in which a fugue must fit in. There are considerable differences between fugues in terms of their constructions. Such scenario is particularly correct in fugues of Bach. Indeed, as Kirkpatrick (1984, p. 34) points out, "if one wants to find a typical Bach fugue, one has to fabricate it". However, it is possible to define, in general terms, fugal form.

Prout (1891, p. 1) defines *fugue* as "a composition founded upon one *subject*[6], announced at first in one part alone, and subsequently imitated by all the other parts in turn, according to certain general principles[7]". A fugue may be divided into three sections — exposition, middle section, and final section (Prout, 1891).

[6] *Subject* is the *theme* of a fugue.

[7] These principles will not be approached here because they lie beyond the scope of the dissertation.

Exposition begins with the single presentation of the theme in one part. After that, the theme is presented, either in its original key or transposed, in all other parts in turn. The *middle section* concedes a great amount of freedom to the composer, and thus, it is the section where he or she elaborates on the thematic material through the parts. The *final section* is that in which a return is made to original key.

Figure 2.5 and 2.6 display respectively the first and second bars of the fifth four-part fugue in D major of the first volume of The Well-Tempered Clavier (Bach, 1989). The theme is introduced unaccompanied in the lowest part in the first bar. In the second, the transposed theme is presented in the lower-middle part.

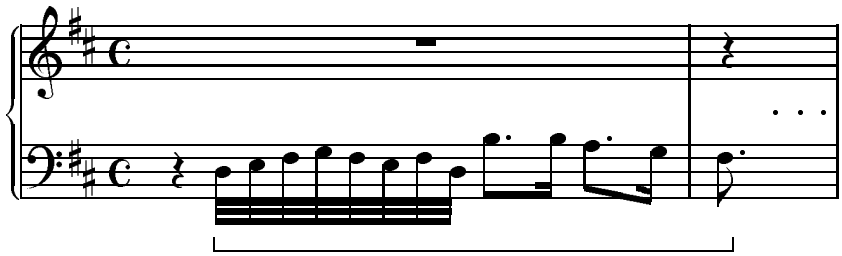


Figure 2.5. First bar of the fifth fugue in D major

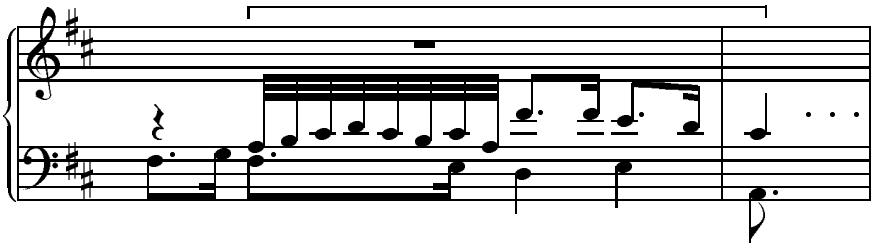


Figure 2.6. Second bar of the fifth fugue in D major

2.4.6 Current open issues in thematic recognition in polyphony

As mentioned before, 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.

We have hitherto presented works loosely related with thematic recognition. For instance, Dowling (section 2.4.1) studied recognition of melodies in the presence of distractors — foreign notes interleaved with melodies — and Palmer and Holleran (section 2.4.4) examined recognition of alterations in melodies in multivoiced music. Unfortunately yet, as far as we know, there have not been experimental works in the domain of real polyphonic music, and thus, important issues concerning thematic recognition in such domain are still open to investigation.

Indeed, there are many more questions than answers relating to thematic recognition in the polyphonic domain. First, it is not known how listeners are able to recognize instances of theme through the voices, and which cognitive mechanisms are involved in the understanding of polyphonic music.

Second, it is not known which kind of and degree of variations are permitted in an instance of theme, so that its recognition remains unaffected. Studies in monophony suggest that listeners are

able to perceive inversions of melodies (see section 2.2.4). They may also perceive a melody even when it arises transposed or altered (see sections 2.2.4 and 2.2.6). However, owing to the complex effects which may be formed by the interactions of the simultaneous voices, one must not assume beforehand, and without experimental results, that such transformations would be perceived in the same extent in polyphony as well.

Third, performers usually reinforce notes of the theme in polyphonic music. Kirkpatrick (1984) even suggests that pianists play such notes louder. Nevertheless, it is not known whether or not, and to what extent, listeners rely on reinforcement in order to recognize properly instances of theme.

Fourth, it is not known how listeners' performance in thematic recognition is affected by the number of voices which sound simultaneously with the theme. Huron (1989) showed that the more voices sounding simultaneously, the greater the difficulty listeners had in perceiving voice entrances. So, that might suggest a similar position respecting to thematic recognition.

Fifth, it is not known how listeners' performance in thematic recognition is affected by the amount of onset synchrony. As Huron (1993) writes, "experimental evidence has shown that the perceptual segregation of concurrent auditory streams is enhanced when tone onsets are asynchronous rather than synchronous". Based on the analysis of the 15 two-part inventions, he concludes that "Bach endeavors to minimize simultaneous note onsets between concurrent voices". Lower amounts of onset synchrony between voices might thus ease the listeners' task of identifying instances of theme.

Finally, although segmentation may be imperative to the complete musical understanding in the homophonic domain, it is not known whether or not it is necessary for thematic recognition in the polyphonic domain. If segmentation be necessary, it is performed under circumstances different from those present in homophony, and so, it leads to different results.

In the monophonic domain, grouping rules of different type and strength, as those proposed by Lerdahl and Jackendoff (section 2.3.4), may conflict between themselves. However, competition between rules arises only in the melodic line since it is the leading line. Two theoretical positions exist (Drake & Palmer, 1993). First, listeners apply conflicting grouping rules independently of each other, and as a result, the perception of one accent produced by one rule is unaffected by the juxtaposition of other accents produced by other rules. Second, listeners apply conflicting grouping rules interactively, and as a consequence, the perception of one accent produced by one rule is affected by the juxtaposition of other accents produced by other rules. In this position, rhythmic rules are the most stable, and dominate other types of rules (Drake & Palmer, 1993). Despite the two existent theoretical positions, it is worth mentioning here that "no-one has yet attempted a framework for a comprehensive analysis of competition of grouping principles" (West et al., 1991).

In the polyphonic domain, in its turn, segmentation is more complicated because grouping rules either of same type and strength or not may conflict between themselves. That arises from the fact that all lines have now the same importance. Two theoretical positions similar to those in monophony might exist here as well. First, listeners apply conflicting grouping rules, either in the same or in different lines, independently of each other. This is the position adopted by Kirkpatrick (1984). The second position is that where conflicting grouping rules interact. The interaction of grouping rules may cause impasses once rules of the same type and strength applied to different lines may require group boundaries at different serial positions. Figure 2.7 illustrates this case in a *stretto*[8] taken from the fifty second and fifty third bars of the eighth three-part fugue in D sharp minor of the first volume of The Well-Tempered Clavier of Bach (Bach, 1989). Figure 2.8 displays its corresponding visual analogy. The second theoretical position demands thus, either an extension of the Lerdahl and Jackendoff's theory (section 2.3.4) to polyphony or the development of an alternative theory (Carpinteiro, 1993).

[8] *Stretto* is a musical passage where two or more instances of theme overlap.

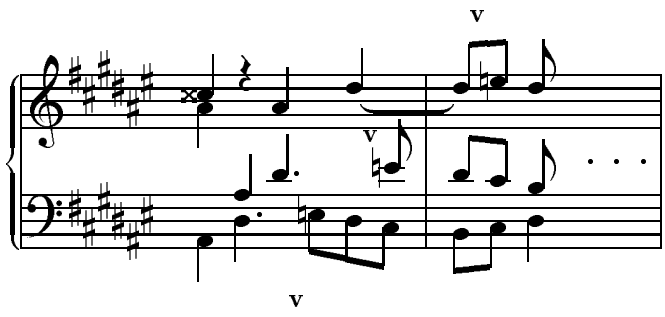
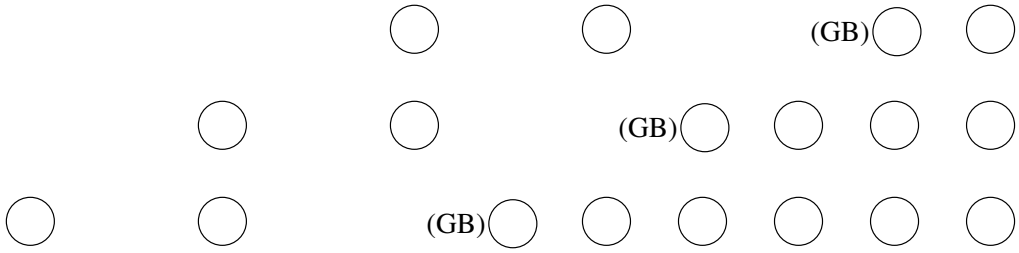


Figure 2.7. A stretto from the eighth fugue in D sharp minor



GB - Group Boundary

Figure 2.8. A stretto from the eighth fugue in D sharp minor (visual analogy)

2.5 Summary

Listeners rely on mental representations of contour and interval sizes in storing and identifying melodic patterns. Nevertheless, contour and interval informations are stored independently.

Interval information is difficult to extract from novel melodies, mainly when key is not established, yet is resistant to decay. Moreover, in tasks involving long-term memory, as to identify familiar melodies, listeners make use mainly of information concerning interval sizes. Thus, interval information seems to relate to long-term memory. On the other hand, contour information seems to be related to short-term memory. Contour information is easily extracted from novel melodies, yet is easily lost. Listeners rely on it in recognizing transpositions and inversions.

Neuropsychological research has lent support to the dichotomy between contour and interval information. The cerebral left and right hemispheres are entailed in tasks which require interval and contour information respectively. That has been evidenced by studies of brain-damaged patients. 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. Thus, melodic contour seems to provide a necessary sketch for encoding interval information.

Listeners do not grasp a musical piece in its entirety, but rather, they segment it into parts. Segmentation is necessary for two main reasons. First, the capacity of human memory is limited. Second, segmentation produces grouping, and groups produced by segmentation become musical units.

Lerdahl and Jackendoff 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, and hence, the formal rules describe how listeners parse

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 instance, issues related to cognitive mechanisms which are involved in the understanding of polyphonic music, issues related to thematic reinforcement, and to segmentation. Segmentation may be or not necessary for thematic recognition in polyphonic music. If it be necessary, either an extension of the Lerdahl and Jackendoff's theory to polyphony or the development of an alternative theory may be required.

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 (Rumelhart et al., 1986b) as the learning algorithm, were employed in two series of experiments.

[1] Robustness is the ability to perform a task in noisy conditions.

[2] Soft constraint satisfaction is the ability to not be limited to all or nothing decisions.

The first experiment was carried out in nets which modelled chord perception. The output units of the nets represented the interval structure of chords. The training set consisted of 36 chords — 12 major, 12 minor, and 12 diminished triads — and learning took place on a pattern-by-pattern basis.

The experiment was divided into three parts. Each part consisted in evaluating one of three different representations of pitch — pitch class, harmonic complex, and subharmonic complex representations.

In the first part, the best performance was reached in a net with 12 input units, 25 hidden units, and 3 output units. The input units represented the 12 pitch classes in an octave. The net was able to identify 34 out of the 36 chords in the training set. However, it was sensitive to the initial weights and learning parameters chosen.

In the second part, the input units represented harmonics of tones. Laden and Keefe tried this representation because it is believed that listeners are able to extract musical pitch from harmonics which compose a complex tone. Thus, each musical tone was represented by five pitch classes which approximated its five first harmonics. Each chord activated 15 input units — five harmonics for each chord tone. They made use of a net with 47 input, 25 hidden, and 3 output units. The net was able to identify all 36 chords, and displayed a better performance, in terms of number of epochs required, than that of the net in the first part.

In the third part, the input units represented subharmonics of tones. Subharmonics are elements of a cognitive model proposed by Terhardt (1974). In this model, each tone generates a number of descending musical intervals which are subharmonics of the tone. Each tone was represented by six subharmonics, and consequently, each chord activated 18 input units. A net with 50 input, 25 hidden, and 3 output units was employed. It classified correctly all but one of the 36 chords.

Laden and Keefe investigated the ability of the net of the second part to generalize. They presented it incomplete patterns, chord inversions, and input shaped after a spectrum of harmonics of a fixed tone.

The study in generalization with incomplete patterns was motivated by the fact that "human listeners can extract pitch and classify chords in the absence of complete harmonic template patterns". In fact, "even the fundamental[3] of a complex tone does not need to be present in order for it to be perceived as the pitch of the complex tone". The net was presented incomplete chord patterns where one or more harmonics of each chord tone were missing. It was able to identify the incomplete patterns with performance above chance levels. Moreover, it was observed that the net generalizes better in absence of lower harmonics than in absence of upper ones.

The study in generalization employing chord inversions was motivated by the fact that both root position and inversions of a chord share a common harmonic basis. The net was presented first and second inversions of major, minor, and diminished chords. It was verified that recognition of second inversions was better than that of first inversions. However, both inversions were identified with performance levels above chance.

The last study in generalization employed input shaped according to a spectrum of harmonics of a fixed tone. In this case, the input units representing the five first harmonics of a tone hold different activations, proportional to amplitude levels of the corresponding five first harmonics of that fixed tone. The motivation behind this study was the fact that "human listeners can successfully identify chord type while the power spectra[4] of the individual tones are varied, such as when different trios of instruments play the same chord". It was observed that chord recognition was more difficult when amplitude levels of upper harmonics were reduced. However, the net performance exceeded chance levels.

The second experiment was carried out in a net which modelled pitch perception. The net held 64 input, 30 hidden, and 36 output units. The input units represented the five first harmonics of a

[3]The fundamental is the first harmonic of a complex tone.

[4]The power spectrum of a tone is the spectrum of harmonics of that tone. In the spectrum, each harmonic is described by its frequency and amplitude levels.

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 compositions. Therefore, the net carried out two separate tasks.

The first task was that of segmenting the compositions into small groups of notes. Gjerdingen realized the importance of performing segmentation by writing that "for a network to retain a memory of a long series of events, a long melody for instance, it must in some way segment the series into parts short enough for each to fit within the limits of short-term memory" (Gjerdingen, 1990). By performing segmentation, the net is thus modelling an important human cognitive task necessary for musical understanding (see section 2.3.3). However, segmentation performed by the net is deficient, for it was not performed after Gestalt principles of proximity and similarity (see section 2.3.1), but rather, merely after the size of its short-term memory.

The second task carried out by the net was that of classifying the groups of notes produced by segmentation into categories. Unfortunately, nothing was mentioned about existence or not of similar sequences of patterns in the training set, so that one could not assess performance of the net in terms of classifications and misclassifications of such sequences.

Although results obtained by Gjerdingen were of some importance, it was his proposals which were of a much greater significance. He proposed another model as displayed in figure 3.1.

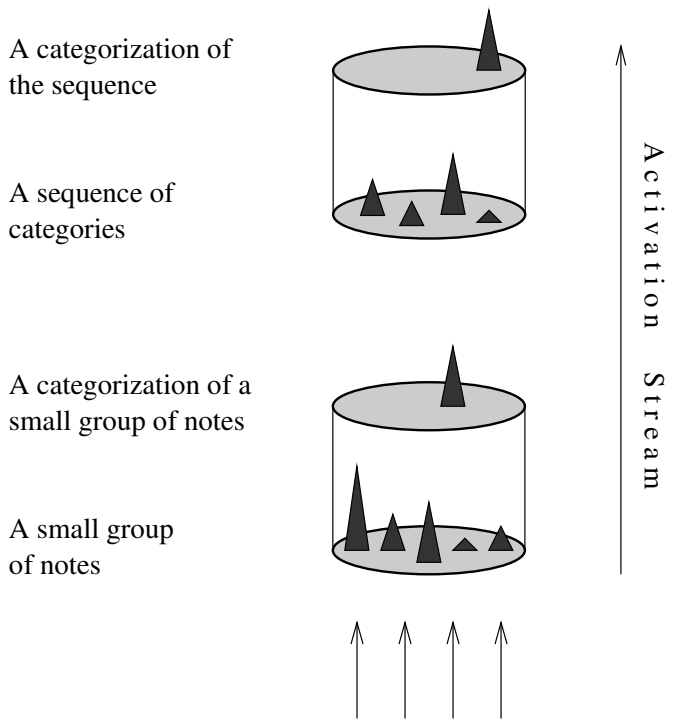


Figure 3.1. Gjerdingen's proposed model

The model would consist of two ART2 nets, one on top of the other. The bottom net would be responsible for segmenting musical pieces into small groups of notes, and classifying them into categories. These categories would then be passed up, as a corresponding sequence of activations decayed in time, to the top net. By classifying the sequence, the top net would thus be classifying a larger group of notes into a novel category.

As human listeners do, the proposed model was therefore intended to segment musical pieces

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 backpropagation learning algorithm (Rumelhart et al., 1986b) 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 representative approaches in the field are described below.

3.3.1 Sejnowski and Rosenberg's model

Sejnowski and Rosenberg (1987) designed an artificial neural model to pronounce English text. The model, which is named *NETtalk*, is displayed in figure 3.2.

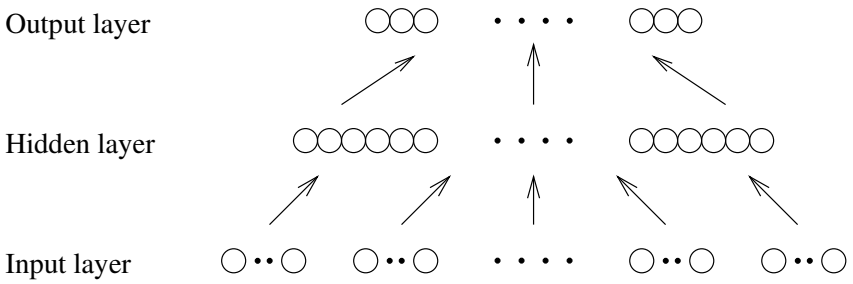


Figure 3.2. Sejnowski and Rosenberg's model

The input layer held 7 groups of 29 units which made up a temporal window. The 29 units locally represented the 26 letters of the alphabet plus 3 termination symbols. Each group encoded either one letter or one termination symbol of text, and so, seven characters — letters or symbols — were presented to the input units at once. The text moved through the window character by character in time. The output layer held 26 units representing phonemes in a distributed representation. The desired output of the model was the correct phoneme associated with the fourth character in the window.

Sejnowski and Rosenberg performed two experiments. In the first, the training set consisted of 1024 words taken from recorded speech of children. After 50 epochs training, the model could learn 95% of the training set. It was tested on a novel set of 439 words as well. Its performance was 78% for this set. In the second experiment, the training set consisted of 1000 commonly occurring words selected from a dictionary. Performance of the model on the training set and on a novel set containing 20012 words was 98% and 77% respectively.

Sejnowski and Rosenberg's model has a particular characteristic. It does not easily achieve invariance under translation in time. For instance, let us consider a window with 5 units, and three sequences — 11100, 01110, and 00111. Despite of incorporating the same basic pattern 111, the sequences have little resemblance for the model. This is a disadvantage when the application

requires the sequences to be classified under the same category. Nevertheless, it is an advantage when the opposite is the case.

Sejnowski and Rosenberg's model has two further drawbacks. First, the window size must be large enough to fit the longest possible input sequence. As a consequence, the size of such sequence must be known in advance. Second, the model becomes computationally expensive as wider windows are required.

The model has an important virtue as well. It is capable of holding a precise memory of past events. The larger the window size, the longer the memory for past events.

3.3.2 Rumelhart, Hinton, and Williams's learning algorithm

Rumelhart, Hinton, and Williams (1986b) extended the backpropagation learning algorithm to recurrent artificial neural models. The extension of the algorithm is usually known as backpropagation in time. Backpropagation in time is thus a learning algorithm which may be applied to any recurrent model. Yet, from now on, in an abuse of meaning in order to simplify, the learning algorithm will be referred to as *backpropagation in time algorithm*, and the recurrent model displayed in figure 3.3 will be referred to either as *backpropagation in time model* or simply as *backpropagation in time*.

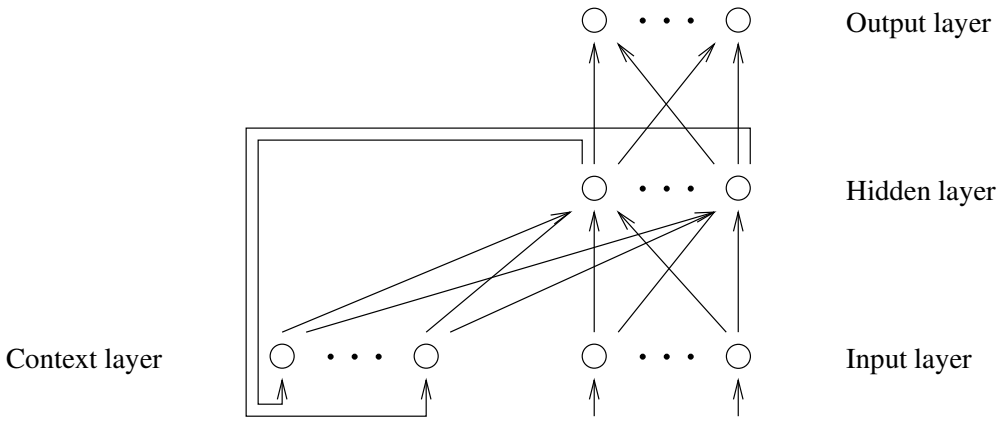


Figure 3.3. Backpropagation in time

The model consists of one recurrent net in which each hidden unit delivers activation to its corresponding context unit which, in turn, sends activation back to all hidden units. Recurrent models may be unfolded in time. Figure 3.4 unfolds in time the model presented in figure 3.3.

In the unfolded model, there is one feedforward non-recurrent net for each step in time. The weights connecting each layer of units to the next share identical values across the nets. The backpropagation in time algorithm consists then in following three main phases. First, to input a sequence, so that each pattern of the sequence be input to one single non-recurrent net. Second, to backpropagate the error to the very first net. Third, to update the weights.

Rumelhart, Hinton, and Williams tested the algorithm on a model slightly different from that shown in figure 3.3. In their model, apart from recurrent connections in the hidden layer identical to that displayed in the model of figure 3.3, each output unit was also recurrently connected to each other output unit, and to itself. Training set consisted of 20 sequences with two letters and four numbers where the first two letters set up the subsequent four numbers. The model was required to predict the numbers. Training was carried out in 260 epochs, and took place on an epoch-by-epoch basis. Results showed that the model was able to learn the entire training set, and to generalize on a novel set as well.

The main advantage of backpropagation in time model is that it possesses a precise memory

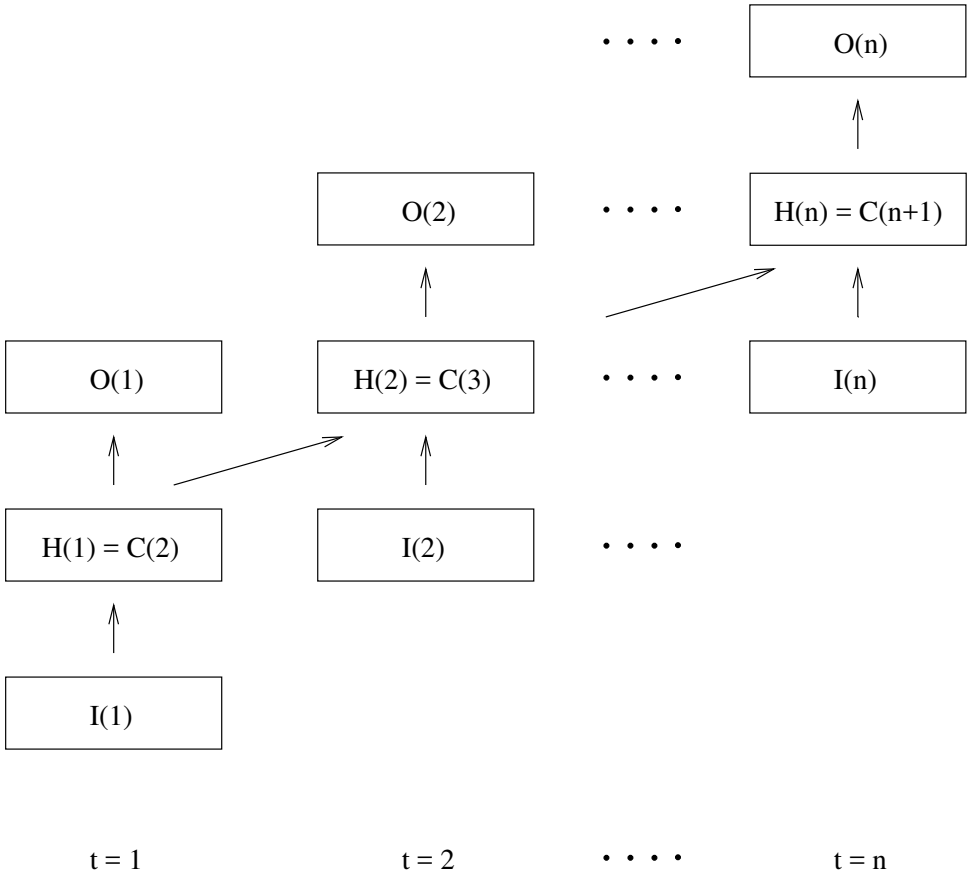


Figure 3.4. Backpropagation in time unfolded

of previous inputs, and so, is theoretically capable of computing any function of the inputs. Its main drawback, however, is the amount of memory required. Indeed, the model requires the storage of weight changes whilst error is backpropagated in time, and of the sequence of unit activations. Owing to this, Ghahramani and Allen (1991) point out that backpropagation in time is not a "plausible [cognitive] model of temporal processing".

3.3.3 Mozer's model

Mozer (1989) proposed a model which is a simplification of backpropagation in time (section 3.3.2). It is shown in figure 3.5.

The model consists of a recurrent net in which context units receive activation from, and send activation back to their corresponding hidden units. Owing to this simplification, memory requirements of the model are lower than those of backpropagation in time. Indeed, for each hidden unit, the error signals of previous steps in time may be easily calculated as a function of the current error signal. Nonetheless, as backpropagation in time, the model also requires the storage of net activations through time.

Mozer carried out a set of experiments. In one of them, he trained the model to reproduce sequences. The training set consisted of six sequences. Each sequence was made up of three parts — three letters in any order at its beginning, a number of zeros corresponding to a fixed delay, and three zeros at its ending. The task of the net was, when a sequence was input, to exhibit in its output units a sequence of zeros followed by the three letters of the input sequence.

The net was able to learn to reproduce the sequences in the training set. Nonetheless, the task of learning became more arduous when the fixed delay in training sequences was increased.

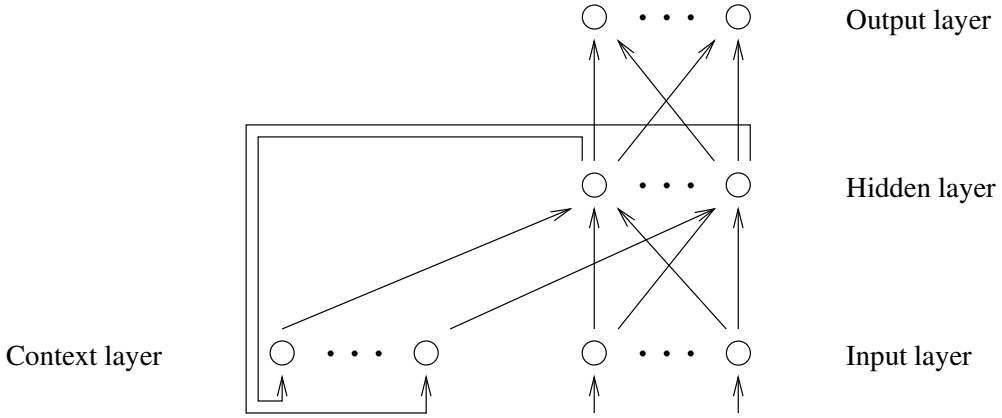


Figure 3.5. Mozer's model

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.

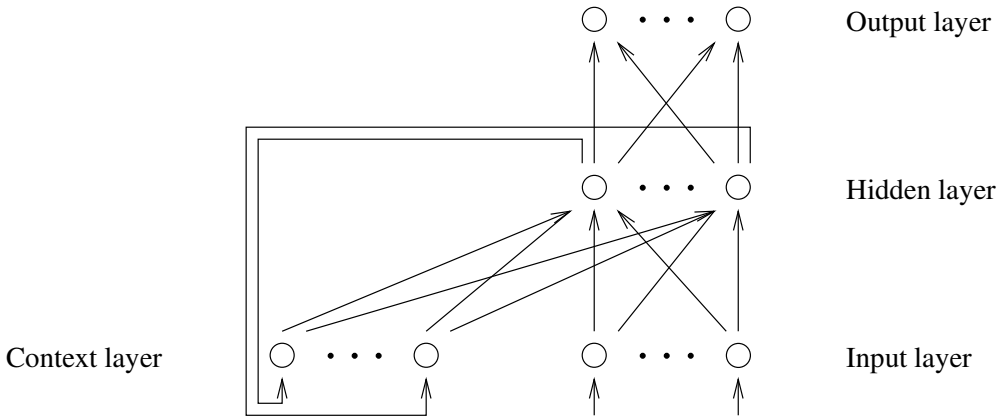


Figure 3.6. Elman's model

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 described below.

3.4.1 Kangas' model

Kangas' model (1991, 1994) is a part of a Finnish speech recognition system. The model is shown in figure 3.7.

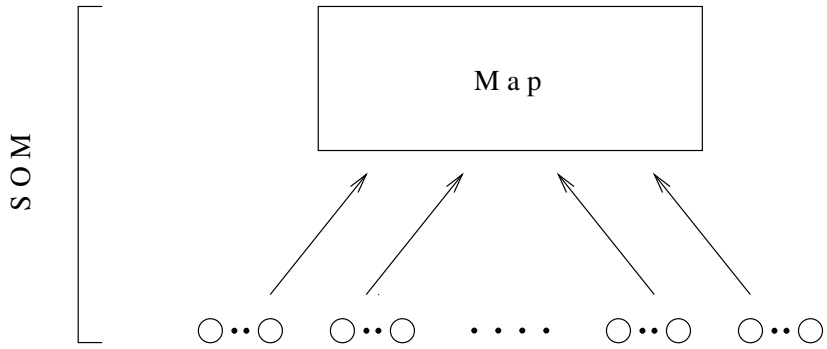


Figure 3.7. Kangas' model

The input layer holds up to 10 groups of 22 units making up a fixed-size window which slides in time. The 22 units represent features of a digitised speech sound. The map holds 216 units, and self-organizes into phonetic classes.

Kangas replaced the standard self-organizing map model with his in the speech recognition system. He could verify that results achieved with his model significantly improved those achieved with standard self-organizing map model. Notwithstanding the good results, as a windowed model, Kangas' model has virtues and drawbacks similar to those of Sejnowski and Rosenberg's

[7]We performed a set of experiments, not presented in the dissertation, with Elman's model in such a domain. Results obtained corroborated the drawback.

(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.

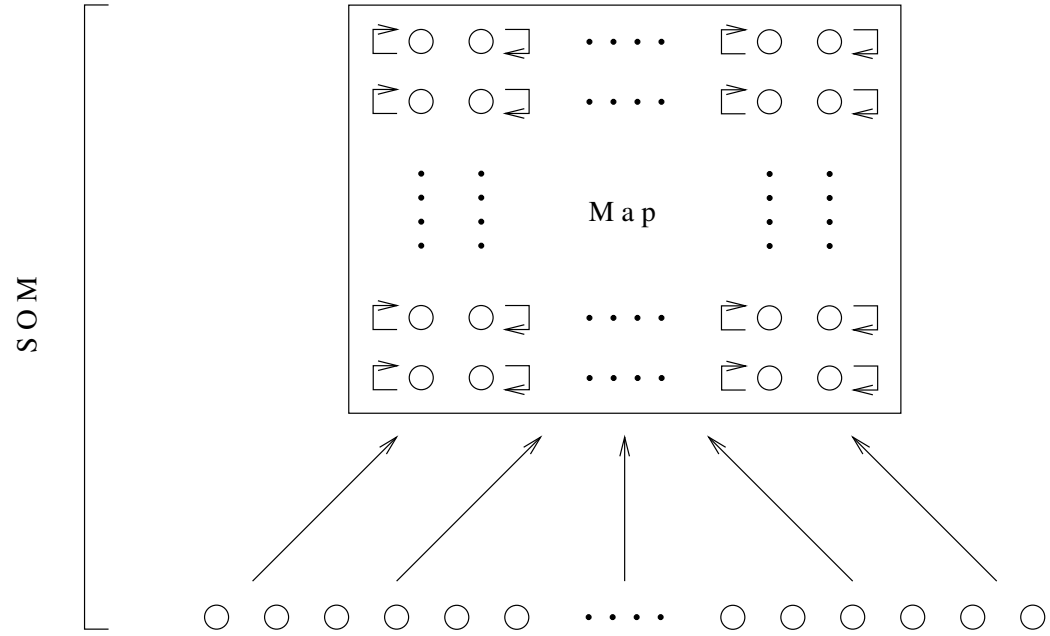


Figure 3.8. Chappell and Taylor's model

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 the word 'net' which occurred in the three sentences in three different units in the map. Thus, Chappell and Taylor claimed that their model was sensitive to the different contexts in which the word 'dry' occurred. Despite their claim, it is worth noticing that those contexts did not share much similarity at all, and this obviously helped the model in that distinct classification.

The main virtue of the model is its biological plausibility. Artificial neural units endowed with time integrators simulate the behaviour of real neurons, which retain an electrical potential with time decay on their membranes. The main pitfall of the model is that it suffers from loss of

[8]Also known as leaky integral approach.

context. Sequences which possess a similar context by slightly differing in their initial elements (e.g., abccc, baccc, and bbccc) would probably have identical classification.

3.4.3 James and Miikkulainen's model

Another recent model to classify sequences in time is James and Miikkulainen's (1995). The model is displayed in figure 3.9.

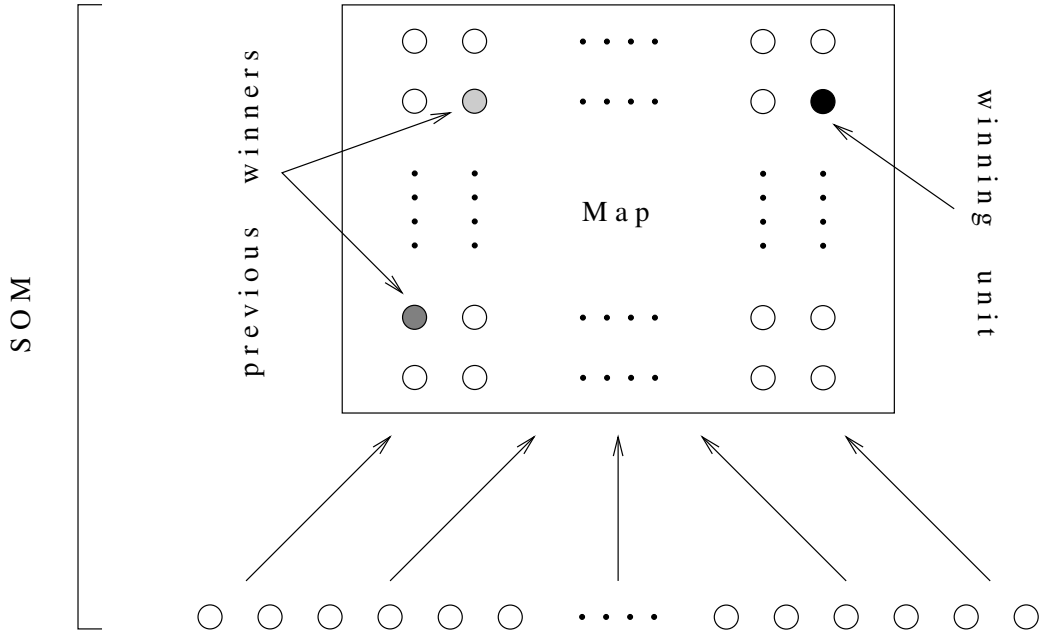


Figure 3.9. James and Miikkulainen's model

When a vector in the sequence is input, the output unit in the map which wins the competition for that vector is disabled for further competition, and its activation decays in time to indicate which vector in the sequence it is representing. Each winning output unit represents just a single vector in the input sequence, and so, the representation for the whole input sequence is distributed, for it is given by a sequence of winning output units in the map.

James and Miikkulainen trained and tested the model on sequences of phonetic representations for English words. The words varied from five to twelve phonemes in length, and each phoneme was represented by a vector with five coordinates. There was a total of 43 phonemes which made up the words in three training sets. The sets containing 713, 988, and 1628 words were used to train, respectively, three net configurations with 81, 84, and 88 output units in the map. The results showed that the three configurations were able to learn how to classify the words in the training sets.

Any output unit in the map may participate in the representation of many input sequences, once each winning unit does not represent a whole input sequence, but rather, just one of its vectors. This distributed representation concedes a high representational power to the model, for even small maps are able, in principle, to represent a large number of input sequences. On the other hand, the distributed representation is more complicated because it has to take in consideration not only the winning units but also their activations. For example, input sequences with the same elements but in different order (e.g., abcd and dcba) will have the same winning units, and so, one has to look at their activations to verify which input sequence the map is representing. Another disadvantage of the model is that it is unable to recognize sub-sequences inside a large and unique input sequence.

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. They are class representatives of many current models ranging from windowed models to recurrent ones. We paid special attention to their main virtues and limitations. In particular, we paid attention focusing the virtues of Sejnowski and Rosenberg's model. Their model is capable of holding a precise memory of past events, and therefore, is suitable in cognitive domains, such as musical segmentation, which demand such requirement.

Finally, the last three models reviewed were unsupervised. They are extensions of the self-organizing feature map to classification of sequential information. Their advantages and disadvantages deserved special attention. It was noted that the models possess serious drawbacks, such as high computational cost, loss of context, and inability to recognize sub-sequences within an unique input sequence. Such drawbacks prevent them from being employed in the cognitive domain of thematic recognition in polyphony.

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 sections. The first section is this introduction. The second section introduces the representation developed for rhythmic sequences. The third section describes the neural model. The fourth, fifth, and sixth sections are concerned with experimentation. Finally, the seventh section provides a summary of the chapter.

4.2 Representation for rhythmic sequences

If we set the small figure in a musical sequence to be the *time interval (TI)*, all other figures become multiples of TI. For example, if TI is an eighth note, a quarter note lasts two TIs, a half note lasts four TIs, and so on. We may also define a counter, *time interval counter (TIC)*. One TIC lasts one TI.

TIC is the unit in which the musical sequence is measured. Therefore, at each TIC, either there is a *rest*, or a note onset, or a note sustained. We refer to *note onset* as the onset of the note, and to *note sustained* as the note continuous sounding after its onset.

Figure 4.1 illustrates the representation. It shows a sequence which lasts nine TICs, and whose TI is an eighth note. In the experiments (sections 4.4, 4.5, and 4.6), each of the three events was represented by a pair of neural input units. A rest was represented by (00). Note sustained was represented by (10), and note onset by (11).

Figure 4.1. A musical sequence lasting nine TICs

4.3 The model

The cases of rhythmic segmentation demand that the model have long memory of past events. Also, they demand it to establish the precise position where boundaries ought to occur. These requirements directed us towards the windowed models domain.

The model presented here has a topology similar to that of Sejnowski and Rosenberg's model (section 3.3.1). It is shown in figure 4.2.

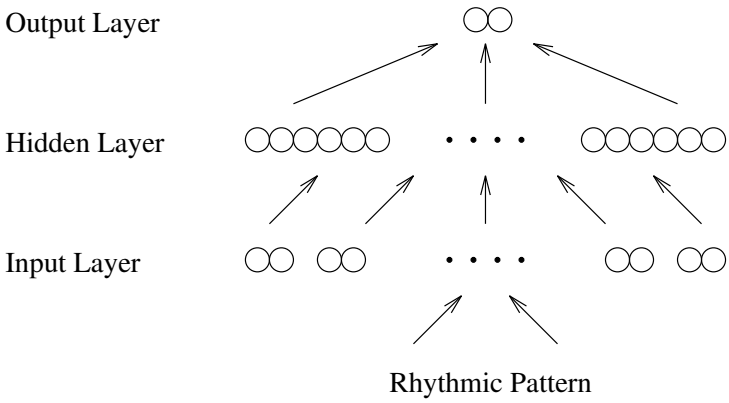


Figure 4.2. The model

The input layer holds a number of pairs of units which make up a window. Each pair represents one of the three events — rest, note sustained, and note onset. The activations of pairs of units in the window represent a rhythmic pattern. For instance, let TI be an eighth note, and the size of the window be nine pairs of units. Figure 4.3 would then represent the example in figure 4.1.

11 10 11 10 10 10 00 00 11

Figure 4.3. Representation for the musical sequence in figure 4.1

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 modified, and total error is calculated. Learning rate is reduced by 50% when total error increases, and increased by 2% when error decreases. Momentum is disabled until the end of training if total error increases. Total error E is given by

$$E = \sum_p \sum_i \delta_i^2(p) \tag{4.5}$$

where subscript p indexes pattern number, and δ_i is the error signal for output unit i.

Error signal δ_i, for an output unit i, is given by

$$\delta_i = t_i - a_i \tag{4.6}$$

where t_i is the desired activation value and a_i is the activation obtained. For a hidden unit i, δ_i is given by

$$\delta_i = a_i(1 - a_i) \sum_k \delta_k w_{ki} \tag{4.7}$$

where w_{ki} is the weight from hidden unit i to output unit k.

Two output units were used in all experiments. 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.

[1] *Flat spots* are points in which the derivative of the sigmoid function approaches zero. The recovery of a non-linear output unit becomes extremely slow when it displays an incorrect output value on a flat spot.

[2] The existence of bias is not necessary, for the output units are linear. We have, however, decided to keep them.

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

4.4 First experiment

The first experiment was on recognizing cases of segmentation given by rests. We extended the Lerdahl and Jackendoff's rule presented in the second chapter (section 2.3.4.4) to include early voice entrances in polyphonic music. The extension of the rule was acknowledged by Kirkpatrick (1984). It states that segmentation given by rests takes place whenever a rest is followed by at least two notes.

Three sets of patterns were initially generated for the experiment. One of the sets was training set, and the other two were test sets. Each set contained 1000 negative and 1000 positive patterns, which were generated by means of using all pattern templates described in table 4.1.

The structure of the templates was based on the rule of segmentation given by rests. It is made up by three parts, which are represented by the three groups of square brackets occurring in each template displayed in table 4.1. The first and last parts contain one of the three events — rest, note sustained, and note onset. The middle part contains either one of these events or one note onset

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
1	1	1
2	1	2
3	1	8
4	1	24
5	5	74
6	15	222
7	47/48	668

event followed by any number, between one and seven, of note sustained events.

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 in which the first, last, and middle parts are given respectively by one rest, one note onset, and either one single note onset or one note onset followed by notes sustained. The negative templates are all those otherwise.

Many of the templates were shorter than the input window, and consequently, a number of slots remained available on them. These free slots were filled up randomly with the three events[3], producing an amount of patterns as displayed in table 4.2. For instance, 15 and 222 patterns were generated respectively for each negative and positive pattern template containing six free slots.

We might illustrate the generation of a pattern in an example. Let us consider the positive template in which the first, middle, and last parts are given respectively by one rest, one note onset followed by three notes sustained, and one note onset. There are four free slots at the end of the template, and thus, according to table 4.2, 24 positive patterns are generated from the template. All of these generated patterns consist of the events in the three parts of the template, followed by four events which are determined randomly.

The first of the three sets of generated patterns was training set. Every 20 epochs, 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.

Figures 4.4, 4.5, and 4.6 display the training and testing on the three sets for four net configurations. For each configuration, the window in the input layer held 10 pairs of units, and initial weights were set randomly. Best performance was given by a configuration with 20 hidden units, trained for 80 epochs.

[3]In spite of the slots were filled up randomly, musical constraints were observed. Thus, for example, a note sustained event was not allowed to follow a rest event.

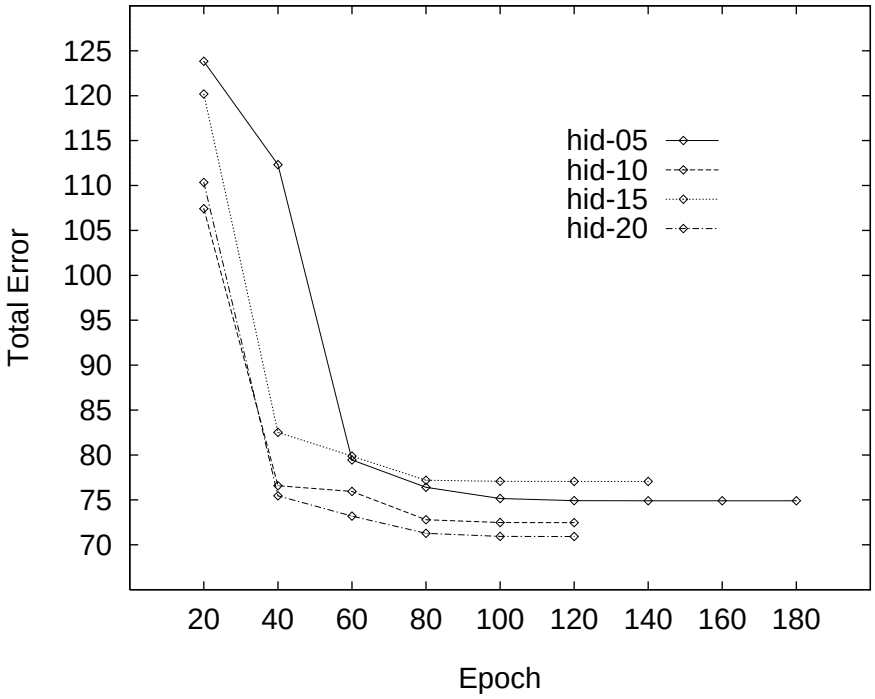


Figure 4.4. Training on the first set (first experiment)

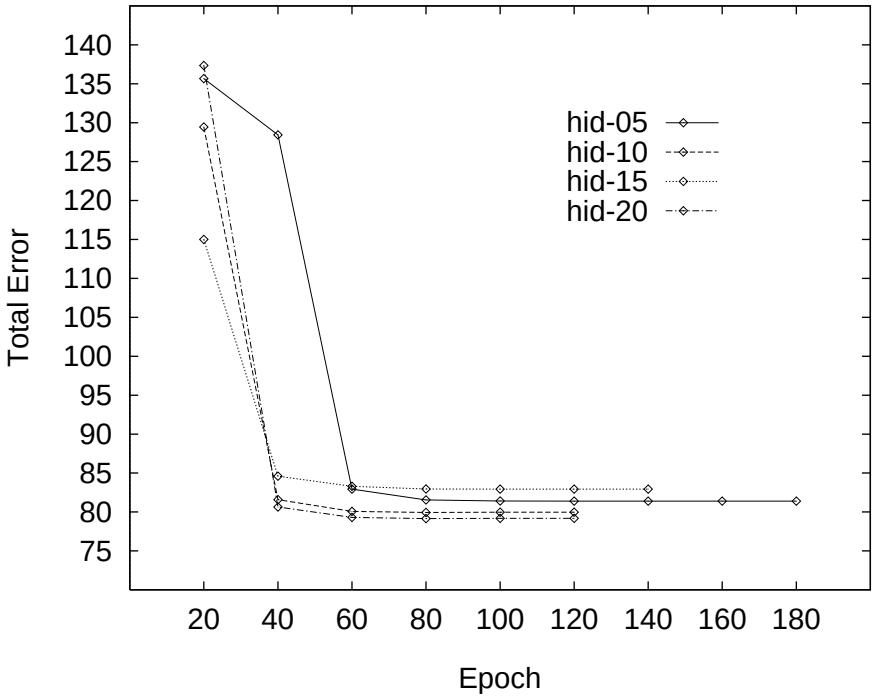


Figure 4.5. Testing on the second set (first experiment)

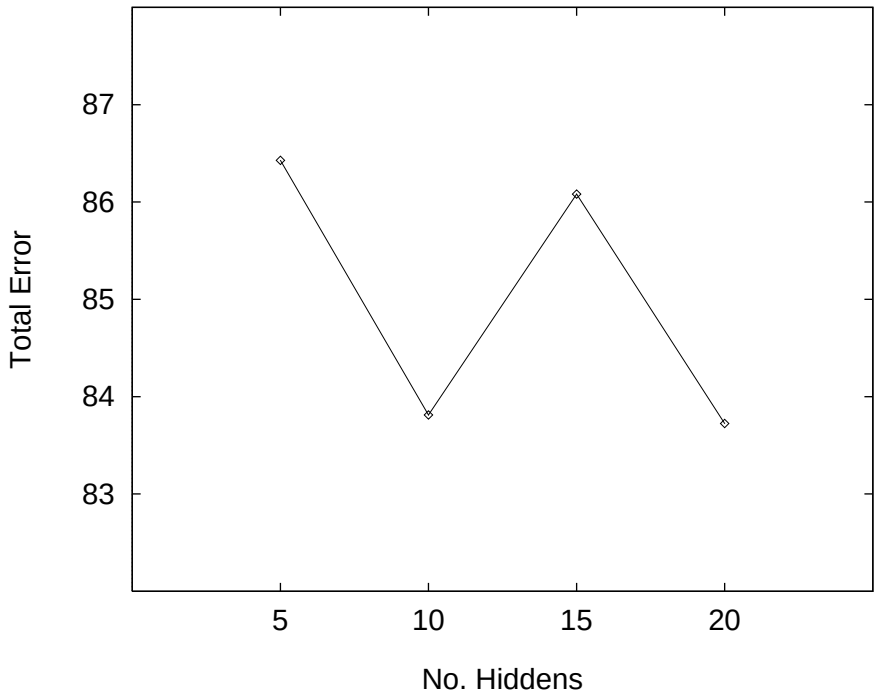


Figure 4.6. Testing on the third set (first experiment)

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 activations of the hidden units given by each pattern in the fourth set. Figure 4.7 plots the two first principal components (PCs) for the patterns in the fourth set. Figures 4.8 and 4.9 plot, respectively, the two first principal components (PCs) for the negative and positive patterns in the fourth set which were correctly classified by the neural model. It can be verified that there is no correlation between negative and positive patterns, for they are placed in different clusters. Therefore, the internal representations stored in the hidden units make a distinction between the two types of patterns.

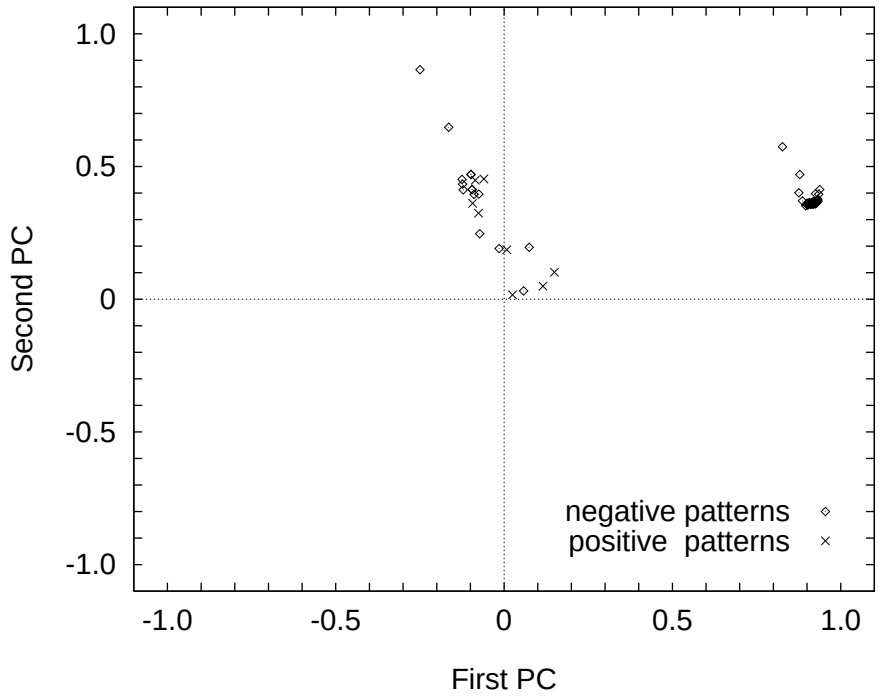


Figure 4.7. Two first PCs for the patterns in the fourth set (first experiment)

[4]Principal component analysis is a technique to change the system of coordinates of a multivariate data set. The new coordinates, which are linear transformations of the original ones, account for decreasing degrees of variation of the data.

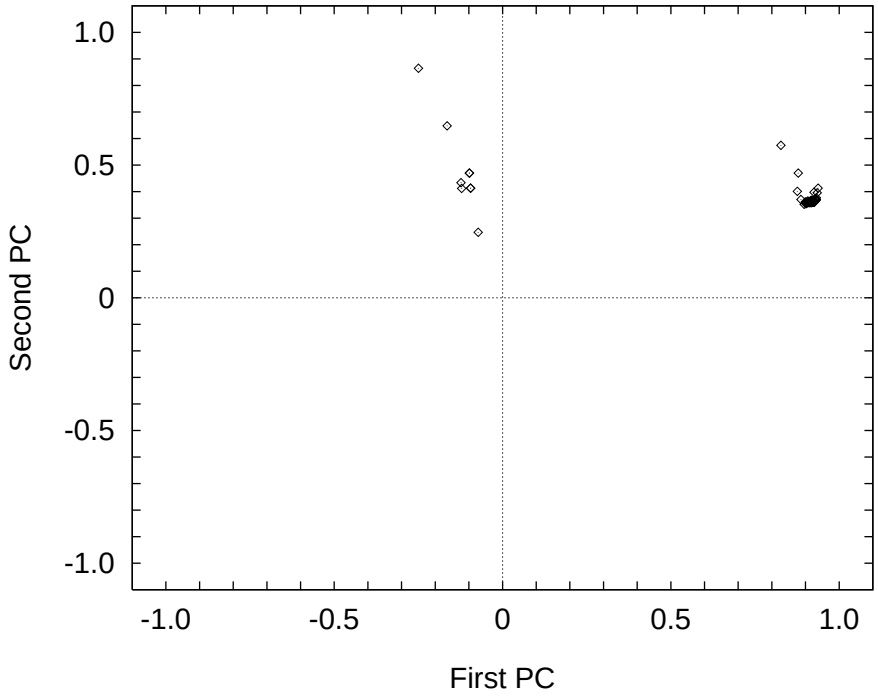


Figure 4.8. Two first PCs for the negative patterns correctly classified (first experiment)

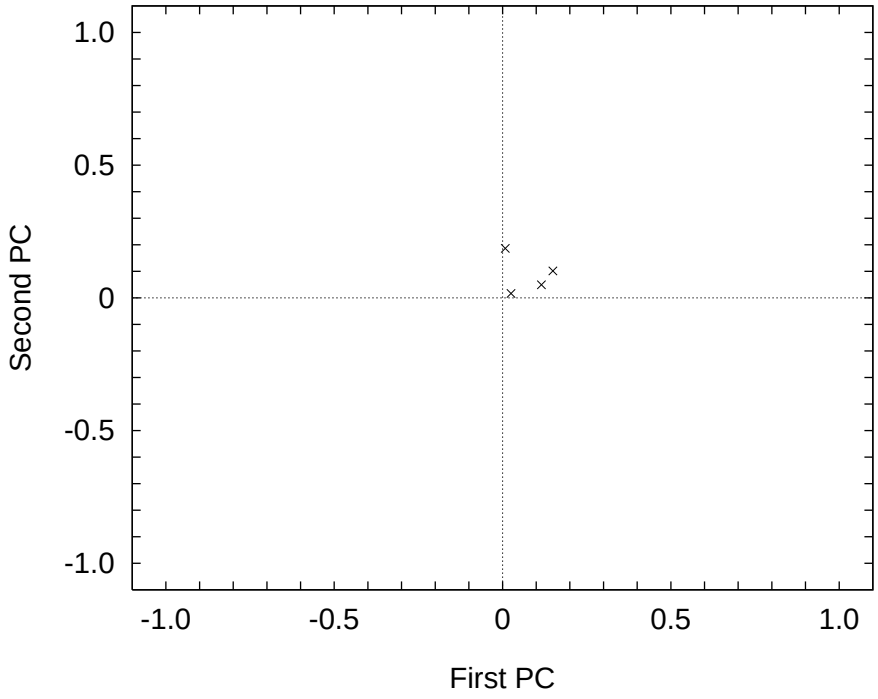


Figure 4.9. Two first PCs for the positive patterns correctly classified (first experiment)

Two negative patterns from each cluster of the plot in figure 4.8, and one positive pattern from the cluster of the plot in figure 4.9 were selected. The internal representations stored in the hidden units for the three patterns are shown in Hinton's diagram of figure 4.10. It can be noticed that the representations are not distributed, and that few units contribute effectively to them.

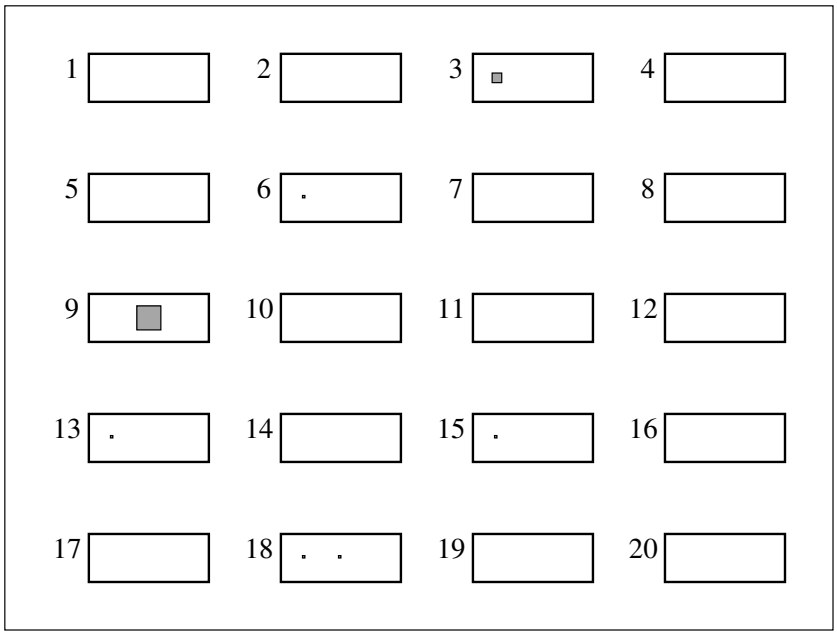


Figure 4.10. Hinton's diagram for three patterns (first experiment) — rectangles correspond to hidden units; squares on the left correspond to the negative pattern in the cluster on the left of figure 4.8; squares on the centre correspond to the negative pattern in the cluster on the right of figure 4.8; squares on the right correspond to the positive pattern in the cluster of figure 4.9;

The model was evaluated on six musical pieces from Bach. The first two pieces were the ninth and thirteenth two-part inventions in F minor and A minor (see section 2.4.5.5). The third and fourth were the third and fourteenth three-part inventions in D major and B flat major (see section 2.4.5.5). The last two pieces were the fourth and seventeenth fugues in C sharp minor and A flat major of the Well-Tempered Clavier (see section 2.4.5.6).

Each voice of each piece was input separately. The size of the window was wide enough to cover 99% of the instances of segmentation present in the pieces. Activations less than 0.4 and greater than 0.6 were considered as 0 and 1 respectively in the output units. Results are displayed in table 4.3.

Apart from misclassifications in the fourth piece, the overall percentage of misclassifications is very low. The high percentage of misclassifications in positive patterns in the fourth piece were due to the occurrence of many instances of one pattern, which was not learnt by the neural model. These misclassifications could be avoided by including more instances of such pattern in the training set.

4.5 Second experiment

The second experiment was on recognizing cases of segmentation given by longer durations. According to Lerdahl and Jackendoff's rule presented in the second chapter (section 2.3.4.4), segmentation given by longer durations occurs whenever, in a group of four notes, the duration of the second note is longer than those of the first and third notes.

As in the first experiment (section 4.4), three sets of patterns were generated, each containing

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 of note sustained events contained by the second part is greater than that contained by the first and the third parts. The negative templates are all those otherwise.

As in the first experiment, we might illustrate the generation of a pattern in an example. Let us consider the negative template in which the first, second, third, and last parts are given respectively by one note onset followed by three notes sustained, one note onset followed by two notes sustained, one note onset followed by one note sustained, and one note onset. There are nine free slots — seven at the beginning, and two at the end of the template — and thus, according to table 4.5, one negative pattern is generated from the template. The generated pattern consists of seven events which are determined randomly, followed by the events in the four parts of the template, followed by two events which are determined randomly.

The training method was identical to that followed in the previous experiment. The training and testing on the sets for four configurations is shown in figures 4.11, 4.12, and 4.13. The window in the input layer held 19 pairs of units, and initial weights were set randomly for each configuration. Best performance was given by a configuration with 10 hidden units, whose training lasted for 240 epochs.

Table 4.4. Generative templates of the pattern sets (second experiment) — R: rest; NS: note sustained; NO: note onset; *FS*: free slot;

Pattern Templates
(12×FS) [R] [R] [R] [R] *(3×FS)*
(12×FS) [R] [R] [R] [NO] *(3×FS)*
(12×FS) [R] [R] [NO] [R] *(3×FS)*
(12×FS) [R] [R] [NO] [NO] *(3×FS)*
(12×FS) [R] [R] [NO NS] [R] *(2×FS)*
(12×FS) [R] [R] [NO NS] [NO] *(2×FS)*
(12×FS) [R] [R] [NO 2×NS] [R] *(FS)*
(12×FS) [R] [R] [NO 2×NS] [NO] *(FS)*
(12×FS) [R] [R] [NO 3×NS] [R]
(12×FS) [R] [R] [NO 3×NS] [NO]
(12×FS) [R] [NO] [R] [R] *(3×FS)*
(12×FS) [R] [NO] [R] [NO] *(3×FS)*
... ...
(12×FS) [R] [NO] [NO 3×NS] [R]
(12×FS) [R] [NO] [NO 3×NS] [NO]
... ...
(3×FS) [R] [NO 9×NS] [R] [R] *(3×FS)*
(3×FS) [R] [NO 9×NS] [R] [NO] *(3×FS)*
... ...
(3×FS) [R] [NO 9×NS] [NO 3×NS] [R]
(3×FS) [R] [NO 9×NS] [NO 3×NS] [NO]
... ...
(9×FS) [NO 3×NS] [R] [R] [R] *(3×FS)*
(9×FS) [NO 3×NS] [R] [R] [NO] *(3×FS)*
... ...
(9×FS) [NO 3×NS] [R] [NO 3×NS] [R]
(9×FS) [NO 3×NS] [R] [NO 3×NS] [NO]
(9×FS) [NO 3×NS] [NO] [R] [R] *(3×FS)*
(9×FS) [NO 3×NS] [NO] [R] [NO] *(3×FS)*
... ...
(9×FS) [NO 3×NS] [NO] [NO 3×NS] [R]
(9×FS) [NO 3×NS] [NO] [NO 3×NS] [NO]
... ...
[NO 3×NS] [NO 9×NS] [R] [R] *(3×FS)*
[NO 3×NS] [NO 9×NS] [R] [NO] *(3×FS)*
... ...
[NO 3×NS] [NO 9×NS] [NO 3×NS] [R]
[NO 3×NS] [NO 9×NS] [NO 3×NS] [NO]

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	—

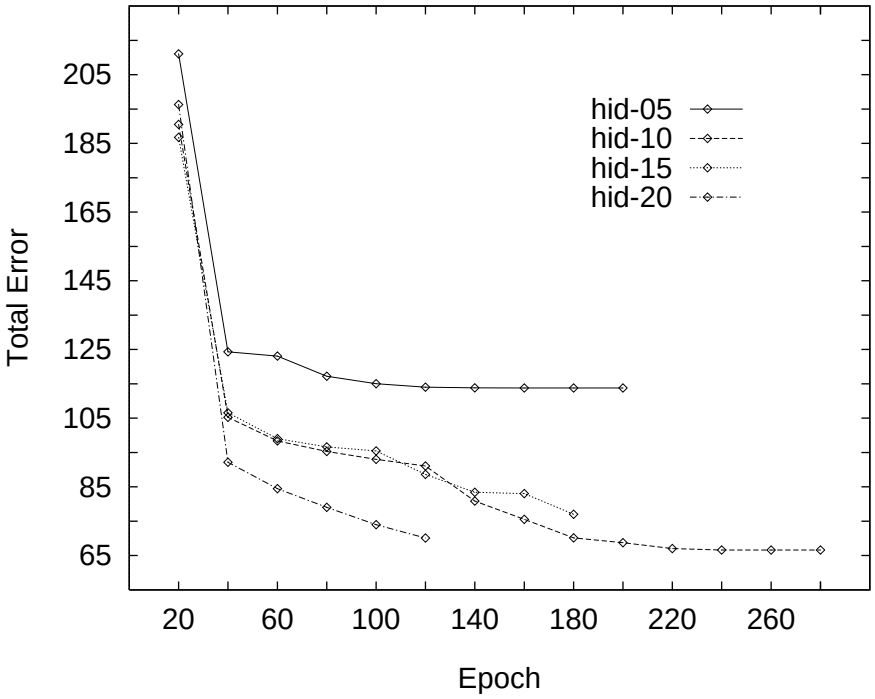


Figure 4.11. Training on the first set (second experiment)

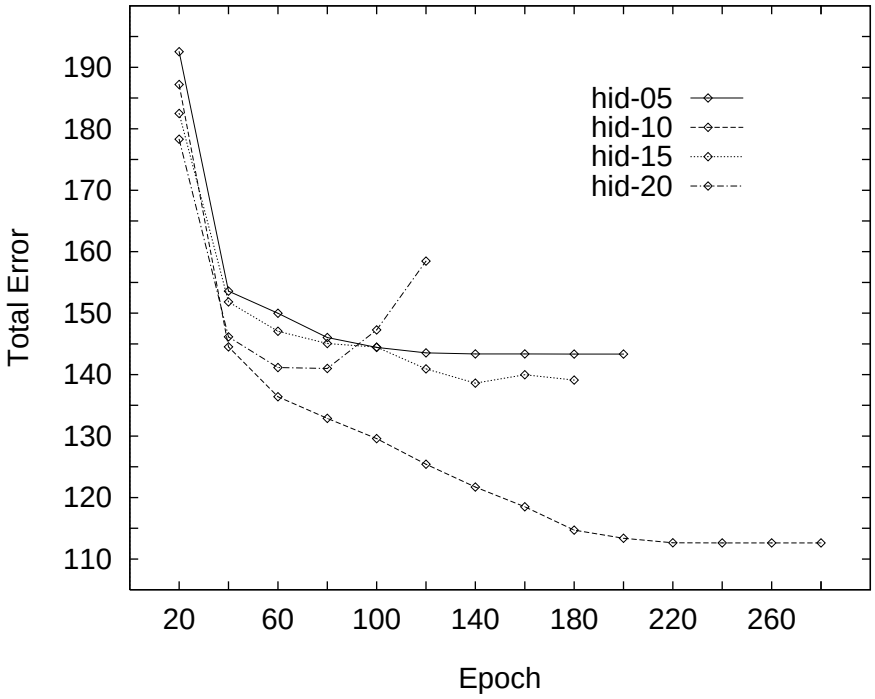


Figure 4.12. Testing on the second set (second experiment)

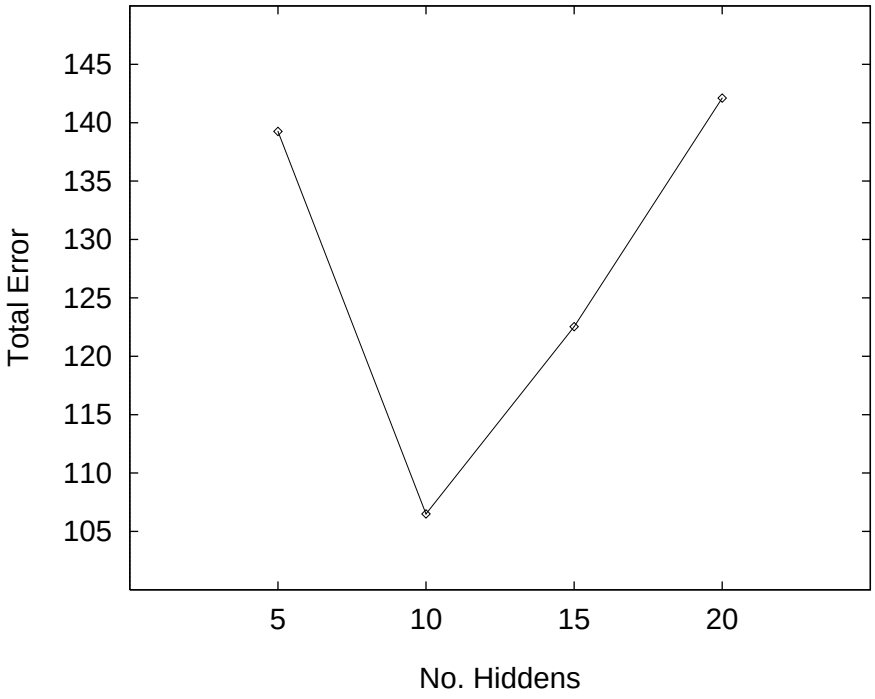


Figure 4.13. Testing on the third set (second experiment)

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.

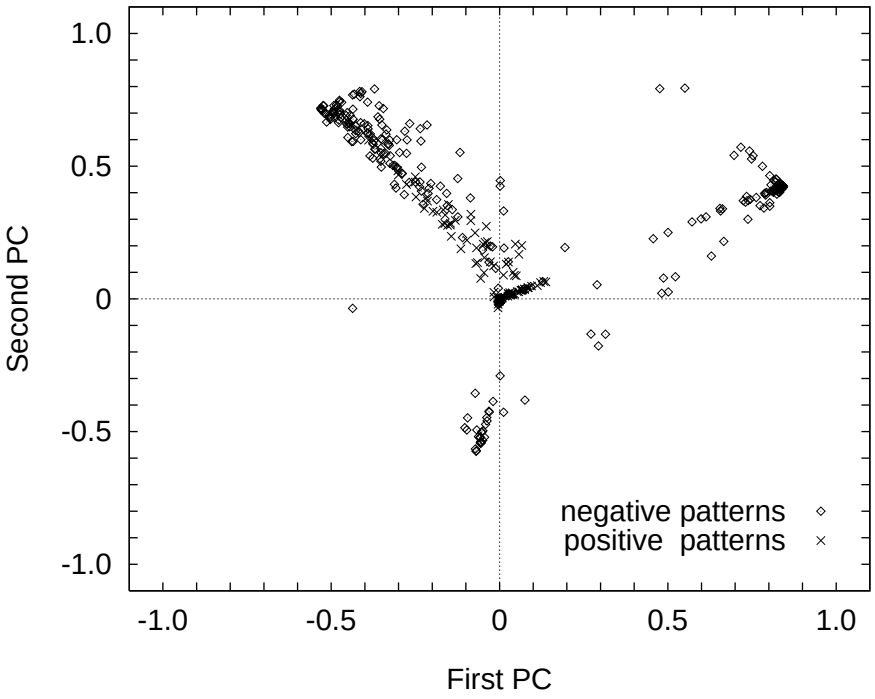


Figure 4.14. Two first PCs for the patterns in the fourth set (second experiment)

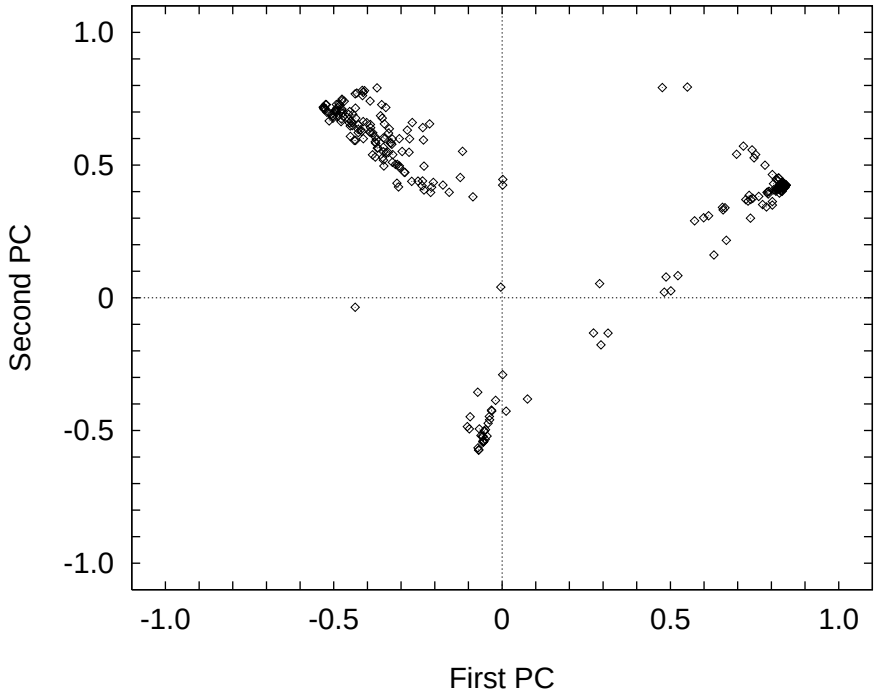


Figure 4.15. Two first PCs for the negative patterns correctly classified (second experiment)

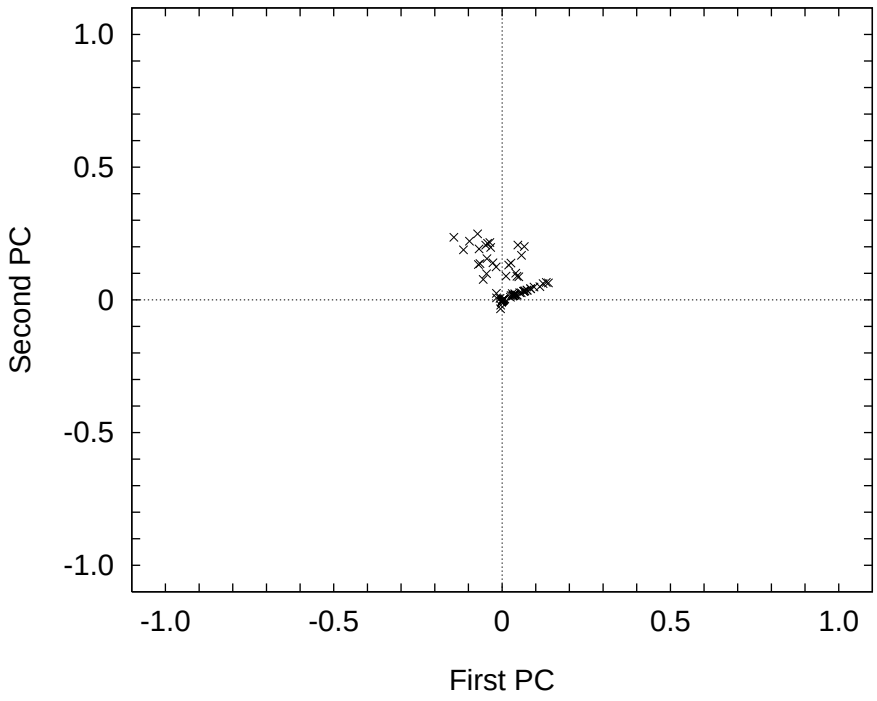


Figure 4.16. Two first PCs for the positive patterns correctly classified (second experiment)

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	2253	39	18	8
5	4483	98	26	15
6	2185	48	22	6

Three negative patterns from each cluster of the plot in figure 4.15, and one positive pattern from the cluster of the plot in figure 4.16 were selected. Their internal representations, stored in the hidden units, are shown in Hinton's diagram of figure 4.17. Two facts may be observed. Firstly, the representations are not distributed, and secondly, few units contribute effectively to them.

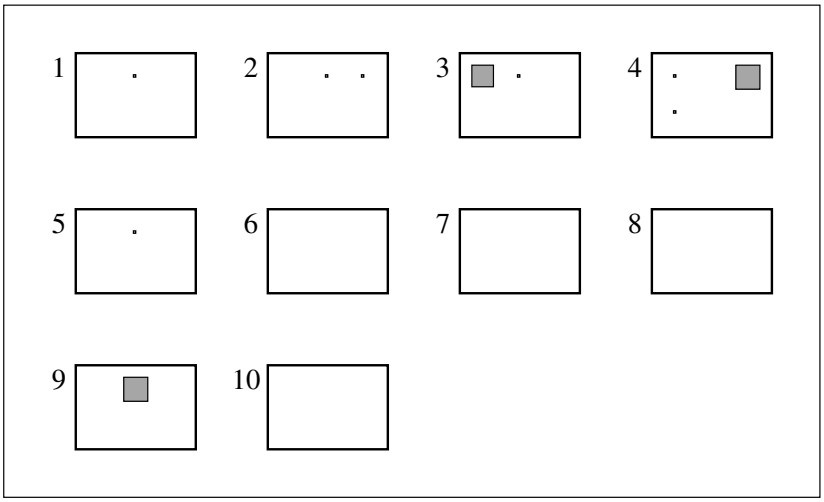


Figure 4.17. Hinton's diagram for four patterns (second experiment) — rectangles correspond to hidden units; squares on the top left correspond to the negative pattern in the cluster on the left of figure 4.15; squares on the top centre correspond to the negative pattern in the cluster on the bottom of figure 4.15; squares on the top right correspond to the negative pattern in the cluster on the right of figure 4.15; squares on the bottom left correspond to the positive pattern in the cluster of figure 4.16;

The model was evaluated on the same musical pieces as in the first experiment (section 4.4). Each voice of each piece was input separately. The window size was able to cover 87% of the instances of segmentation which occur in the pieces. Activations less than 0.4 and greater than 0.6 were regarded as 0 and 1 respectively in the output units.

Results are presented in table 4.6. The percentage of misclassifications is low. However, it could be further reduced or even avoided by increasing the number of patterns in the training set.

4.6 Third experiment

The third experiment was on recognizing cases of segmentation given by breaks of similarity. According to Lerdahl and Jackendoff's rule presented in the second chapter (section 2.3.4.5), seg-

Table 4.7. Generative templates of the pattern sets (third experiment) — R: rest; NS: note sustained; NO: note onset; *FS*: free slot;

Pattern Templates
(2×FS) [R] [NO] [NO NS] [NO NS] [NO NS] [NO NS] [NO] [NO NS] *(FS)*
(2×FS) [NO] [R] [NO NS] [NO NS] [NO] [NO] [NO NS] [NO NS] *(2×FS)*
(2×FS) [NO NS] [NO NS] [R] [NO] [NO] [NO] [NO NS] [NO NS] *(2×FS)*
(2×FS) [NO NS] [NO] [NO NS] [R] [NO] [NO] [NO] [NO NS] *(3×FS)*
(2×FS) [NO] [NO] [NO NS] [NO NS] [R] [NO NS] [NO NS] [NO] *(2×FS)*
(2×FS) [NO] [NO] [NO NS] [NO NS] [NO NS] [R] [NO NS] [NO NS] *(FS)*
(3×FS) [NO] [NO] [NO NS] [NO] [NO] [NO NS] [R] [NO] *(3×FS)*
(FS) [NO] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [R] *(FS)*
(3×FS) [R] [NO NS] [NO] [R] [NO NS] [NO] [NO] [NO] *(3×FS)*
(4×FS) [R] [NO] [NO] [NO] [NO] [NO NS] [R] [NO] *(3×FS)*
(2×FS) [NO NS] [R] [NO NS] [NO] [R] [NO NS] [NO NS] [NO NS] *(FS)*
(3×FS) [NO] [R] [NO] [NO NS] [NO NS] [NO] [NO NS] [R] *(2×FS)*
(2×FS) [NO] [NO NS] [R] [NO NS] [NO] [R] [NO] [NO NS] *(3×FS)*
(2×FS) [NO] [NO NS] [R] [NO NS] [NO] [NO] [NO NS] [R] *(3×FS)*
(2×FS) [NO NS] [NO] [NO NS] [R] [NO] [NO] [R] [NO NS] *(3×FS)*
(3×FS) [NO] [NO] [NO] [NO NS] [R] [R] [NO NS] [NO NS] *(2×FS)*
(2×FS) [R] [NO NS] [NO NS] [NO] [R] [NO] [NO] [R] *(4×FS)*
(3×FS) [NO] [R] [NO NS] [R] [NO NS] [NO NS] [R] [NO NS] *(FS)*
(2×FS) [NO] [NO NS] [R] [NO NS] [NO NS] [R] [NO NS] [R] *(2×FS)*
(FS) [NO NS] [NO NS] [NO NS] [R] [R] [R] [NO NS] [NO NS] *(2×FS)*
(2×FS) [R] [NO NS] [NO NS] [R] [R] [NO NS] [NO] [R] *(3×FS)*
(4×FS) [NO] [R] [R] [NO] [NO NS] [R] [R] [NO NS] *(2×FS)*
(3×FS) [R] [NO] [R] [NO NS] [R] [NO] [R] [NO NS] *(3×FS)*
(2×FS) [NO NS] [R] [NO NS] [R] [NO NS] [R] [NO NS] [R] *(2×FS)*
(3×FS) [R] [NO NS] [R] [R] [R] [NO] [NO NS] [R] *(3×FS)*
(2×FS) [NO NS] [R] [R] [NO NS] [R] [R] [R] [NO] *(4×FS)*
(3×FS) [R] [NO NS] [R] [NO] [R] [NO] [R] [R] *(4×FS)*
(3×FS) [R] [R] [NO NS] [R] [NO] [R] [NO] [R] *(4×FS)*
(4×FS) [R] [NO] [R] [R] [R] [R] [NO NS] [R] *(3×FS)*
(4×FS) [NO] [R] [R] [R] [R] [R] [R] [NO NS] *(3×FS)*
(4×FS) [R] [R] [R] [R] [R] [R] [NO NS] [R] *(3×FS)*
(3×FS) [NO NS] [R] [R] [R] [R] [R] [R] [R] *(4×FS)*
(4×FS) [R] [R] [R] [R] [R] [R] [R] [R] *(4×FS)*
(4×FS) [NO] [NO] [NO] [NO] [NO] [NO] [NO] [NO] *(4×FS)*
(4×FS) [NO] [NO] [NO] [NO] [NO] [NO] [NO] [NO NS] *(3×FS)*
... ...
[NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO] *(FS)*
[NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS] [NO NS]

mentation given by breaks of similarity occurs whenever, in a group of eight notes, the durations of the first four notes are identical, the durations of the last four notes are also identical, and the durations of the first four notes are different from those of the last four.

As in previous experiments (sections 4.4 and 4.5), three sets of patterns were generated, each containing 1500 negative and 1500 positive patterns. The sets were generated from the pattern templates described in table 4.7. The number of patterns produced from a template was also determined by 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 4.8.

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, trained for 180 epochs.

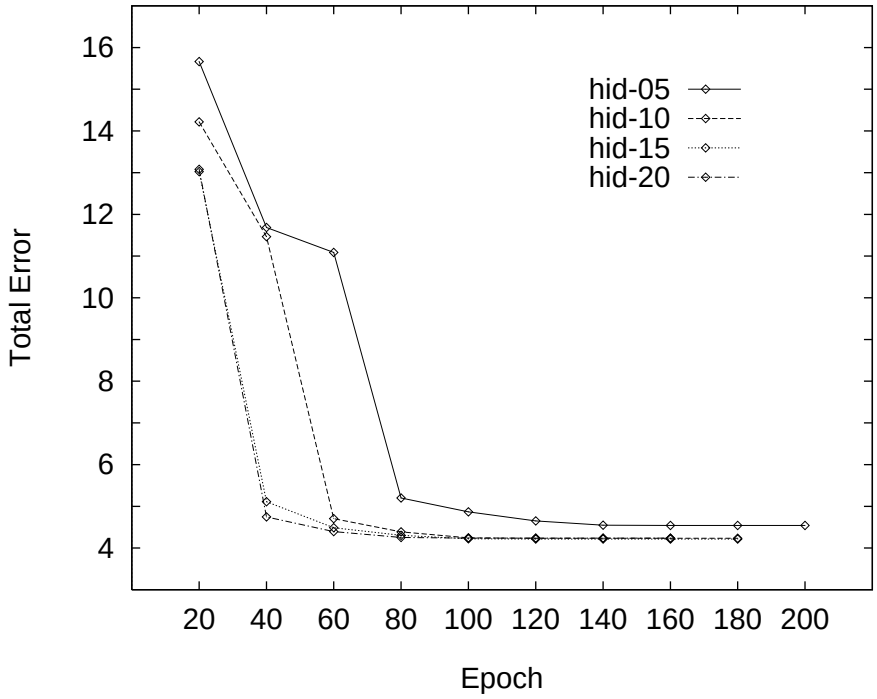


Figure 4.18. Training on the first set (third experiment)

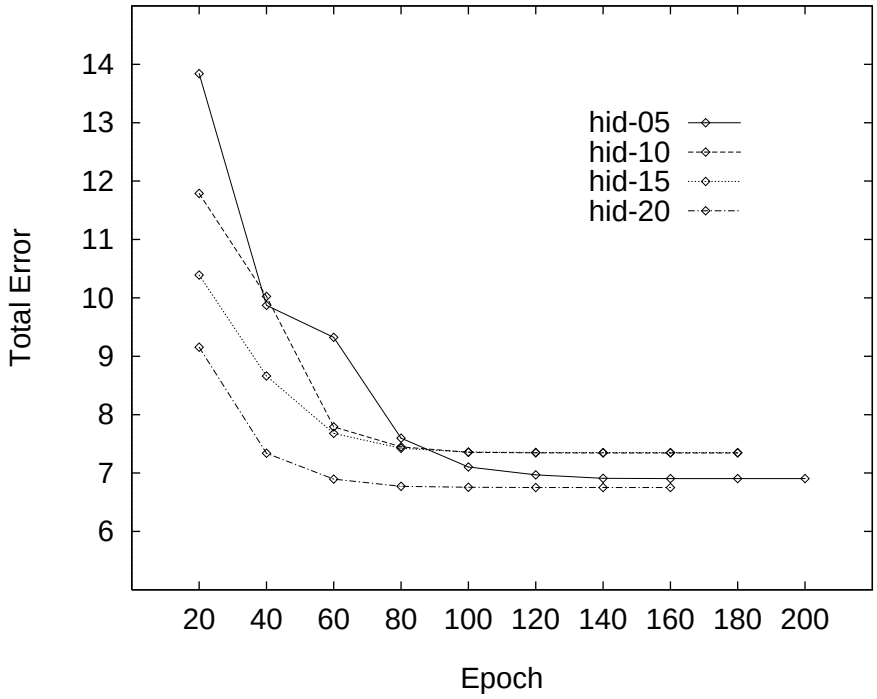


Figure 4.19. Testing on the second set (third experiment)

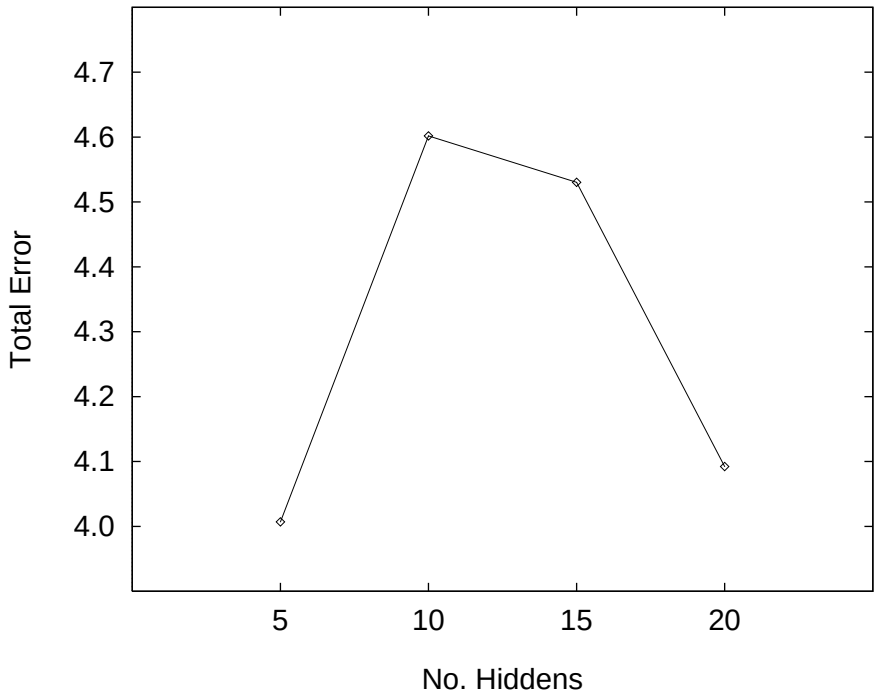


Figure 4.20. Testing on the third set (third experiment)

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.

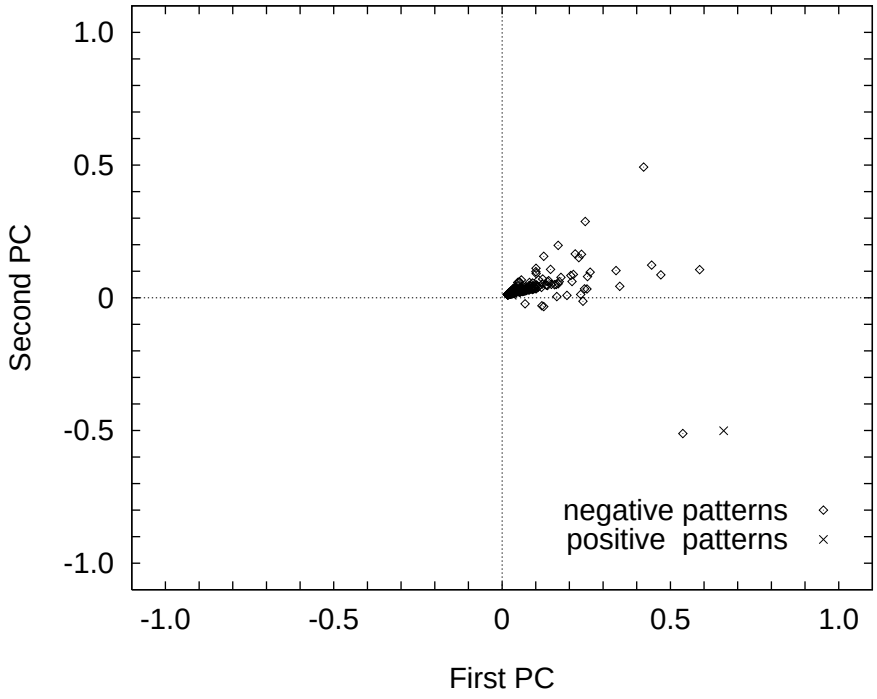


Figure 4.21. Two first PCs for the patterns in the fourth set (third experiment)

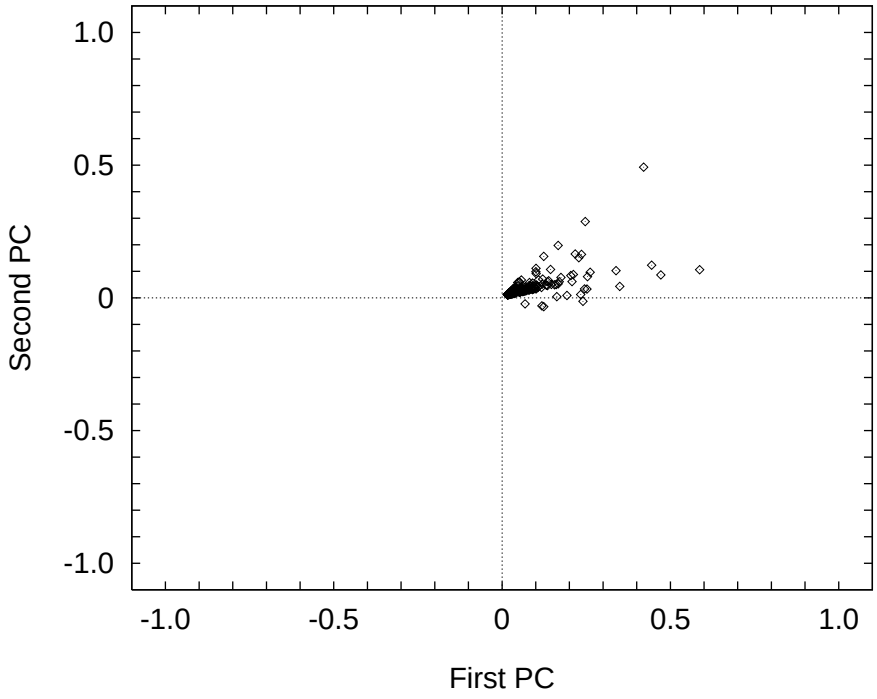


Figure 4.22. Two first PCs for the negative patterns correctly classified (third experiment)

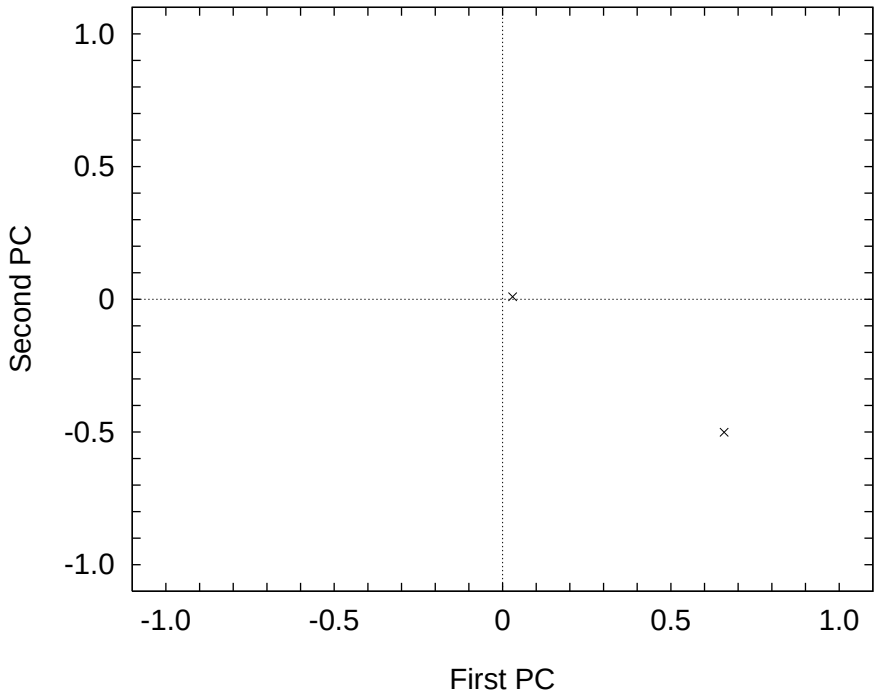


Figure 4.23. Two first PCs for the positive patterns correctly classified (third experiment)

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 distributed, and that most of the units take part effectively in the representations.

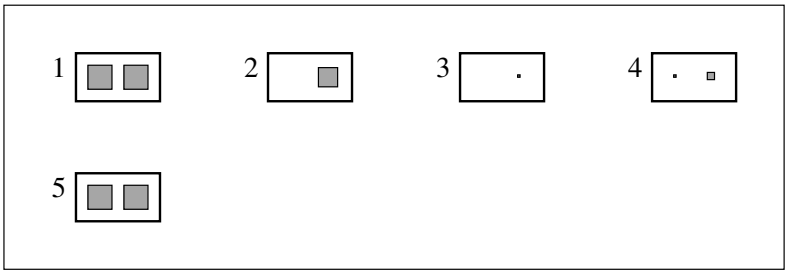


Figure 4.24. Hinton's diagram for two patterns (third experiment) — rectangles correspond to hidden units; squares on the left correspond to the negative pattern in the cluster of figure 4.22; squares on the right correspond to the positive pattern farthest from origin in figure 4.23;

The model was evaluated on the same musical pieces as in former experiments (sections 4.4 and 4.5). Each voice of each piece was input separately. The window size was wide enough to cover 93% of the instances of segmentation existent in the pieces. Activations less than 0.4 and greater than 0.6 were considered as 0 and 1 respectively in the output units.

Results are displayed in table 4.9. The percentage of misclassifications is very low. It could be further reduced or even avoided, by increasing the number of patterns in the training set.

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;

Pieces	#NP	#PP	%NM	%PM
1	813	3	3	0
2	772	28	6	0
3	1196	4	5	0
4	2302	1	0	0
5	4589	8	2	0
6	2233	7	3	0

4.7 Summary

This chapter introduces a novel representation for rhythmic sequences, and a neural model to segment musical pieces according to three cases of rhythmic segmentation. The model has a topology which is similar to that of Sejnowski and Rosenberg's model, that is, a three-feedforward-layered backpropagation network with windowed input. It was trained and tested on sets of contrived patterns, and successfully applied to six musical pieces from Bach. The results presented here suggest that cognitive mechanisms which recognize the three cases of rhythmic segmentation and perform segmentation according to the three Lerdahl and Jackendoff's grouping rules can be modelled by a windowed artificial neural model with supervised learning (Carpinteiro, 1995).

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) concerning recognition of melodies in the presence of distractors — foreign notes interleaved with melodies — and Palmer and Holleran's (1994) studies (section 2.4.4) concerning recognition of alterations in melodies in multivoiced music, have provided a few clues for research in perception of themes in polyphonic music.

The studies suggest that recognition of a theme presented in a polyphonic voice is affected when it overlaps with other voices. Yet, active search for a theme can lead to discerning it even in midst of overlapping voices. The studies also suggest that listeners attend more readily to the voice presenting the theme, specially when, in comparison with other polyphonic voices, it occurs in the higher-frequency range. In the light of these studies therefore, for our experiments presented in this chapter, we assume that listeners do have the ability to detect thematic material in polyphonic music.

This chapter is concerned with thematic recognition in polyphonic music. It proposes original representations for musical sequences as well as an original artificial neural model for thematic recognition.

The chapter comprises nine sections, the first of which is this introduction. The second, third, and fourth sections introduce representations developed for binary and musical sequences. The fifth section describes the artificial neural model. The sixth, seventh, and eighth sections are concerned with experimentation. Finally, the ninth section provides a summary of the chapter.

5.2 Representation for binary sequences

The representation described here was employed to represent input data in the first experiment (section 5.6). The data consisted of binary sequences, which were input one bit at a time. Two neural units and a time integrator were used in the representation. The time integrator provides the input units with their former activation values decayed in time. Bits 0's of the sequences were represented as (1,0), whereas bits 1's were represented as (0,1).

As an example, let us consider the six-bit sequence **S** = 101100, and a time integrator with decay rate of 0.5 applied to input units. The integrated value of activations of the units is then given by the values in table 5.1.

Table 5.1. Representation for a binary sequence

Time	Input Bit	Activation of Unit 1	Activation of Unit 2
1	1	—	1.0
2	0	1.0	0.5
3	1	0.5	1.25
4	1	0.25	1.625
5	0	1.125	0.813
6	0	1.563	0.406

5.3 Representation for univoiced musical sequences

The representation described in this section is based on interval representations. It is plausible because is supported by studies reviewed in the second chapter (section 2.2). The studies indicate that individuals hold internal representations for intervals and contour. However, the studies do not specify the type of those representations.

The representation makes use of the concept behind the representation for rhythmic sequences employed in the fourth chapter (section 4.2), that is, the division of a musical piece into equal size time intervals. It may be used to represent any univoiced musical sequence (see section 2.4.5.1), although it has been used, in particular, to represent the input data in the second experiment (section 5.7).

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

Fifteen neural units are used in the representation. Each unit represents one musical interval ranging from an octave down to an octave up. As mentioned in section 4.2, at each TIC, either there is a rest, or a note is onset, or a note is sustained. If there is a rest, none of the input units receives activation. If a note is onset, as it makes up an interval, the unit corresponding to that interval receives an activation of 1.0. If it is sustained, the unit which corresponds to the interval receives 0.5.

The reason for conceding activation to notes sustained is to preserve a certain amount of activation in units corresponding to intervals which were formerly onset, so that, the amount of activation in those units be significant with relation to the activation value of the unit corresponding to the interval currently onset. We might illustrate these ideas in an example.

Let us consider the univoiced musical sequence in figure 5.1, and a time integrator with decay rate of 0.5 applied to input units. The sequence lasts 12 TICs, and TI (see section 4.2) is a sixteenth note. Table 5.2 displays the integrated value of activations of the units when considering the case in which notes sustained receive activation. This case was employed in the second experiment. Table 5.3, in its turn, displays the opposite case, which was not considered in any of the experiments. By setting the values in table 5.3 against those in table 5.2, one may verify that the values in the latter table accounts much better for the memory trace of intervals occurring in past times.

5.4 Representation for multivoiced musical sequences

As the representation introduced in section 5.3, the representation described here is also based on interval representations. The representation is plausible because it is supported by studies reviewed in the second chapter (section 2.2). The studies indicate that individuals hold internal representations for intervals and contour, although they do not specify the type of those representations.

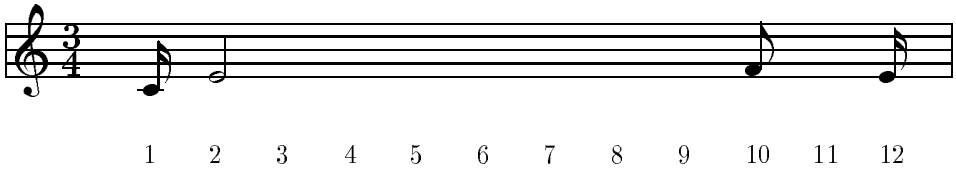


Figure 5.1. An univoiced musical sequence

Table 5.2. Representation for the musical sequence in figure 5.1 (case I)

TIC	Activations of Units Representing Intervals								
	-8	...	-3	-2	0	+2	+3	...	+8
1	—	...	—	—	—	—	—	...	—
2	—	...	—	—	—	—	1.0	...	—
3	—	...	—	—	—	—	1.0	...	—
4	—	...	—	—	—	—	1.0	...	—
5	—	...	—	—	—	—	1.0	...	—
6	—	...	—	—	—	—	1.0	...	—
7	—	...	—	—	—	—	1.0	...	—
8	—	...	—	—	—	—	1.0	...	—
9	—	...	—	—	—	—	1.0	...	—
10	—	...	—	—	—	1.0	0.5	...	—
11	—	...	—	—	—	1.0	0.25	...	—
12	—	...	—	1.0	—	0.5	0.125	...	—

The representation makes use of the concept behind the representation for rhythmic sequences employed in the fourth chapter (section 4.2), that is, the division of a musical piece into equal size time intervals. It may be used to represent any multivoiced musical sequence (see section 2.4.5.1),

Table 5.3. Representation for the musical sequence in figure 5.1 (case II)

TIC	Activations of Units Representing Intervals								
	-8	...	-3	-2	0	+2	+3	...	+8
1	—	...	—	—	—	—	—	...	—
2	—	...	—	—	—	—	1.0	...	—
3	—	...	—	—	—	—	0.5	...	—
4	—	...	—	—	—	—	0.25	...	—
5	—	...	—	—	—	—	0.125	...	—
6	—	...	—	—	—	—	0.063	...	—
7	—	...	—	—	—	—	0.031	...	—
8	—	...	—	—	—	—	0.016	...	—
9	—	...	—	—	—	—	0.008	...	—
10	—	...	—	—	—	1.0	0.004	...	—
11	—	...	—	—	—	0.5	0.002	...	—
12	—	...	—	1.0	—	0.25	0.001	...	—

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 listeners are able to separate out voices if intervals between notes in the voices are greater than a fifteenth. 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.

As opposed to the representation in section 5.3, the representation for multivoiced musical sequences has to take into consideration all musical voices, and therefore, is more complex because the voices interact. The representation assumes three facts. First, any note onset occurring in a TIC makes up an interval with all notes onset or sustained which occurred in the TIC immediately before. Second, the representation does represent the intervals which occur in a TIC, but does not represent multiple instances occurring through the voices in the TIC. Third, at any given TIC, an interval corresponding to a note onset masks any occurrence of the same interval corresponding to a note sustained. The example below illustrates the representation, and the three assumptions.

Let us consider the multivoiced musical sequence in figure 5.2, and a time integrator with decay rate of 0.5 applied to input units. The sequence lasts 8 TICs, and TI (see section 4.2) is a sixteenth note. Units corresponding to intervals made up by a note onset or note sustained receive an activation of 1.0 and 0.5 respectively. Table 5.4 displays then the integrated value of activations of the units when inputting, TIC by TIC, the sequence in figure 5.2.

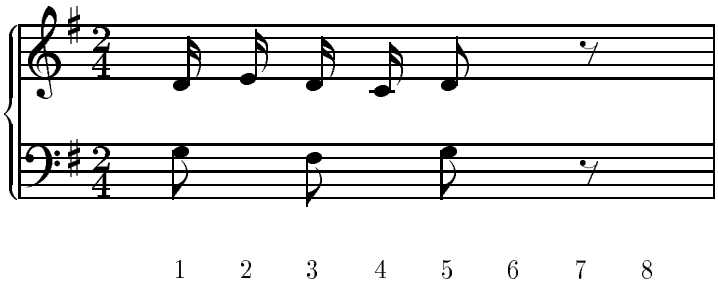


Figure 5.2. A multivoiced musical sequence

Table 5.4. Representation for the musical sequence in figure 5.2

TIC	Activations of Units Representing Intervals												
	-8	-7	...	-4	-3	-2	0	+2	...	+5	+6	+7	+8
1	—	—	...	—	—	—	—	—	...	—	—	—	—
2	—	—	...	—	—	—	—	1.0	...	—	1.0	—	—
3	—	1.0	...	—	—	1.0	—	0.5	...	1.0	0.5	—	—
4	—	1.0	...	—	—	1.5	—	0.25	...	1.5	0.25	—	—
5	—	0.5	...	1.0	—	0.75	—	1.125	...	0.75	1.125	—	—
6	—	0.25	...	1.0	—	0.375	—	1.063	...	0.375	1.063	—	—
7	—	0.125	...	0.5	—	0.188	—	0.531	...	0.188	0.531	—	—
8	—	0.063	...	0.25	—	0.094	—	0.266	...	0.094	0.266	—	—

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 consequently, is biologically more plausible. The time integrator, nevertheless, is applied to input units as opposed to output units as in their model.

Thematic recognition sets three conditions on the model. First, it demands that the model be able to recognize both a set of input sequences and a set of sub-sequences within a large and unique input sequence. The model is required to recognize a set of input sequences when segmentation is performed on the musical piece. Otherwise, when segmentation is not performed, the entire piece consists of a unique input sequence, and the model is thus required to recognize sub-sequences of that sequence.

Second, it demands that the model classify sequences (or sub-sequences) properly in the presence of noise. The reason follows from the fact that any two sequences which differ slightly must achieve similar classifications.

Third, it demands that the model recognize sequences (or sub-sequences) in a very precise form. The reason for the latter is that any two sequences which share either some intervals, or even all intervals, but in an alternative order or rhythm, are musically different, and as a consequence, must be recognized as distinct.

The model is shown in figure 5.3. It is made up of two self-organizing maps (SOMs), one on top of the other. The idea of having neural nets placed in such a hierarchical way is not original. Gjerdingen has proposed it before, although he has not explored it fully (see section 3.2.3). Our approach differs from his in that we explore the idea using SOM nets, whereas he proposed it for ART2 (Carpenter & Grossberg, 1987) nets.

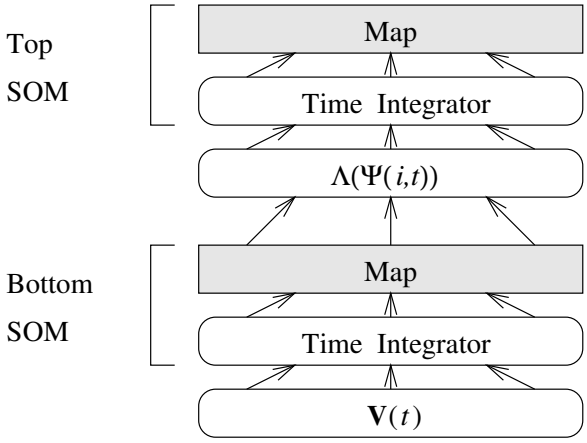


Figure 5.3. The model

The input to the model is a sequence in time of m-dimensional vectors, $\mathbf{S_1} = \mathbf{V}(1), \mathbf{V}(2), \ldots, \mathbf{V}(t), \ldots, \mathbf{V}(z)$, where the components of each vector are non-negative real values. The sequence is presented to the input layer of the bottom SOM, one vector at a time. The input layer has m units, one for each component of the input vector $\mathbf{V}(t)$, and a time integrator. The activation $\mathbf{X}(t)$ of the units in the input layer is given by

$$\mathbf{X}(t) = \mathbf{V}(t) + \delta_1 \mathbf{X}(t-1) \tag{5.1}$$

where $\delta_1 \in (0,1)$ is the decay rate. The winning unit i^* in the map[1] is the unit which has the

[1] Also known as array, grid, or output layer.

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 Ψ of the units i, no matter how close or far away they be from the winning unit i^*. The disadvantage is that the accuracy of the input delivered to the top SOM relies heavily upon the quality of the classifications in the bottom SOM. For instance, it is difficult to find a suitable radius ρ to the gaussian when, for each vector $\mathbf{V}(t)$, the distances of the winning units $\Psi(i^*,t)$ vary on a wide range. Alternatively, Λ may be defined as

$$\Lambda(\Psi(i,t)) = \begin{cases} 1 - \kappa\Phi(i,i^*) & \text{if } i \in N^* \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

where κ is a constant, and N^* is a neighbourhood of the winning unit. The main advantage is that it is simple to use this type of function. Yet, depending on the radius of the neighbourhood chosen, either pertinent information can be discarded or non-pertinent information be included. We have not considered the situation in which information given by units far away from the neighbourhood N^* is lost, for, as mentioned before, the neighbourhood interaction function avoided the existence of local winning units.

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 in sequences is not original. Wickelphones (Wickelgren, 1969) and Wickelfeatures (Rumelhart & McClelland, 1988) are examples of such a representation. Model II also makes use of the representation, and that is the reason why its performance is much superior than that of model I. The top SOM of model II sees bits and over all, context in its input. As opposed to Wickelphones and Wickelfeatures, the representations in the input layer of the top SOM 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.

The size of context is the size of memory of past inputs, that means, the maximum number of past input bits that the bottom SOM may recognize. The size of context is directly dependent of the decay rate in the bottom SOM. If the decay rate is very large, the size of the map ought to be increased to recognize properly the large number of different contexts in the representations in the input layer. However, in most cases, a large memory of the past inputs is not necessary. If the decay rate is very small, the existence of the bottom SOM is unnecessary because it merely maps

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 input layer of the bottom SOM. A corresponding sequence $\mathbf{S_a^*}$ of winning units will be activated in the map. A second sequence $\mathbf{S_b}$ = 101111 which differs from the first only in the first bit is now presented. A corresponding sequence $\mathbf{S_b^*}$ of winning units is also activated in the map. By comparing the distances Φ (equation 5.5) between the winning units in $\mathbf{S_a^*}$ and $\mathbf{S_b^*}$, we may trace the memory size for past inputs of the SOM.

Figure 5.4 displays the distances when using different decay rates in the SOM. Using a decay rate of 0.6, we verify that after entering with the third bit, the SOM is still capable of distinguishing the difference in the first bit between $\mathbf{S_a}$ and $\mathbf{S_b}$. The size of the memory is then three bits. The memory size is two bits for decay rates of 0.4 and 0.5. We have presented sequences of up to three bits to the bottom SOM with decay rate of 0.6 to confirm that the SOM is able to classify them separately. The map is shown in figure 5.5.

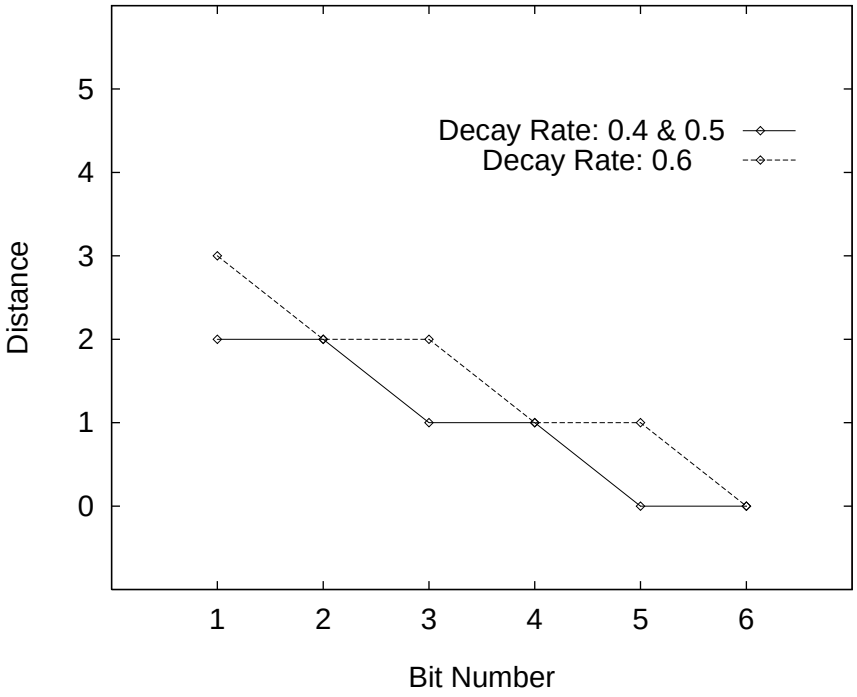


Figure 5.4. Distances between the winning units of two binary sequences

The decay rate of the bottom SOM also affects the outcomes of the top SOM, for the input of the later depends upon the classifications made by the former. We might verify this effect in the experiment. Despite the identical number of misclassifications (table 5.6), the misclassified sequence changes when varying the decay rate of the bottom SOM. Reducing the decay rate makes the memory size shorter. So, by reducing the decay rate, we would expect that the differences between the misclassified and the referential sequences would move more and more to the first bits. The experiment has indeed shown this behaviour. For decay rates of 0.4, 0.5, and 0.6, the misclassified sequences were 101011, 001001, and 001010 respectively.

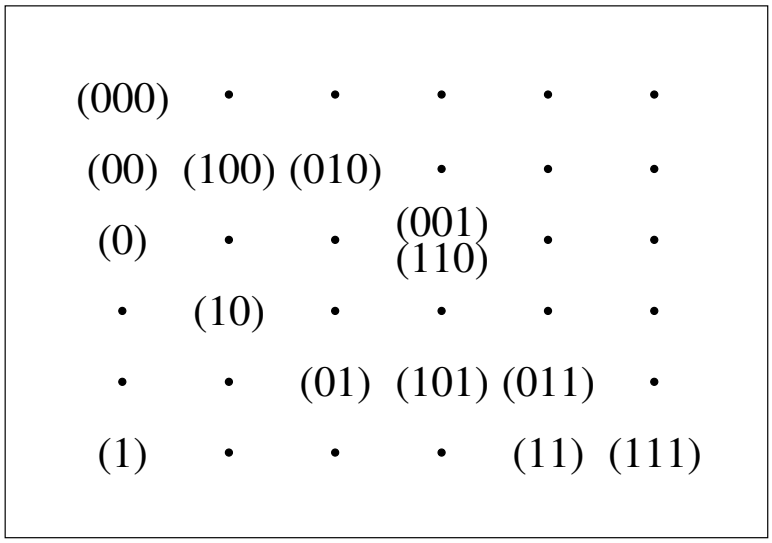


Figure 5.5. The map for three-bit binary sequences

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*.

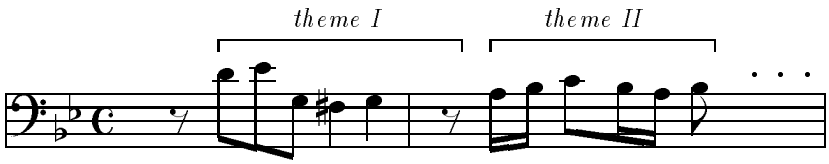


Figure 5.6. Theme of the sixteenth fugue in G minor

The fugue in G minor was chosen for several reasons. First, as many of fugues of Bach, it has four voices. Second, it possesses several perfect and modified instances of themes I and II. 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 twenty eighth and thirtieth bars, in which three instances of theme overlap. Fourth, the thematic material of theme II is extensively developed throughout the fugue. Such developments[2], although quite similar to theme II, are not instances of theme II. Fifth, very common intervals, as seconds up and down, occur extensively in the theme as well as in many passages in the fugue.

All facts above are usually present in real situations, in which humans are asked to perform thematic recognition in a polyphonic domain. Apart from providing a typical situation in a real domain, such facts also increase very much the level of difficulty of the domain to which the artificial neural model is applied.

The experiment pursued two aims. First, to determine whether models I and II recognize all instances of theme I and II in the third voice of the fugue. Second, to determine whether any other sequence (or sub-sequence), which was not an instance, was not misclassified as theme I or II.

The training of the two SOMs of model II and the SOM of model I was identical to that of the first experiment. They were trained in two phases — coarse-mapping and fine-tuning — with the same initial and final learning rates and sizes of the neighbourhood used in the first experiment. Again, in the coarse-mapping phase, the learning rate and the radius of the neighbourhood were reduced linearly whereas in the fine-tuning phase, they were kept constant. The coarse-mapping

[2]In the fugal domain, the developments of thematic material take place in parts called *episodes*.

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.

As in the first experiment (section 5.6), we have tested different values for the decay rate. In the bottom SOM of model II, they varied from 0.4 to 0.6, and in the top SOM, from 0.7 to 0.9. In model I, the decay rate ranged from 0.7 to 0.9. We present here only the results using decay rates of 0.5 and 0.85 respectively for the bottom and top SOM of model II, and 0.85 for the SOM of model I. Such results, reached with those decay rates, were the best for both input sets I and II.

The SOM of model I was tested with map size of 18×18, and was trained in 850 epochs. In model II, the map sizes were set to 15×15 (trained in 700 epochs) and 18×18 (trained in 850 epochs) to the bottom and top SOMs respectively. The transfer function Λ was a gaussian type function, given by equation 5.6, with κ and ρ set to 10 and 0.05.

The input layer of the model I and of the bottom SOM of model II held fifteen units, one for each musical interval ranging from an octave down to an octave up. The representation employed in these units is fully described in section 5.3.

A sequence (or sub-sequence) $\mathbf{S_a}$ is said to have the same classification as that of the theme $\mathbf{S_t}$ if the distance $\Phi(i_a^*, i_t^*) < 2$, where i_a^* and i_t^* are the (last) winning units of $\mathbf{S_a}$ and $\mathbf{S_t}$. In fact, if $\mathbf{S_a}$ is also an instance of the theme, when $\mathbf{S_a}$ and $\mathbf{S_t}$ have the same classification, not only i_a^* and i_t^* are adjacent but the winning units of $\mathbf{S_a}$ converge TIC by TIC to the winning units of the theme $\mathbf{S_t}$. The error of the instance $\mathbf{S_a}$ is given then by calculating the sum of the distances between each winning unit of $\mathbf{S_a}$ and its corresponding in $\mathbf{S_t}$. The mean error is given by the sum of the errors of each instance divided by the number of instances.

Figures 5.7 to 5.10 plot, for each instance of theme I, the distances between each winning unit of the instance and its corresponding in theme I for models I and II. Similarly, figures 5.11 to 5.16 plot, for each instance of theme II, the distances by applying the same method to theme II. Tables 5.8 and 5.9 display respectively the classifications and misclassifications of both models.

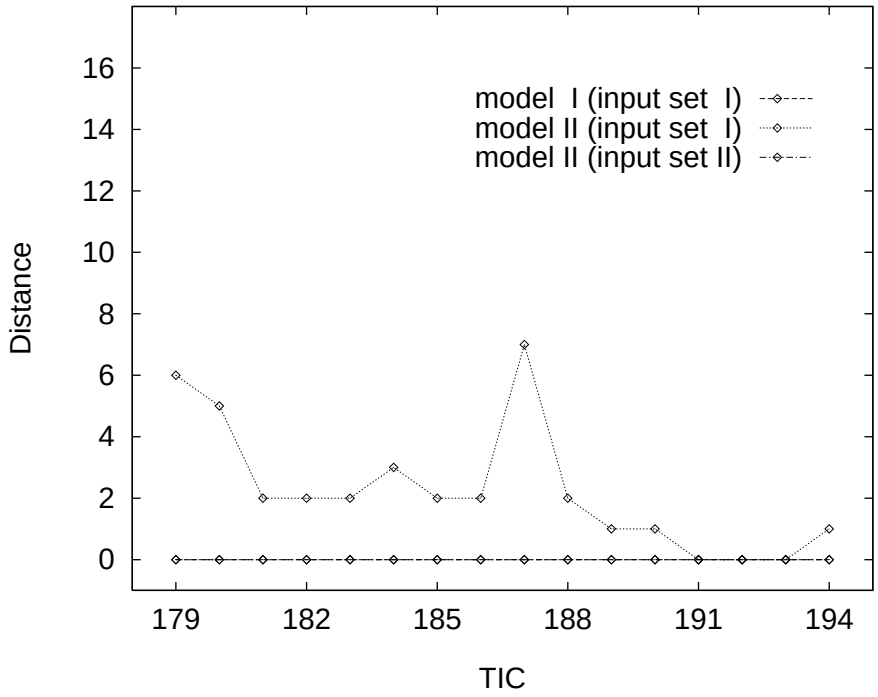


Figure 5.7. Classifications of the first instance of theme I (TICs 179 – 194) relative to theme I. The instance is a perfect copy of the theme I. The third voice remained inactive for more than five bars. Therefore, the context preceding the instance is insignificant when using input set I, and non-existent when using input set II. As a result, the activation of the input units immediately before the entrance of the instance is very low when using input set I, and non-existent when using input set II. The classifications of the instance yielded by model I and II converge to that of the theme I.

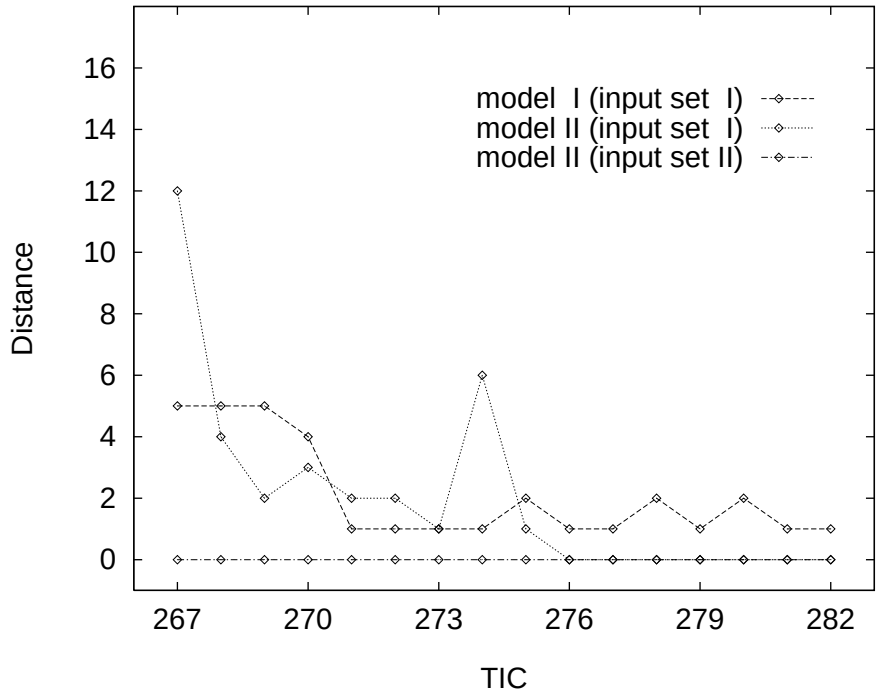


Figure 5.8. Classifications of the second instance of theme I (TICs 267 – 282) relative to theme I. The instance is a perfect copy of the theme I. There is a rest before the entrance of the second instance, and therefore, the activation of the input units immediately before the entrance of the instance 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.

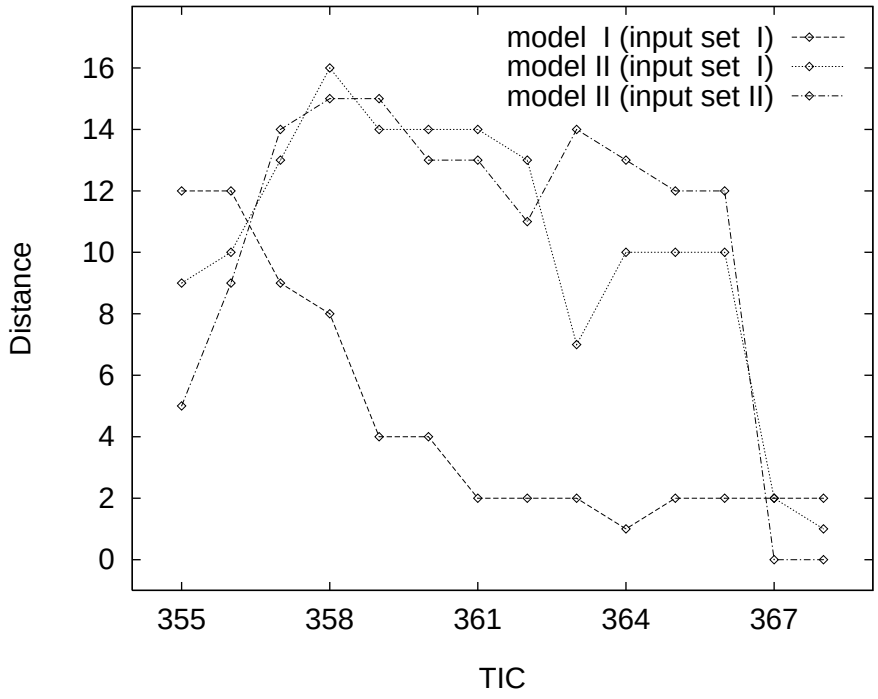


Figure 5.9. Classifications of the third instance of theme I (TICs 355 – 368) relative to theme I. The instance differs from the theme I in its first two TICs. There are no rests before the entrance of the instance, and as a consequence, the level of activation of the input units immediately before the entrance of the instance is high when using input sets I or II. The classifications of the instance yielded by model II converge to that of the theme I. On the contrary, the classification produced by model I does not.

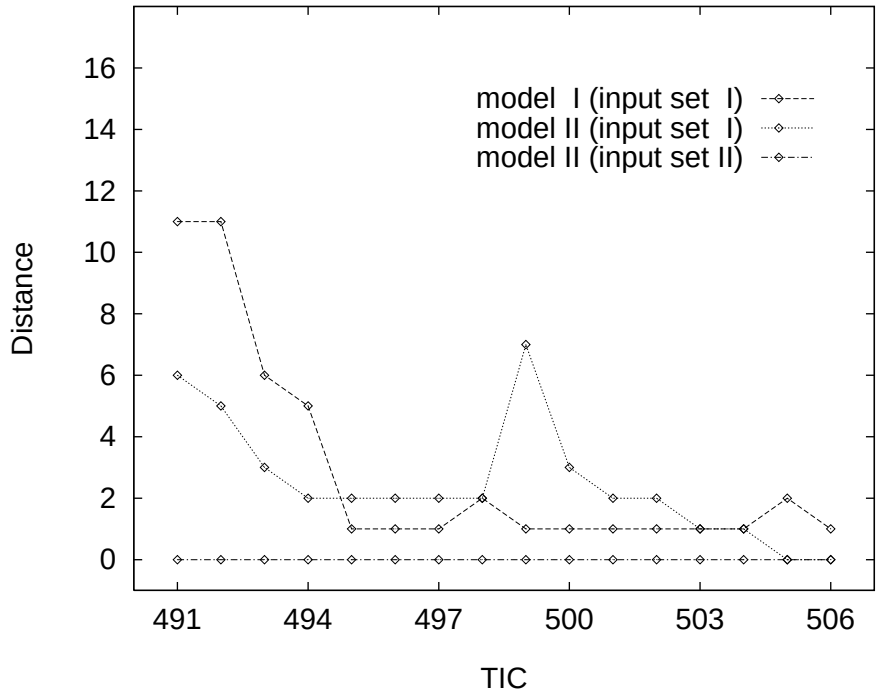


Figure 5.10. Classifications of the fourth instance of theme I (TICs 491 – 506) relative to theme I. The instance is a perfect copy of the theme I. There is a rest before the entrance of the fourth instance, and therefore, the activation of the input units immediately before the entrance of the instance 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.

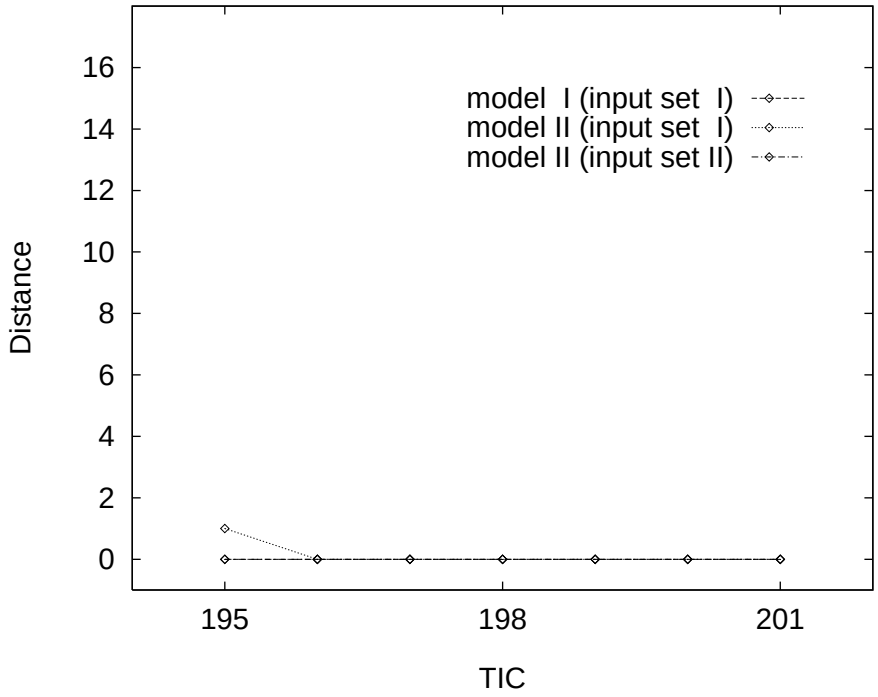


Figure 5.11. Classifications of the first instance of theme II (TICs 195 – 201) relative to theme II. The instance is a perfect copy of the theme II. The first instance of theme II is preceded by a perfect instance of theme I. Moreover, theme I ends with a rest. Two cases exist, therefore. When input 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 turn, when input set II is employed, there is no context preceding the instance and theme II, and thus, the activation of the input units immediately before the entrance of the instance is identical to that immediately before the entrance of theme II. The classifications of the instance yielded by model I and II converge to that of the theme II.

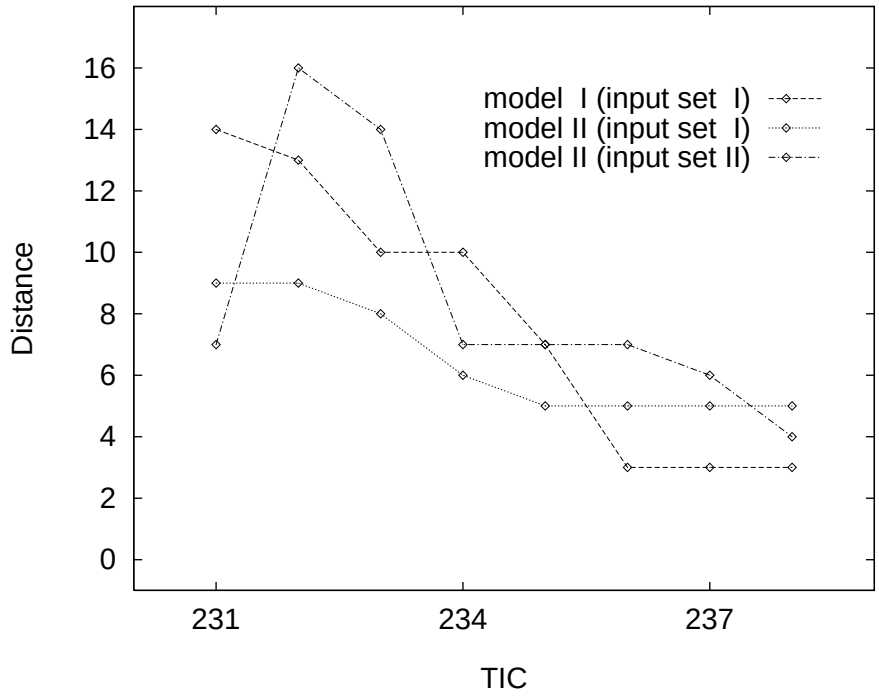


Figure 5.12. Classifications of the second instance of theme II (TICs 231 – 238) relative to theme II. The instance is a perfect copy of the theme II. The instance of theme II is not preceded by an instance of theme I, and as a result, the activation of the input units immediately before the entrance of the instance is totally different from 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.

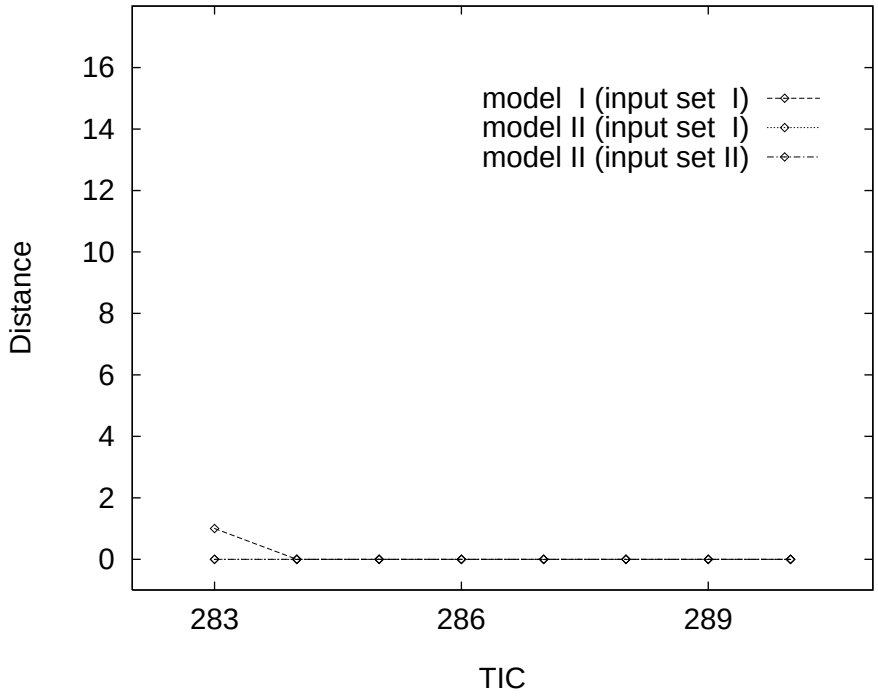


Figure 5.13. Classifications of the third instance of theme II (TICs 283 – 290) relative to theme II. The instance is a perfect copy of the theme II. The third instance of theme II is preceded by a perfect instance of theme I. Moreover, theme I ends with a rest. Two cases exist, therefore. When input 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 turn, when input set II is employed, there is no context preceding the instance and theme II, and thus, the activation of the input units immediately before the entrance of the instance is identical to that immediately before the entrance of theme II. The classifications of the instance yielded by model I and II converge to that of the theme II.

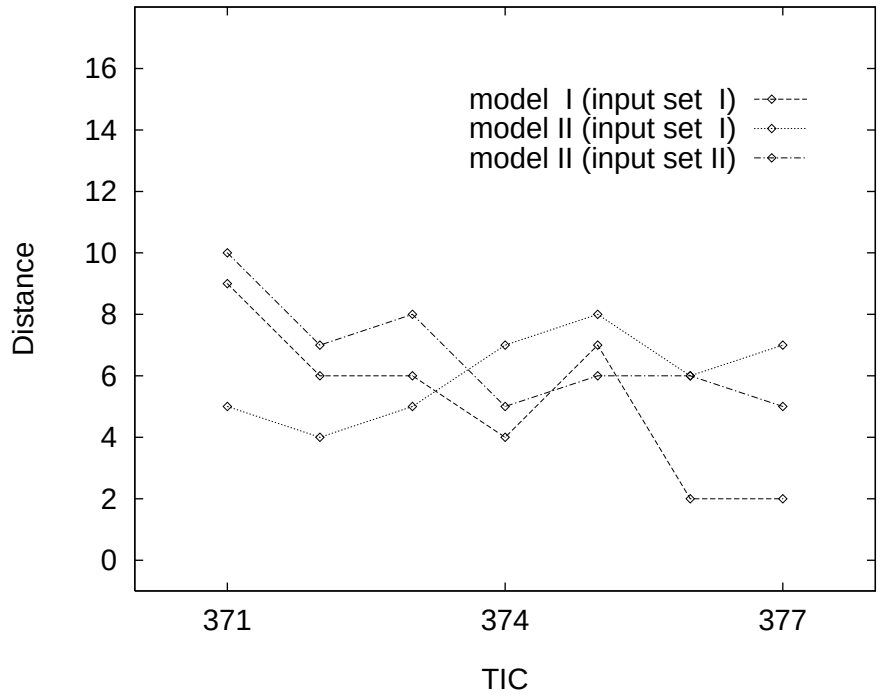


Figure 5.14. Classifications of the fourth instance of theme II (TICs 371 – 377) relative to theme II. The instance is a perfect copy of the theme II. The instance of theme II is preceded by an instance of theme I, which occurs altered in its first and last two TICs. Consequently, the activation of the 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.

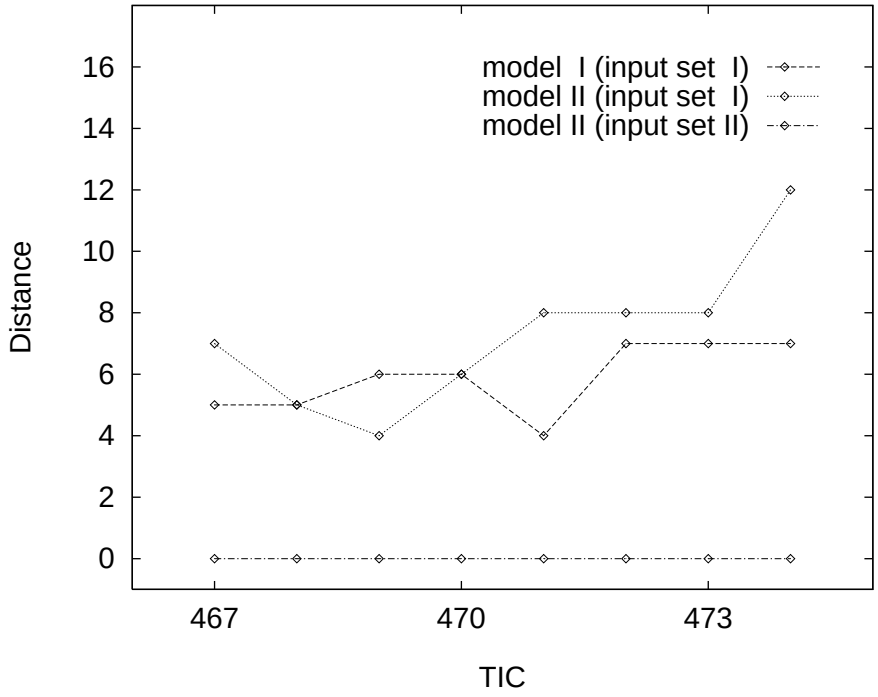


Figure 5.15. Classifications of the fifth instance of theme II (TICs 467 – 474) relative to theme II. The instance is a perfect copy of the theme II. The instance of theme II is not preceded by an instance of theme I. However, it is preceded by a rest. As theme I ends also in a rest, the activation of the input units immediately before the entrance of the instance is identical to that immediately before the entrance of theme II when using input set II. Nevertheless, when using input set I, the activation immediately before the entrance of the instance is very different from that immediately before the entrance of theme II. The classification of the instance produced by model II on the input set II converge to that of the theme II. On the contrary, the classifications produced by model I and II on the input set I do not.

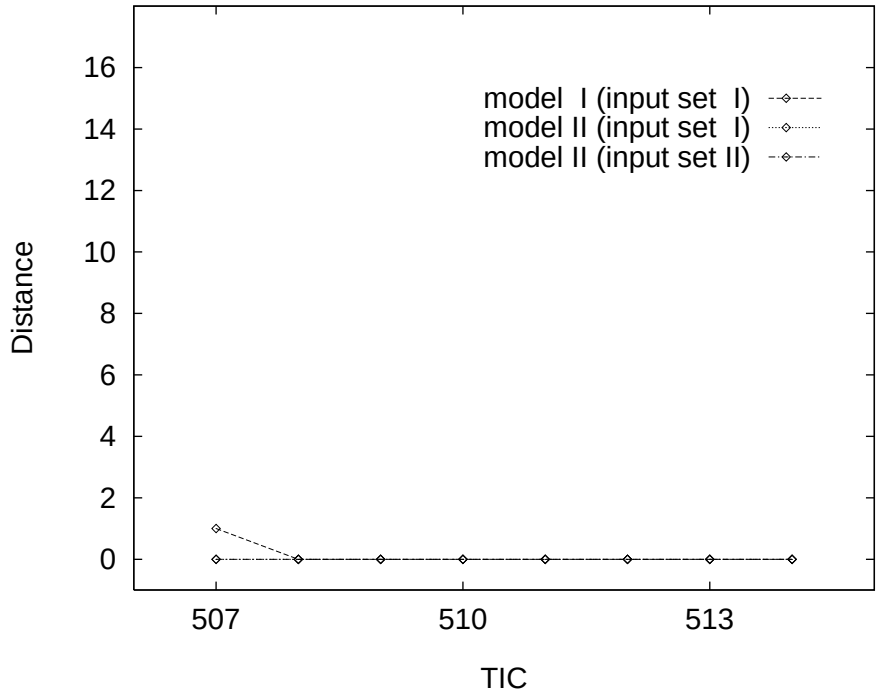


Figure 5.16. Classifications of the sixth instance of theme II (TICs 507 – 514) relative to theme II. The instance is a perfect copy of the theme II. The sixth instance of theme II is preceded by a perfect instance of theme I. Moreover, theme I ends with a rest. Two cases exist, therefore. When input 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 turn, when input set II is employed, there is no context preceding the instance and theme II, and thus, the activation of the input units immediately before the entrance of the instance is identical to that immediately before the entrance of theme II. The classifications of the instance yielded by model I and II converge to that of the theme II.

Table 5.8. Classifications of model I and II (second experiment)

Theme	No. Instances	Model	Input Set	No. Failures	Mean Error
I	4	I	I	1	36.25
		II	I	0	63.00
			II	0	36.50
II	6	I	I	3	24.67
		II	I	3	25.50
			II	2	19.17

Table 5.9. Misclassifications of model I and II (second experiment)

Theme	Model	Input Set	No. Minor Miscl.	No. Major Miscl.
I	I	I	4	1
	II	I	0	0
		II	0	0
II	I	I	3	4
	II	I	2	0
		II	1	0

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 any sequence (or sub-sequence) which have either some intervals of the theme in any order or rhythm, or all the intervals of the theme but in a different order or rhythm is not an instance of the theme, and so, must not be classified as such. The number of occurrences of this kind of sequences (or sub-sequences) in the fugue is high, and so is the probability that any neural model has of making misclassifications. Model II, however, considering input set I, had only two cases of minor misclassification[3]. Model I suffered from loss of context, and seriously misclassified five sub-sequences which contained intervals which were also present in the theme.

When performing on input set II, model II reaches even better results, whether they be in a fewer number of failures in classification, or in a lesser value of mean error, or in a fewer number of misclassifications. The better results may be explained by the fact that, as opposed to set I, input set II contains many sequences. Thus, model II discards previous context when starting inputting the following sequence, and consequently, the overall level of unit activations is reduced as well. The results reached by the model on this set suggest that segmentation does facilitate thematic recognition when performed on a univoiced musical domain.

[3]We consider a case of minor misclassification when the model keeps on classifying as theme the next few TICs which follow the theme.

5.8 Third experiment

The third experiment was on thematic recognition on a polyphonic musical sequence. As introduced in the second chapter (section 2.4.5.2), a polyphonic musical sequence is a sequence which contains multiple voices. The voices are sounded simultaneously, and share the same importance.

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 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 fugue. Segmentation was performed after the position adopted by Kirkpatrick (see section 2.4.6), that is, segmentation by rests was applied to each voice independently.

Polyphonic domains set much higher demands on neural models than univoiced musical domains. For this reason, and owing to the poor results of model I in sections 5.6 and 5.7, we decided to study the performance of model II only in this section. Thus, when set I was used, model II was trained and assessed on the recognition of sub-sequences within the fugue. Otherwise, when set II was used, model II was trained and assessed on the recognition of sequences produced by segmentation.

As mentioned in section 5.7, the fugue in G minor has 544 TICs, and TI is a sixteenth note. The *theme* of the fugue, which is shown in figure 5.6, was considered here in its entirety, no longer divided into two parts as it was in section 5.7. For this reason, in input set II, segmentation was not performed at any TIC in which instances of the theme occurred, so that any sequence in the set contained either an entire instance of theme or no part of it. The fugue in G minor was chosen as the polyphonic domain of the experiment for the same reasons as those described in section 5.7.

The experiment pursued two aims. Firstly, to determine whether model II recognizes all instances of the theme in the fugue. Secondly, to determine whether any other sequence (or subsequence), which was not an instance, was not misclassified as theme.

The training of the two SOMs of model II was identical to that in previous experiments. They were trained in two phases — coarse-mapping and fine-tuning — with the same initial and final learning rates and sizes of the neighbourhood used in those experiments. Again, in the coarse-mapping phase, the learning rate and the radius of the neighbourhood were reduced linearly whilst in the fine-tuning phase, they were kept constant. 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 both SOMs.

Different values for decay rate were tested. In the bottom SOM of model II, it varied from 0.1 to 0.7, and in the top SOM, from 0.5 to 0.9. We present here, however, only the results using decay rates of 0.3 and 0.7 for the bottom and top SOM respectively. Such results, reached with those decay rates, were the best for all studies performed.

The bottom SOM of model II was tested with map size of 15×15, and was trained in 700 epochs. The top SOM was tested with map size of 18×18, and trained in 850 epochs. Two transfer functions Λ were tested. The first was given by equation 5.7, with neighbourhood $N^* = \{i \mid \Phi(i,i^*) < 2\}$, and $\kappa = 0.5$. The second was also given by equation 5.7, but with neighbourhood $N^* = \{i \mid \Phi(i,i^*) < 4\}$, and $\kappa = 0.25$. We report here studies using the second transfer function only, for it produced the better results.

The input layer of the bottom SOM of model II held fifteen units, one for each musical interval ranging from an octave down to an octave up. The representation employed in these units is fully described in section 5.4.

The experiment comprised five studies. In the last four, in order to study the role of thematic reinforcement in thematic recognition in polyphony, reinforcement in activation was given to input units when representing instances of theme. For example, in the second experiment, note onset and note sustained received activations of 0.1 and 0.07 respectively. When corresponding to instances of the theme, they received instead, activations of 0.5 and 0.35 respectively. Table 5.10 shows the

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 corresponding five intervals in $\mathbf{S_t}$. Thus, from the next TIC onwards, each interval in $\mathbf{S_a}$ which matches its corresponding in $\mathbf{S_t}$ receives reinforcement. Reinforcement is stopped if either, in a given TIC, the interval in $\mathbf{S_a}$ does not match its corresponding in $\mathbf{S_t}$ or the last TIC in $\mathbf{S_t}$ is reached. The reinforcement process may be reinitiated again, nevertheless, if after any seven contiguous TICs ahead, the intervals in $\mathbf{S_a}$ start matching their corresponding in $\mathbf{S_t}$.

The reason for waiting seven TICs before starting the reinforcement process is due to the behaviour of human listeners when performing thematic recognition. We assume that the cognitive mechanisms which detect instances of theme in polyphonic music do not start actuating immediately with the first interval of the instance which matches its corresponding in the theme, but rather, they start actuating after a certain amount of time has elapsed. In our studies therefore, we considered this amount of time as being correspondent to seven TICs.

As in section 5.7, a sequence (or sub-sequence) $\mathbf{S_a}$ is said to have the same classification as that of the theme $\mathbf{S_t}$ if the distance $\Phi(i_a^*, i_t^*) < 2$, where i_a^* and i_t^* are the (last) winning units of $\mathbf{S_a}$ and $\mathbf{S_t}$. In case of $\mathbf{S_a}$ be also an instance of the theme, the error of the instance $\mathbf{S_a}$ is then given by calculating the sum of the distances between each winning unit of $\mathbf{S_a}$ and its corresponding in $\mathbf{S_t}$[4]. The mean error is given by the sum of the errors of each instance divided by the number of instances.

Figures 5.17 to 5.32 plot, for each instance of theme, the distances between each winning unit of the instance and its corresponding in the theme for each study carried out on model II. Tables 5.11 and 5.12 display respectively the classifications and misclassifications of the studies. Figure 5.33 plots the mean error of classifications in accordance with reinforcement provided in the first four studies.

[4]The error and the distances of each instance are computed from the seventh common TIC between the instance and the theme.

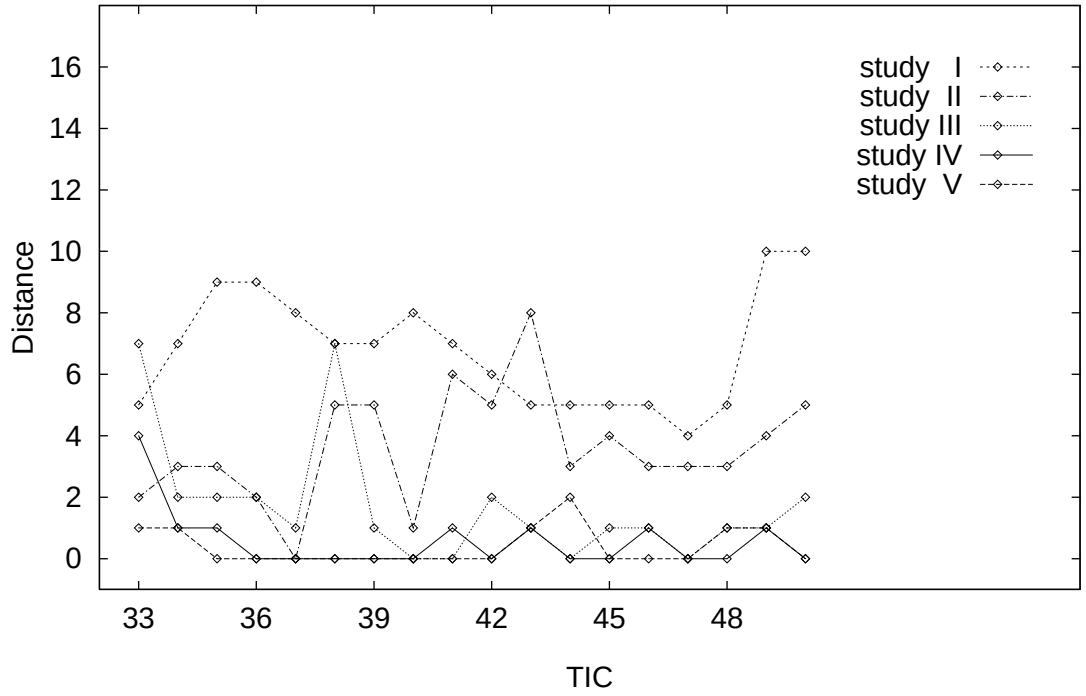


Figure 5.17. Classifications of the first instance of theme (TICs 33 – 50) relative to theme. The instance occurs in the fourth voice — the highest one — concurrently with another voice. The instance differs from the theme in its first two TICs. The classifications of the instance yielded by model II only converge to that of the theme in the fourth and fifth studies.

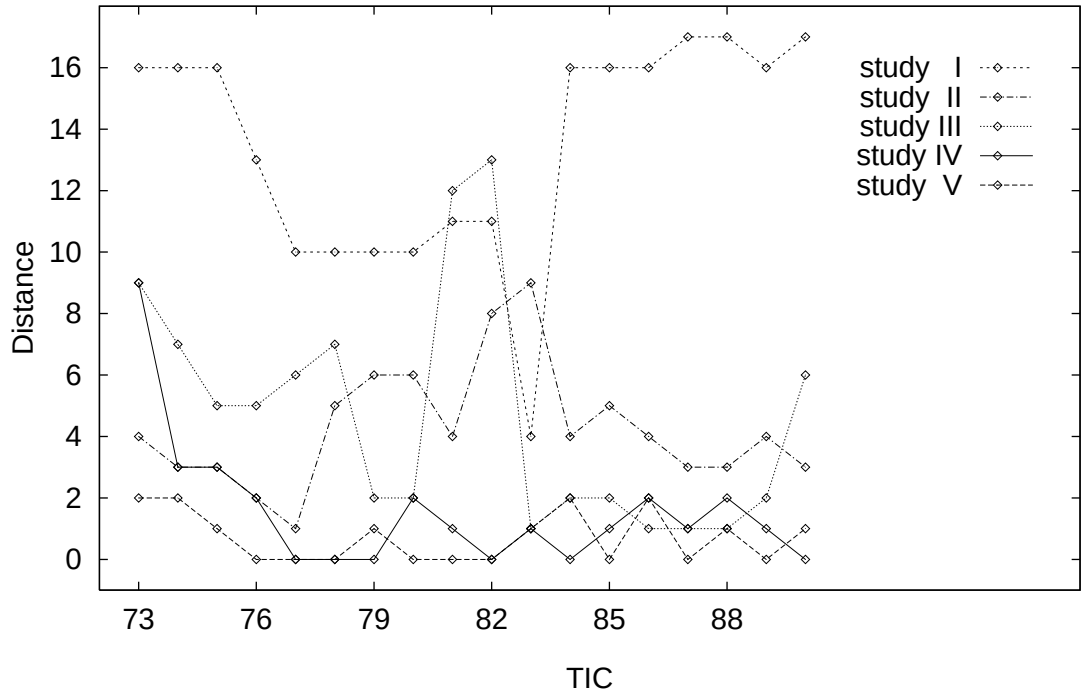


Figure 5.18. Classifications of the second instance of theme (TICs 73 – 90) relative to theme. The instance occurs in the first voice — the lowest one — concurrently with two other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II only converge to that of the theme in the fourth and fifth studies.

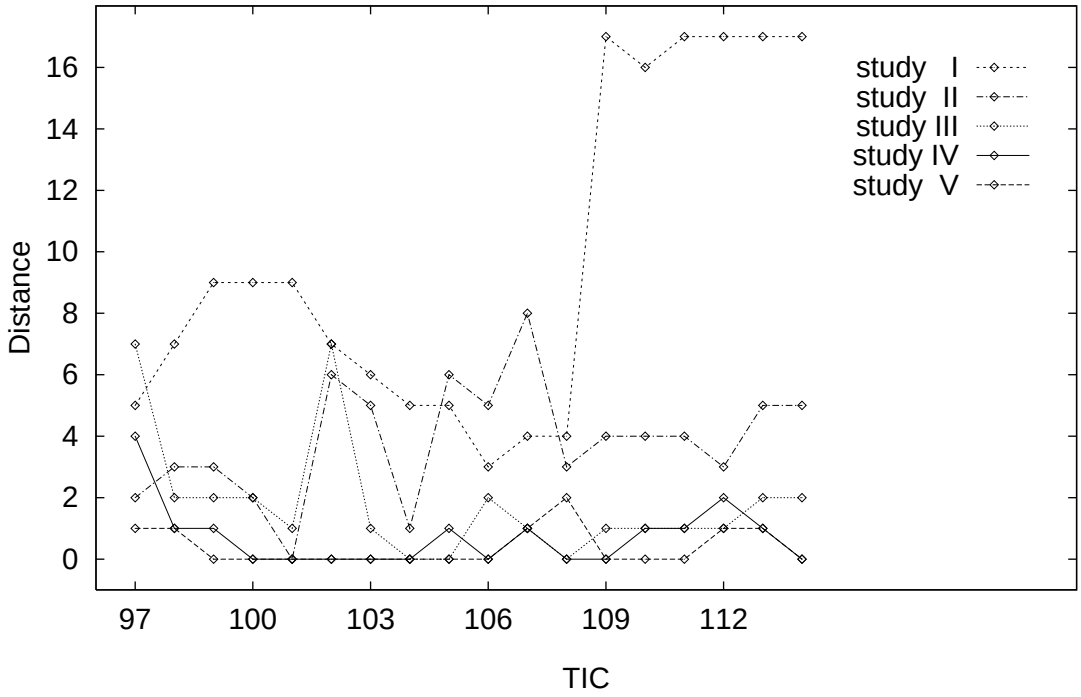


Figure 5.19. Classifications of the third instance of theme (TICs 97 – 114) relative to theme. The instance occurs in the second voice — the lower-middle one — concurrently with two other voices. The instance differs from the theme in its first two TICs. The classifications of the instance yielded by model II only converge to that of the theme in the fourth and fifth studies.

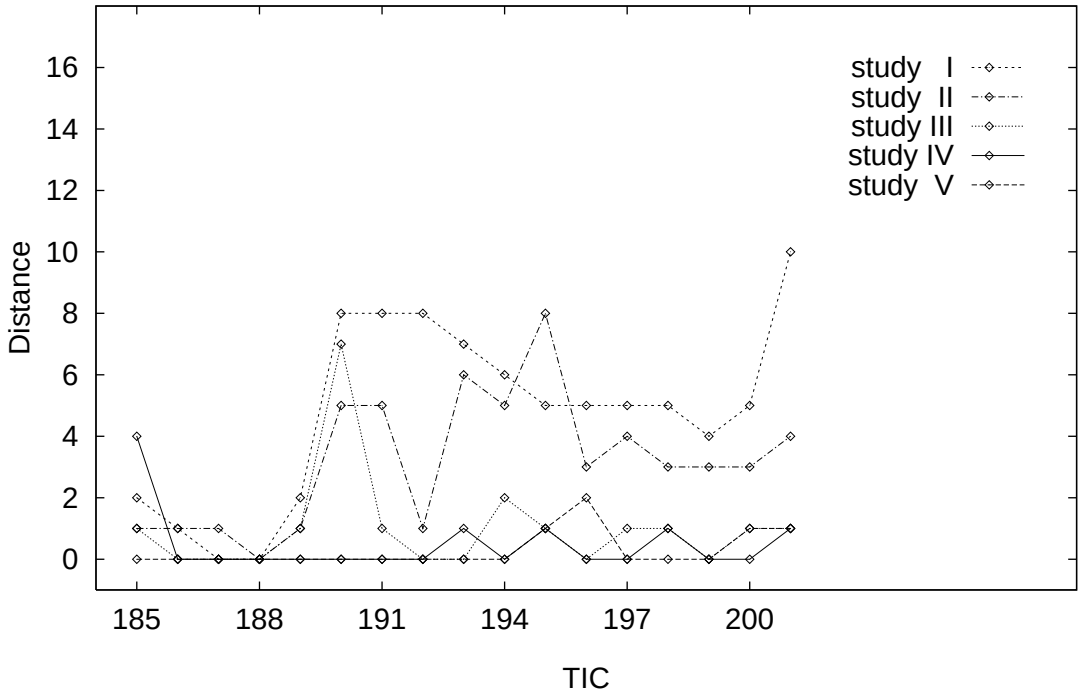


Figure 5.20. Classifications of the fourth instance of theme (TICs 185 – 201) relative to theme. The instance occurs in the third voice — the higher-middle one — concurrently with another voice. The instance is a perfect copy of the theme. The classifications of the instance produced by model II only converge to that of the theme in the third, fourth, and fifth studies.

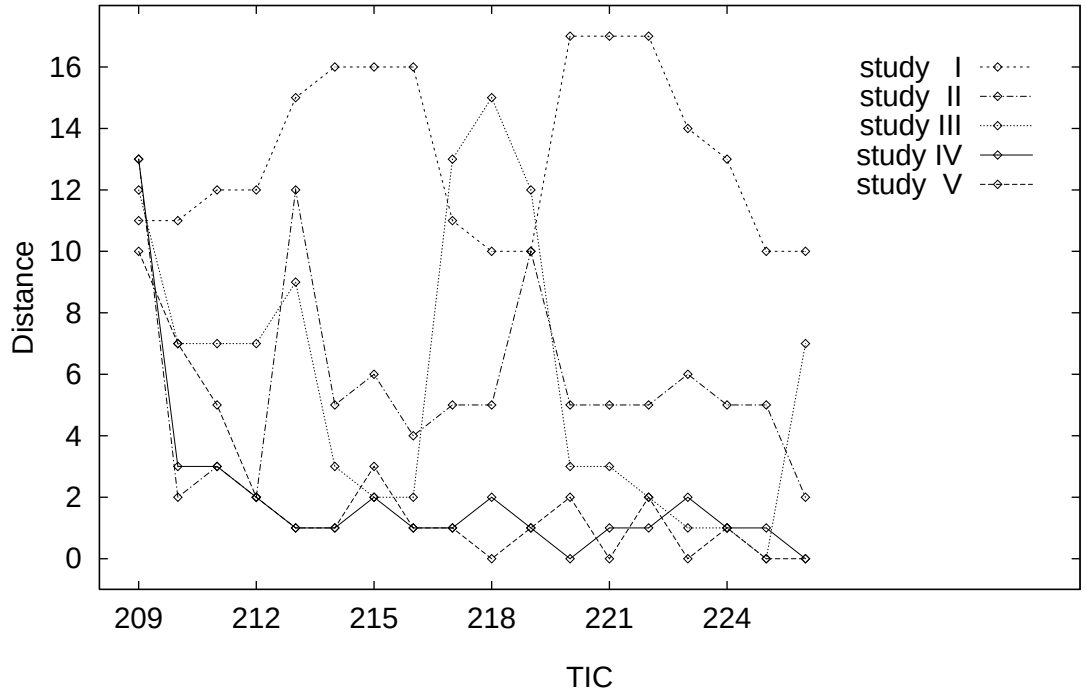


Figure 5.21. Classifications of the fifth instance of theme (TICs 209 – 226) relative to theme. The instance occurs in the first voice, concurrently with two other voices. The instance differs from the theme in its first two TICs. The classifications of the instance yielded by model II only converge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

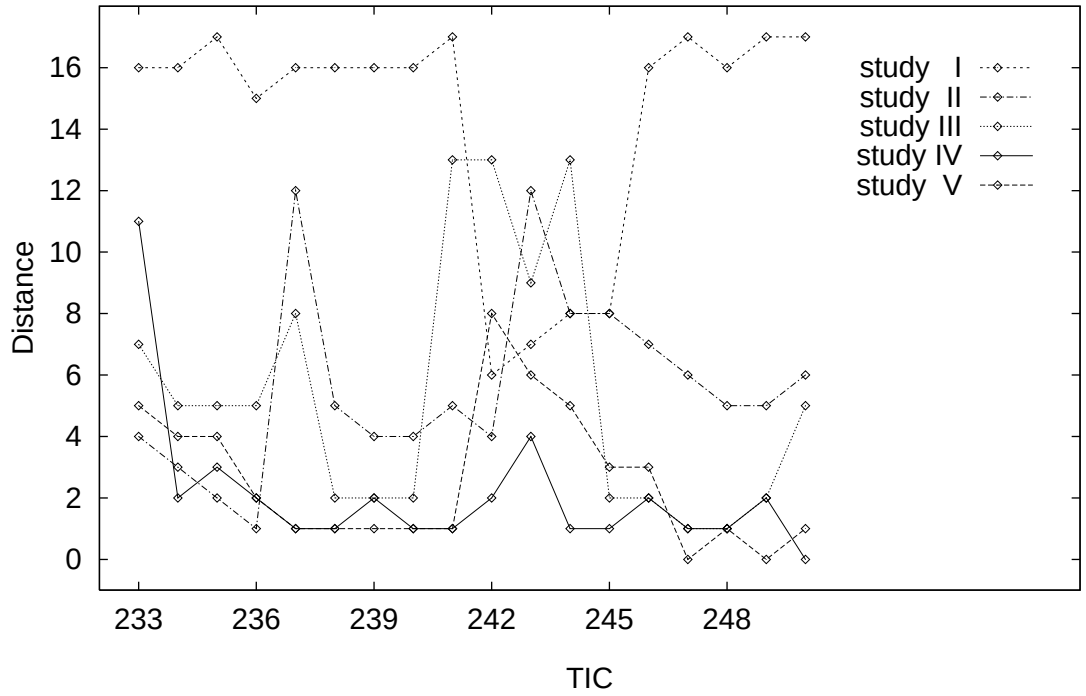


Figure 5.22. Classifications of the sixth instance of theme (TICs 233 – 250) relative to theme. The instance occurs in the fourth voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II only converge to that of the theme in the fourth and fifth studies.

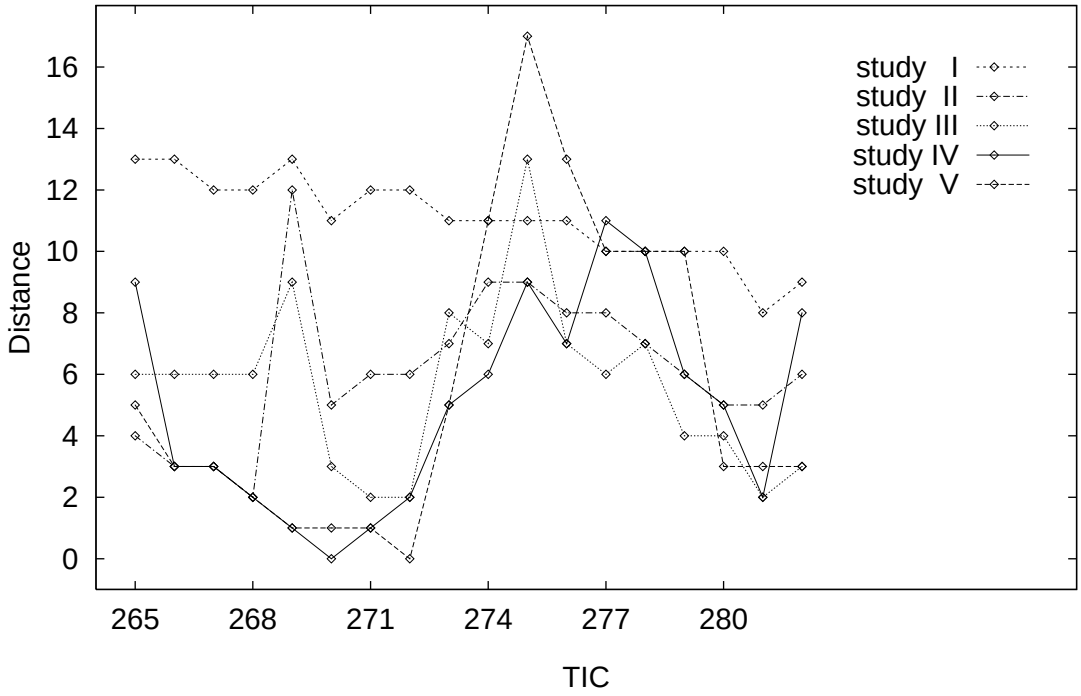


Figure 5.23. Classifications of the seventh instance of theme (TICs 265 – 282) relative to theme. The instance occurs in the first voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance yielded by model II do not converge to that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

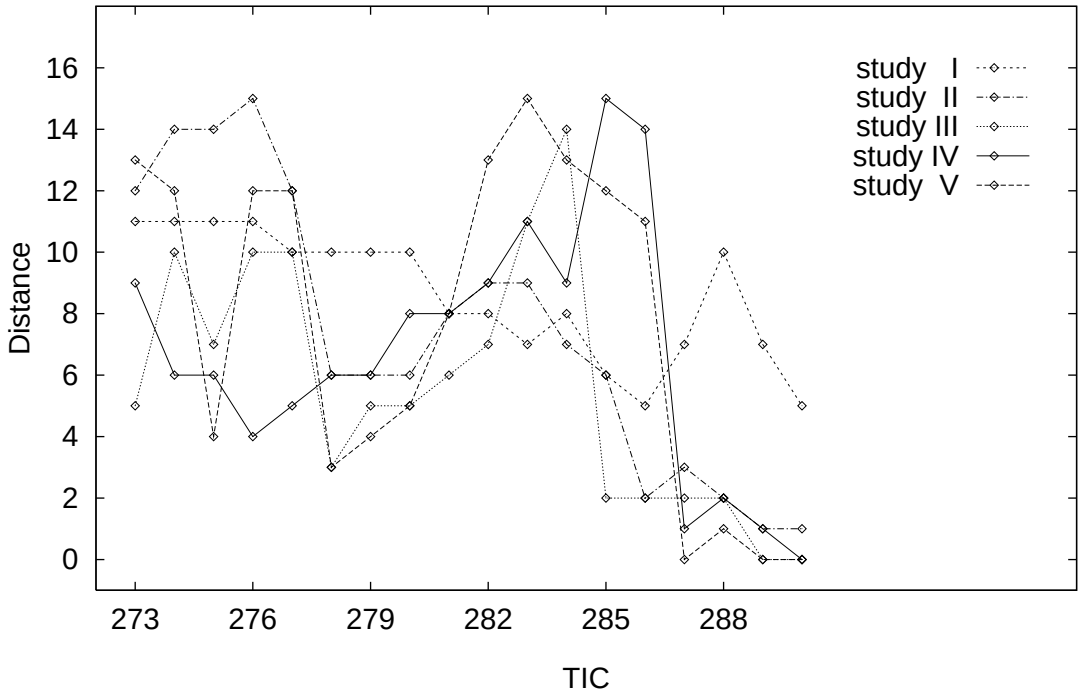


Figure 5.24. Classifications of the eighth instance of theme (TICs 273 – 290) relative to theme. The instance occurs in the third voice, and is a perfect copy of the theme. At its beginning, the eighth instance occurs concurrently with three other voices. From TIC 283 onwards, it occurs concurrently with just one voice. The classifications of the instance produced by model II only converge to that of the theme in the second, third, fourth, and fifth studies.

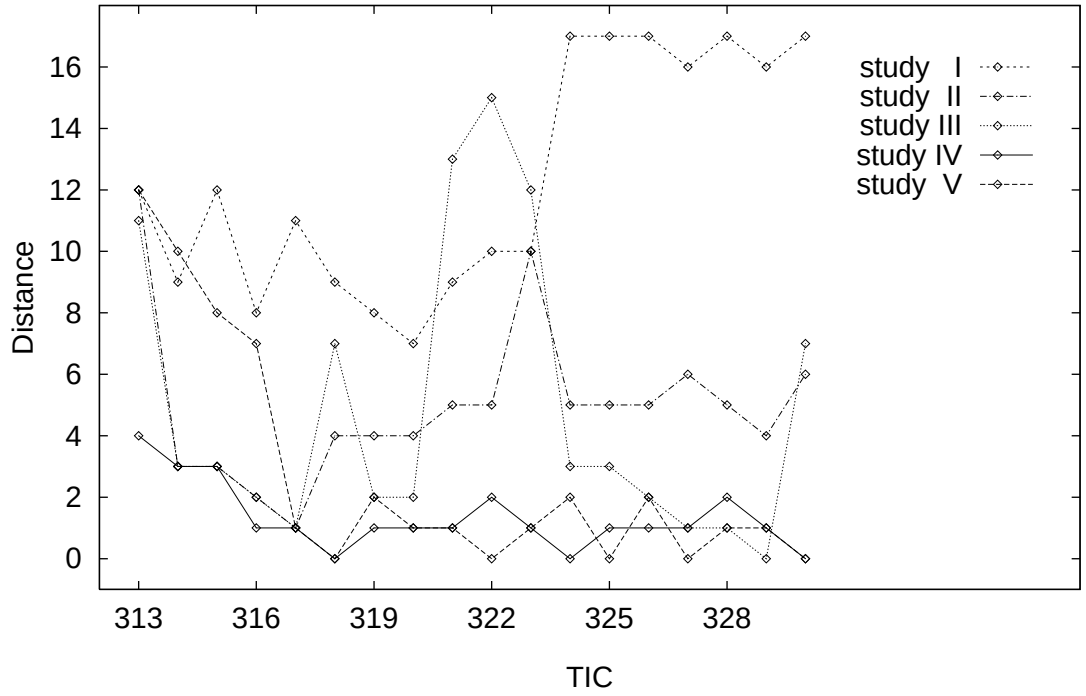


Figure 5.25. Classifications of the ninth instance of theme (TICs 313 – 330) relative to theme. The instance occurs in the first voice, concurrently with two other voices. The instance is a perfect copy of the theme. The classifications of the instance yielded by model II only converge to that of the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

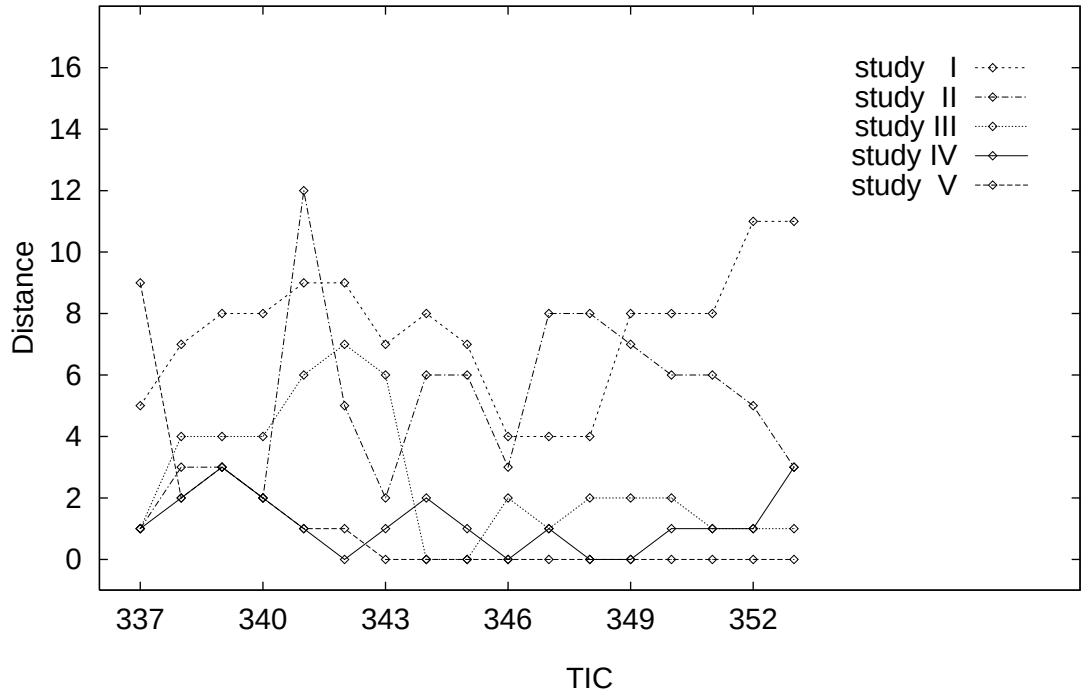


Figure 5.26. Classifications of the tenth instance of theme (TICs 337 – 353) relative to theme. The instance occurs in the fourth voice, concurrently with two other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II only converge to that of the theme in the third and fifth studies.

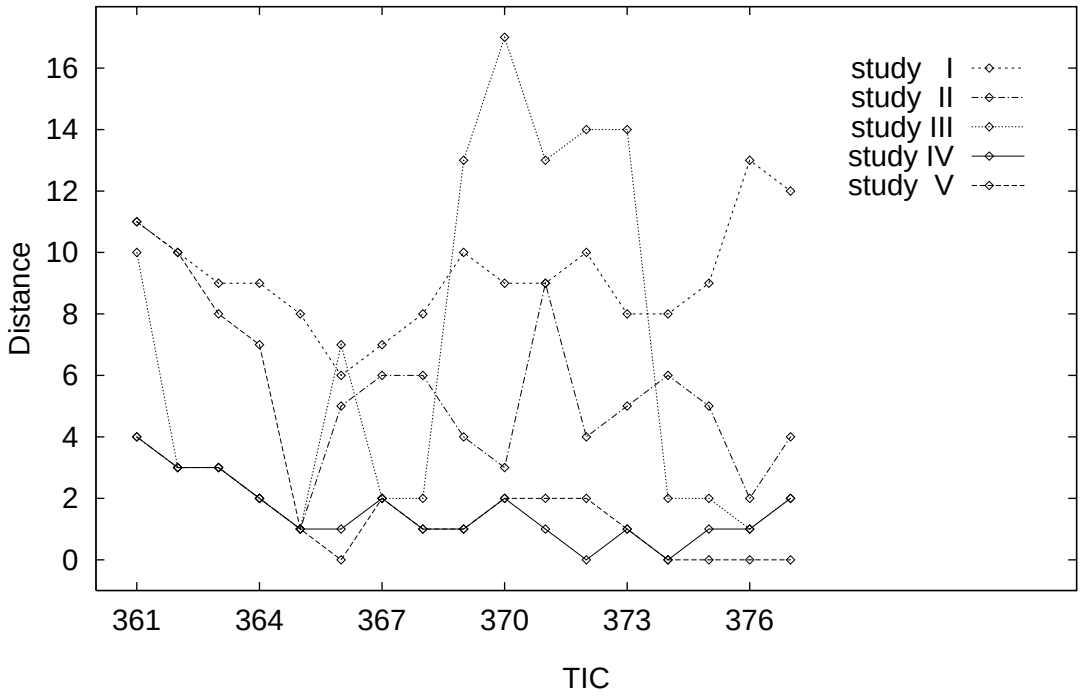


Figure 5.27. Classifications of the eleventh instance of theme (TICs 361 – 377) relative to theme. The instance occurs in the third voice, concurrently with two other voices. The instance differs from the theme in its first two TICs, and in two TICs in its middle. The classifications of the instance yielded by model II only converge to that of the theme in the fifth study.

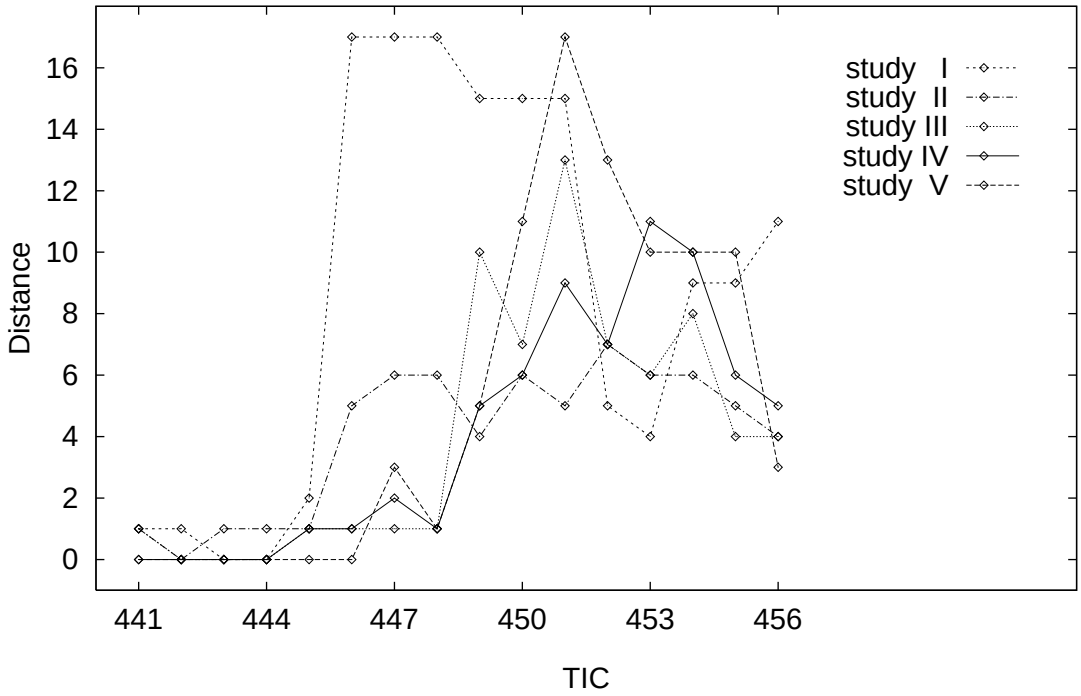


Figure 5.28. Classifications of the twelfth instance of theme (TICs 441 – 456) relative to theme. The instance occurs in the fourth voice, and is a perfect copy of the theme. At its beginning, the instance occurs unaccompanied. From TIC 443 onwards, it occurs concurrently with two other voices, and from TIC 451 onwards, it occurs concurrently with three other voices. The classifications of the instance produced by model II do not converge to that of the theme in any of the studies.

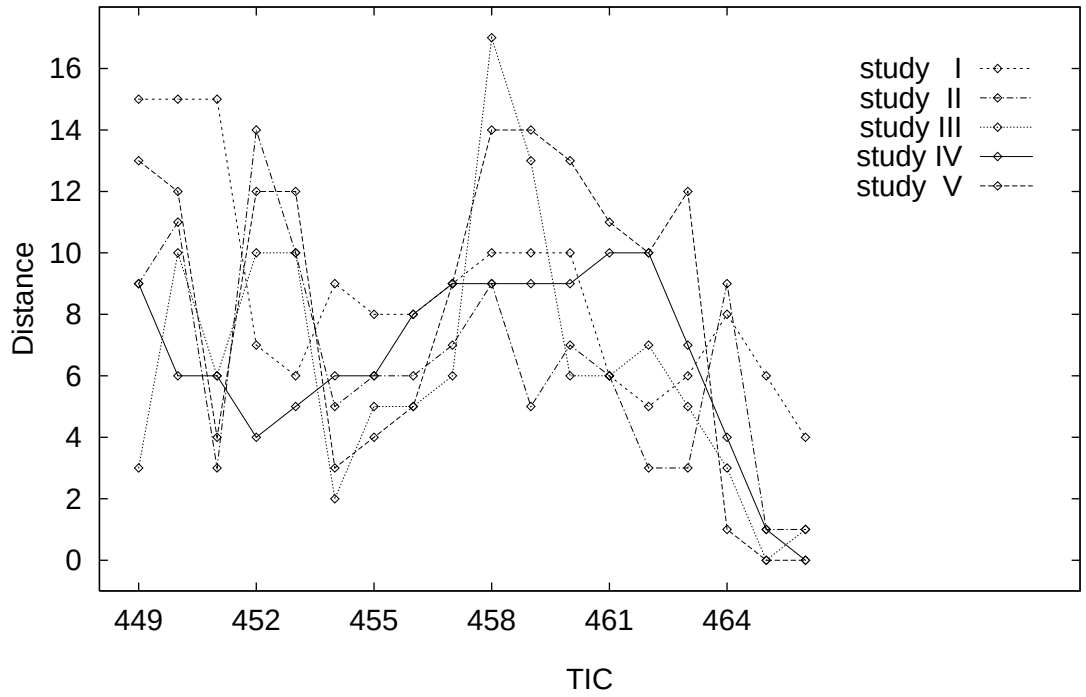


Figure 5.29. Classifications of the thirteenth instance of theme (TICs 449 – 466) relative to theme. The instance occurs in the second voice, and is a perfect copy of the theme. At its beginning, the instance occurs concurrently with two other voices. From TIC 451 onwards, it occurs concurrently with three other voices, and from TIC 459 onwards, it occurs concurrently with just one voice. The classifications of the instance yielded by model II only converge to that of the theme in the second, third, fourth, and fifth studies.

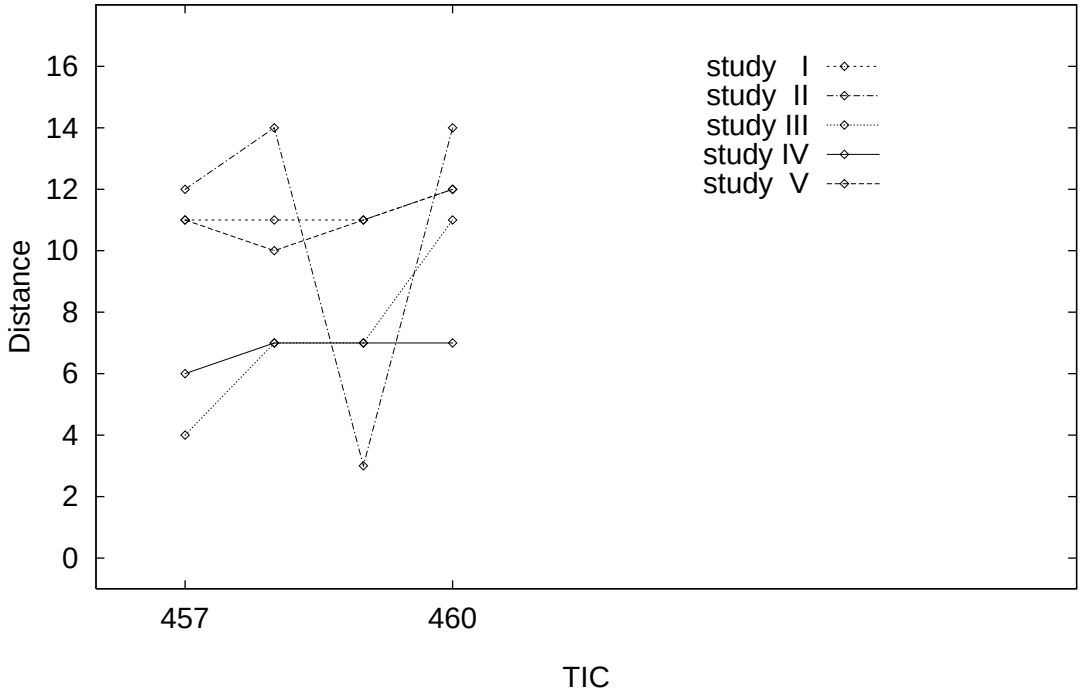


Figure 5.30. Classifications of the fourteenth instance of theme (TICs 457 – 460) relative to theme. The instance occurs in the first voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II do not converge to that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

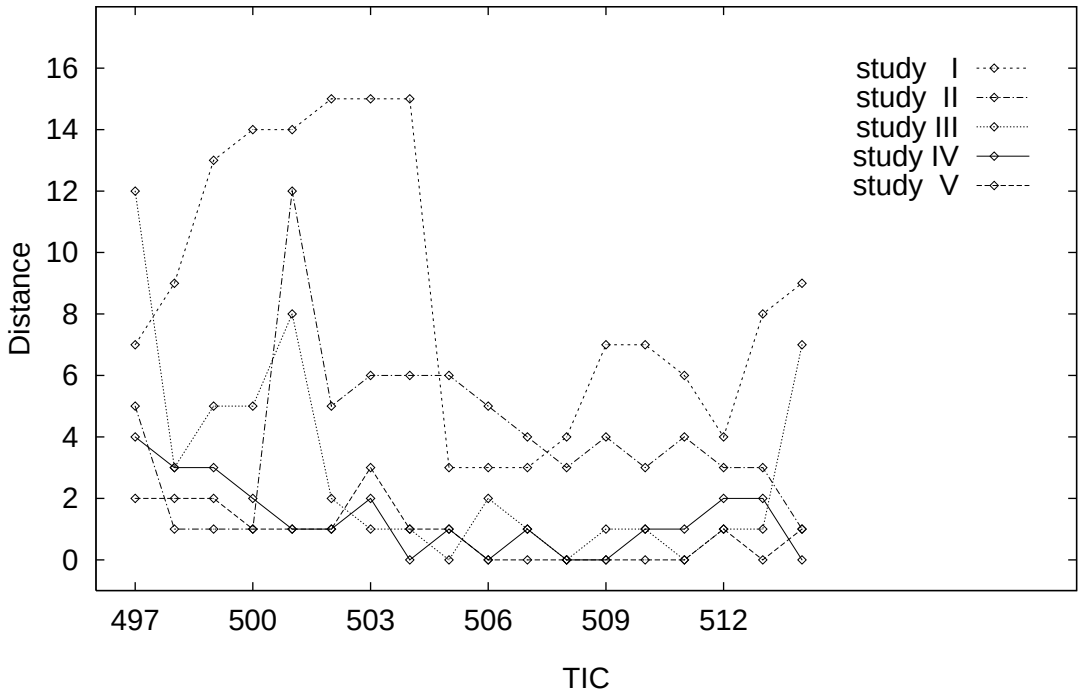


Figure 5.31. Classifications of the fifteenth instance of theme (TICs 497 – 514) relative to theme. The instance occurs in the third voice, concurrently with two other voices. The instance is a perfect copy of the theme. The classifications of the instance yielded by model II only converge to that of the theme in the second, fourth, and fifth studies.

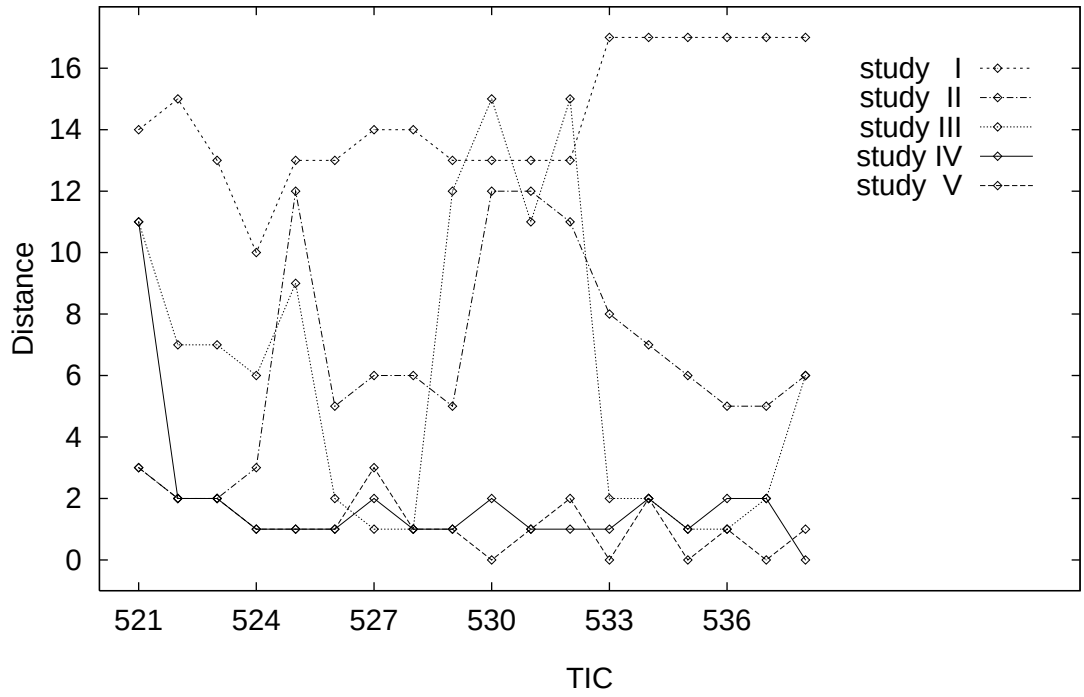


Figure 5.32. Classifications of the sixteenth instance of theme (TICs 521 – 538) relative to theme. The instance occurs in the second voice, concurrently with three other voices. The instance is a perfect copy of the theme. The classifications of the instance produced by model II only converge to that of the theme in the fourth and fifth studies.

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

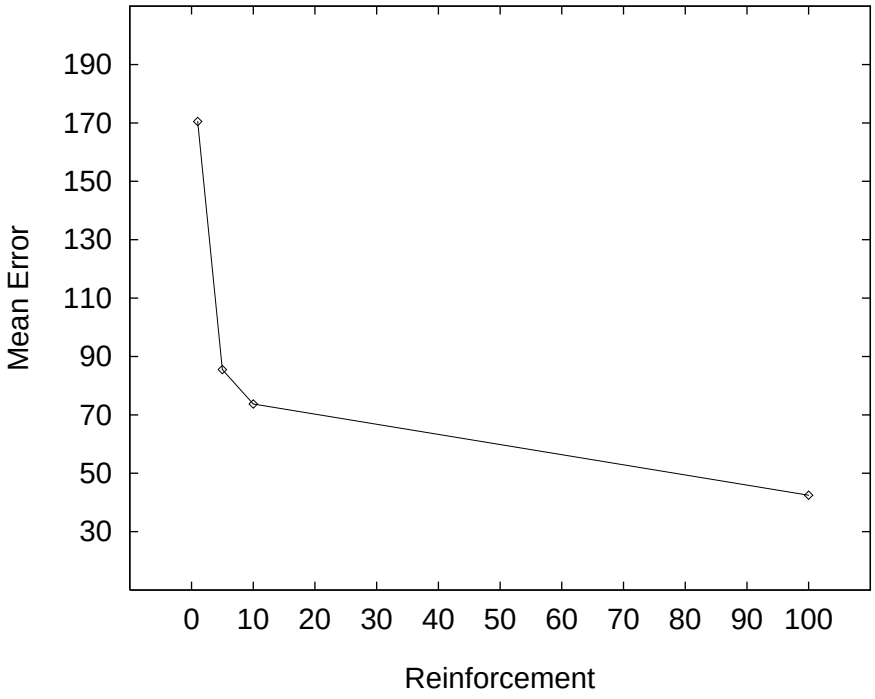


Figure 5.33. Mean error of classifications

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 polyphonic music. The connectionist model, which was introduced in the first chapter, comprises two stages — musical segmentation, described in the fourth chapter, and thematic recognition, described in the fifth chapter.

6.1.1 Musical segmentation stage

6.1.1.1 Background

It is believed that listeners perform segmentation in order to understand a musical piece. Owing to the limited capacity of their memories, listeners do not grasp a musical piece in its entirety, but rather, they segment it into parts. These parts constitute musical units that can be later analyzed and related to each other.

Three cases of rhythmic segmentation 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). Gestalt principles of proximity and similarity underlie these rules. These cases of segmentation were also acknowledged by Drake and Palmer (1993), and Kirkpatrick (1984), and were supported by experiments performed by Deliege (1987). In our experiments presented in the fourth chapter therefore, we assumed the validity of the three Lerdahl and Jackendoff's rules. We assumed that listeners do have the ability to recognize the cases of segmentation and perform segmentation according to the rules when listening to music.

A novel representation for rhythmic sequences was introduced. The concept behind the representation is the division of a musical piece into equal size time intervals, so that at each time interval, either there is a rest, or a note onset, or a note sustained. The representation is physically plausible since, in terms of rhythmic events, music is heard as sequences of rests, onsets, and sustained sounds. The representation is distributed in a pair of neural input units. A rest was represented by (00). Note sustained was represented by (10), and note onset by (11).

6.1.1.2 Model

The supervised model was trained to segment musical pieces in accordance with the three cases of rhythmic segmentation. It has a topology which is similar to that of Sejnowski and Rosenberg's (1987) model. The input layer holds a number of pairs of units which make up a window. Each pair represents one of the three events — rest, note sustained, and note onset. The activations of pairs of units in the window represent a rhythmic pattern.

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 segmentation given by rests. We extended the Lerdahl and Jackendoff's rule presented in the second chapter (section 2.3.4.4) to include early voice entrances in polyphonic music. The extension of the rule was acknowledged by Kirkpatrick (1984). It states that segmentation given by rests takes place whenever a rest is followed by at least two notes.

The second and third experiments were on recognizing cases of segmentation given by longer durations and breaks of similarity respectively. According to Lerdahl and Jackendoff's rules presented in the second chapter (sections 2.3.4.4 and 2.3.4.5), segmentation given by longer durations occurs whenever, in a group of four notes, the duration of the second note is longer than those of the first and third notes. In its turn, segmentation given by breaks of similarity occurs whenever, in a group of eight notes, the durations of the first four notes are identical, the durations of the last four notes are also identical, and the durations of the first four notes are different from those of the last four.

The model was evaluated on the same musical pieces in each experiment. It was evaluated on two two-part inventions, two three-part inventions, and two fugues of Bach. Each voice of each piece was input separately.

6.1.1.4 Results

The model has proved to be superior to the three recurrent supervised models for sequence classification in time reviewed in the third chapter. The cases of rhythmic segmentation demand that the model have long memory of past events. They also demand it establish the precise position where boundaries ought to occur. The windowed model employed successfully met both conditions. On the contrary, the recurrent supervised models do not fulfil the requirements demanded by rhythmic segmentation. Rumelhart, Hinton, and Williams's (1986b) model is computationally very expensive, and do not establish boundary positions. Mozer's (1989) model and Elman's (1990) model do not establish boundary positions, and do not hold a long memory of past events.

6.1.1.5 Conclusions

The results showed that the model could learn the grouping rules. The overall percentage of misclassifications was very low. However, it could be further reduced by increasing the number of patterns in the training sets. The results suggest that cognitive mechanisms which recognize the three cases of rhythmic segmentation and perform segmentation according to the three Lerdahl and Jackendoff's grouping rules can be modelled by a windowed artificial neural model with supervised learning.

6.1.2 Thematic recognition stage

6.1.2.1 Background

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.

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) concerning recognition of melodies in the presence of distractors — foreign notes interleaved with melodies — and Palmer and Holleran's (1994) studies (section 2.4.4) concerning recognition of alterations in melodies in multivoiced music, have provided a few clues for research in perception of themes in polyphonic music.

The studies suggest that recognition of a theme presented in a polyphonic voice is affected when it overlaps with other voices. Yet, active search for a theme can lead to discerning it even in midst of overlapping voices. The studies also suggest that listeners attend more readily to the voice presenting the theme, specially when, in comparison with other polyphonic voices, it occurs in the higher-frequency range. In the light of these studies therefore, for our experiments presented in the fifth chapter, we assumed that listeners do have the ability to detect thematic material in polyphonic music.

Despite of those studies on interleaved melodies and multivoiced music yet, as far as we know, there have not been experimental studies in the domain of real polyphonic music, and thus, important issues concerning thematic recognition in such domain are still open to investigation. Indeed, there are many more questions than answers relating to thematic recognition in the polyphonic domain.

First, it is not known how listeners are able to recognize instances of theme through the voices, and which cognitive mechanisms are involved in the understanding of polyphonic music. Second, it is not known which kind of and degree of variations are permitted in an instance of theme, so that its recognition remains unaffected. Third, it is not known whether or not, and to what extent, listeners rely on reinforcement in order to recognize properly instances of theme. Fourth, it is not known how listeners' performance in thematic recognition is affected by the number of voices which sound simultaneously with the theme. Fifth, it is not known how listeners' performance in thematic recognition is affected by the amount of onset synchrony. Sixth, although segmentation may be imperative to the complete musical understanding in the homophonic domain, it is not known whether or not it is necessary for thematic recognition in the polyphonic domain.

The fifth chapter was concerned, in particular, with answers to two of those questions. It focused mainly on investigating the role of segmentation and thematic reinforcement in thematic recognition in polyphonic music. In other words, it focused on the question of whether or not cognitive mechanisms of segmentation and thematic reinforcement facilitate thematic recognition in polyphonic music.

Novel representations for univoiced and for multivoiced musical sequences were introduced in the fifth chapter. They are based on interval representations, and are plausible because are supported by studies reviewed in the second chapter (section 2.2). The studies indicate that individuals hold internal representations for intervals and contour. However, the studies do not specify the type of those representations.

The novel representations for univoiced and for multivoiced musical sequences made use of the concept behind the representation for rhythmic sequences employed in the fourth chapter. Thus, a musical piece is divided into equal size time intervals, so that at each time interval, either there is a rest, or a note onset, or a note sustained.

Neural input units endowed with a time integrator were used in the representations for univoiced and for multivoiced musical sequences. The time integrator provides the input units with

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

twenty eighth and thirtieth bars, in which three instances of theme overlap. Fourth, the thematic material is extensively developed throughout the fugue. Such developments, although quite similar to the theme, are not instances of theme. Fifth, very common intervals, as seconds up and down, occur extensively in the theme as well as in many passages in the fugue.

All facts above are usually present in real situations, in which humans are asked to perform thematic recognition in a polyphonic domain. Apart from providing a typical situation in a real domain, such facts also increase very much the level of difficulty of the domain to which the artificial neural model is applied.

6.1.2.4 Results

Model II has proved to be superior to the three unsupervised models reviewed in the third chapter. As well as model II, they are extensions of the self-organizing map to sequence classification in time.

Thematic recognition sets three conditions on the model. First, it demands that the model be able to recognize both a set of input sequences and a set of sub-sequences within a large and unique input sequence. The model is required to recognize a set of input sequences when segmentation is performed on the musical piece. Otherwise, when segmentation is not performed, the entire piece consists of a unique input sequence, and the model is thus required to recognize sub-sequences of that sequence.

Second, it demands that the model classify sequences (or sub-sequences) properly in the presence of noise. The reason follows from the fact that any two sequences which differ slightly must achieve similar classifications.

Third, it demands that the model recognize sequences (or sub-sequences) in a very precise form. The reason for the latter is that any two sequences which share either some intervals, or even all intervals, but in an alternative order or rhythm, are musically different, and as a consequence, must be recognized as distinct.

Model II successfully met the three conditions. Conversely, the three unsupervised models reviewed do not fulfil the requirements demanded by thematic recognition. Kangas' (1994) model is computationally very expensive in consequence of the long window sizes required to deal with musical sequences. Chappell and Taylor's (1993) model suffers from loss of context, and thus, is unable to classify properly long musical sequences. James and Miikkulainen's (1995) model is unable to recognize sub-sequences within an unique input sequence, and consequently, can not be employed when segmentation is not performed on the musical piece.

The results showed that the performance of model I was worse than that of model II in cases of sequence classification. Model II performed thematic recognition successfully, and could classify properly all instances of theme, apart from three cases of *stretto*. The model was fault tolerant to errors, since 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.

The results also showed that the performance of model I was much worse than that of model II in cases of sequence misclassification. The fugue in G minor 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 any sequence which have either some intervals of the theme in any order or rhythm, or all the intervals of the theme but in a different order or rhythm is not an instance of the theme, and so, must not be classified as such. The number of occurrences of this kind of sequences in the fugue is high, and so is the probability that any neural model has of making misclassifications. Model II, nevertheless, had very few cases of minor misclassification. Model I suffered from loss of context, and seriously misclassified several sequences which contained

intervals which were also present in the theme.

6.1.2.5 Conclusions

The results indicate that the original unsupervised model can perform thematic recognition successfully. This suggests that the cognitive mechanisms which recognize instances of theme in polyphonic music can be modelled by an artificial hierarchical neural model with unsupervised learning. By assuming the artificial neural model as a plausible model for the field of music perception, these results then reveal the important role that segmentation and thematic reinforcement play in thematic recognition in polyphonic music, and lead us to relevant conclusions in music perception as well. 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.

6.2 Contributions of the research

The current research situates within the wide field of music perception. It proposes a connectionist model to investigate the mechanisms behind thematic recognition. Research in music perception aims at understanding the mechanisms involved in the perception of musical attributes, such as pitch, rhythm, themes, tonality, and form. We reviewed three works in the field in the third chapter. Laden and Keefe (1991) assessed representations of pitch, and supervised neural nets as models of pitch and chord perception. Leman (1991) performed a set of experiments in tonality by employing a self-organizing map model (Kohonen, 1989) for studying relations between tones in a tonal context. Gjerdingen (1990, 1991) used a self-organizing ART2 net (Carpenter & Grossberg, 1987) to classify musical patterns in six of Mozart's earliest compositions. The models were able to learn important human cognitive tasks, and consequently, the studies put forward a few contributions to the field of music perception. Nevertheless, it is worth noticing the fact that the studies were on simple cognitive tasks, which are hierarchically situated in cognitive levels lower than those of segmentation and thematic recognition.

The current research makes several contributions. First, are the representations for rhythmic sequences, and for univoiced and multivoiced musical sequences described in the fourth and fifth chapters respectively. The idea of employing the small figure to define time interval and time interval counter in a musical piece is original and physically plausible, and so are those representations, which make use of the idea. The representations for univoiced and multivoiced musical sequences are supported by studies on interval representation reviewed in the second chapter (section 2.2). Although the studies do not specify the type of that representation, they indicate that individuals do hold internal representations for intervals.

Second, is the supervised model described in the fourth chapter. A better understanding of a cognitive phenomenon can be achieved by means of employing an explicit model, which can be evaluated and explored. The results showed that the model could learn Lerdahl and Jackendoff's grouping rules. This suggests that cognitive mechanisms which recognize cases of rhythmic segmentation and perform segmentation according to Lerdahl and Jackendoff's grouping rules can be modelled by a windowed artificial neural model with supervised learning.

Third, is the original unsupervised model itself described in the fifth chapter. The hierarchical topology along with time integrators enable 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 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.

Fourth, is the original unsupervised model mentioned above employed as a model in music

perception. The results of the experiments showed that it could perform thematic recognition rightly. The results suggest that cognitive mechanisms which recognize instances of theme in polyphonic music can be modelled by an artificial hierarchical neural model with unsupervised learning. By assuming the artificial neural model as a plausible model for the field of music perception, these results then suggest the next four contributions in music perception.

Fifth, thematic recognition is particularly difficult when performed on passages containing *strettos*. The model performed poorly on such passages, and that might suggest that the listener's attention is disoriented by the presence of overlapping instances of theme. Experiments in cognitive psychology, nevertheless, should be carried out in order to confirm and determine to what extent passages containing *strettos* present an obstacle to listeners' performance on thematic recognition.

Sixth, mechanisms of segmentation seem to facilitate thematic recognition when carried out on an univoiced musical domain. Univoiced musical domains are quite similar to homophonic musical domains. Indeed, by excluding the harmonic accompaniment, a homophonic musical piece would contain the melodic line only, and thus, it could be considered as an univoiced musical piece. Although Lerdahl and Jackendoff's grouping theory is directed to homophonic music, the grouping rules are intended to be applied uniquely to the melodic line in order to define grouping boundaries. Thus, by considering thematic recognition as a part of musical understanding, our results seem to confirm Lerdahl and Jackendoff's generative theory, which declares that segmentation is a necessary step to the understanding of homophonic music.

Seventh, thematic recognition in polyphony appears not to be dependent upon mechanisms of segmentation. This result does not suggest that segmentation is not necessary for the complete musical understanding in polyphony, yet that mechanisms of segmentation do not act on thematic recognition, and consequently, do not play so important role in polyphony as they play in homophony. The result thus suggests that Lerdahl and Jackendoff's generative theory may not be directly applicable to polyphony.

Finally, mechanisms of thematic reinforcement seem to facilitate thematic recognition in polyphony. The mechanisms of thematic reinforcement might be activated under two circumstances. They might be activated by reinforcement in loudness performers yield to notes of theme in polyphonic music. Indeed, as mentioned in section 2.4.6, Kirkpatrick (1984) suggests that pianists reinforce notes of theme in polyphonic music by playing them louder. Alternatively, thematic reinforcement might be activated by memory mechanisms in the brain. Indeed, based on experiments with interleaved melodies, Dowling (1973) (section 2.4.1) claims that listeners are able to recognize a familiar target melody amid overlapping voices if the target is prespecified. Thus, for example, by memorizing the theme of a fugue, which occurs unaccompanied at the beginning of the fugue, listeners should be able to identify its instances whenever they occur throughout the fugue.

6.3 Further work

We outline here some directions for further work. First, classifications of the unsupervised model could be better analysed by comparing cases of instances of theme occurring unaccompanied and accompanied by a varied number of voices. The classifications could thus support and extend Huron's (1989) work (section 2.4.6), which points out that the more voices sounding simultaneously, the greater the difficulty listeners have in perceiving voice entrances.

Second, as presented in section 2.4.6, Huron (1993) has also claimed that perceptual segregation of voices is enhanced when onsets are asynchronous. Again, studies on classifications of the unsupervised model could be carried out in order to verify whether the model performs actually better on passages which contain lower amounts of onset synchrony.

Third, results of the connectionist model point out directions for research in experimental psychology. Experiments in cognitive psychology could be carried out, for instance, to investigate

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*, *5*(1), 1–30.

Bharucha, J. J. (1991). Pitch, harmony, and neural nets: a psychological perspective. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 84–99. The MIT Press, Cambridge, MA.

Bharucha, J. J., & Todd, P. M. (1991). Modeling the perception of tonal structure with neural nets. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 128–137. The MIT Press, Cambridge, MA.

Carpenter, G. A., & Grossberg, S. (1987). ART2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, *26*(23), 4919–4930.

Carpenter, G. A., & Grossberg, S. (1990). ART3: hierarchical search using chemical transmitters in self-organizing pattern recognition architectures. *Neural Networks*, *3*, 129–152.

Carpinteiro, O. A. S. (1993). The grouping rules of the Lerdahl and Jackendoff's theory and polyphony. Tech. rep. CSRP 300, School of Cognitive and Computing Sciences — University of Sussex, Falmer, UK. The Sixth White House Papers.

Carpinteiro, O. A. S. (1995). A neural model to segment musical pieces. In Miranda, E. R. (Ed.), *Proceedings of the Second Brazilian Symposium on Computer Music*, Fifteenth Congress of the Brazilian Computer Society, pp. 114–120. Brazilian Computer Society.

Carpinteiro, O. A. S., & Barrow, H. G. (1996). A self-organizing map model for sequence classification. Tech. rep. CSRP 424, School of Cognitive and Computing Sciences — University of Sussex, Falmer, UK.

Chappell, G. J., & Taylor, J. G. (1993). The temporal Kohonen map. *Neural Networks*, *6*, 441–445.

Cohen, M. A., & Grossberg, S. (1987). Masking fields: a massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. *Applied Optics*, *26*(10), 1866–1891.

Cole, W. (1970). *The Form of Music* (Second edition). The Associated Board of the Royal Schools of Music, London.

Deliege, I. (1987). Grouping conditions in listening to music: an approach to Lerdahl & Jackendoff's grouping preference rules. *Music Perception*, *4*(4), 325–360.

Desain, P., & Honing, H. (1991). The quantization of musical time: a connectionist approach. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 150–167. The MIT Press, Cambridge, MA.

Deutsch, D. (1982). Grouping mechanisms in music. In Deutsch, D. (Ed.), *The Psychology of Music*, chap. 4, pp. 99–134. Academic Press, London.

Deutsch, D. (1986). Auditory pattern recognition. In Boff, K. R., Kaufman, L., & Thomas, J. P. (Eds.), *Cognitive Processes and Performance*, Vol. 2 of *Handbook of Perception and Human Performance*, chap. 32, pp. 1–49. John Wiley & Sons, New York.

Dibben, N. (1994). The cognitive reality of hierarchic structure in tonal and atonal music. *Music Perception*, *12*(1), 1–25.

Dowling, W. J. (1971). Recognition of inversions of melodies and melodic contours. *Perception & Psychophysics*, *9*(3(B)), 348–349.

Dowling, W. J. (1972). Recognition of melodic transformations: inversion, retrograde, and retrograde inversion. *Perception & Psychophysics*, *12*(5), 417–421.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, *5*, 322–337.

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies. *Psychological Review*, *85*(4), 341–354.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, *1*(1), 30–49.

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, *49*(2(II)), 524–531.

Dowling, W. J., Lung, K. M., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics*, *41*(6), 642–656.

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception*, *10*(3), 343–378.

Edworthy, J. (1985). Melodic contour and musical structure. In Howell, P., Cross, I., & West, R. (Eds.), *Musical Structure and Cognition*, pp. 169–188. Academic Press, London.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Everitt, B. S. (1993). *Cluster Analysis* (Third edition). Edward Arnold, London.

Everitt, B. S., & Dunn, G. (1991). *Applied Multivariate Data Analysis*. Edward Arnold, London.

Fahlman, S. E. (1988). An empirical study of learning speed in back-propagation networks. Tech. rep. CMU-CS-88-162, School of Computer Science — Carnegie Mellon University, Pittsburgh, PA.

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 analyses of auditory rhythm patterns (part I). *Scandinavian Journal of Psychology*, *14*, 138–160.

Gallun, E., & Reisberg, D. (1995). On the perception of interleaved melodies. *Music Perception*, *12*(4), 387–398.

Garner, W. R., & Gottwald, R. L. (1968). The perception and learning of temporal patterns. *The Quaterly Journal of Experimental Psychology*, *20*(2), 97–109.

Ghahramani, Z., & Allen, R. B. (1991). Temporal processing with connectionist networks. In *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2, pp. 541–546.

Gjerdingen, R. O. (1990). Categorization of musical patterns by self-organizing neuronlike networks. *Music Perception*, *7*(4), 339–370.

Gjerdingen, R. O. (1991). Using connectionist models to explore complex musical patterns. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 138–149. The MIT Press, Cambridge, MA.

Gjerdingen, R. O. (1992). Learning syntactically significant temporal patterns of chords: a masking field embedded in an ART3 architecture. *Neural Networks*, *5*, 551–564.

Handel, S. (1974). Perceiving melodic and rhythmic auditory patterns. *Journal of Experimental Psychology*, *103*(5), 922–933.

Huron, D. (1989). Voice denumerability in polyphonic music of homogeneous timbres. *Music Perception*, *6*(4), 361–382.

Huron, D. (1993). Note-onset asynchrony in J. S. Bach's two-part inventions. *Music Perception*, *10*(4), 435–444.

Jackendoff, R. (1991). Musical parsing and musical affect. *Music Perception*, *9*(2), 199–230.

Jackendoff, R., & Lerdahl, F. (1981). Generative music theory and its relation to psychology. *Journal of Music Theory*, *25*(1), 45–90.

James, D. L., & Miikkulainen, R. (1995). SARDNET: a self-organizing feature map for sequences. In Tesauro, G., Touretzky, D. S., & Leen, T. K. (Eds.), *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 7. Morgan Kaufmann.

Jones, M. R. (1987). Dynamic pattern structure in music: recent theory and research. *Perception & Psychophysics*, *41*(6), 621–634.

Kangas, J. (1991). Time-dependent self-organizing maps for speech recognition. In Kohonen, T., Mäkisara, K., Simula, O., & Kangas, J. (Eds.), *Proceedings of the International Conference on Artificial Neural Networks*, pp. 1591–1594. Elsevier Science Publishers.

Kangas, J. (1994). *On the Analysis of Pattern Sequences by Self-Organizing Maps*. Ph.D. thesis, Laboratory of Computer and Information Science, Helsinki University of Technology, Rakentajanaukio 2 C, SF-02150, Finland.

Kirkpatrick, R. (1984). *Interpreting Bach's Well-Tempered Clavier, A Performer's Discourse of Method*. Yale University Press, London, UK.

Kitson, C. H. (1924). *The Art of Counterpoint*. Da Capo Press.

Kohonen, T. (1988). An introduction to neural computing. *Neural Networks*, *1*, 3–16.

Kohonen, T. (1989). *Self-Organization and Associative Memory* (Third edition). Springer-Verlag, Berlin.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Laden, B., & Keefe, D. H. (1991). The representation of pitch in a neural net model of chord classification. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 64–83. The MIT Press, Cambridge, MA.

Lashley, K. S. (1954). Dynamic processes in perception. In *Brain Mechanisms and Consciousness*, pp. 422–443. Blackwell Scientific Publications, Oxford.

Leman, M. (1991). The ontogenesis of tonal semantics: results of a computer study. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 100–127. The MIT Press, Cambridge, MA.

Lerdahl, F., & Jackendoff, R. (1983a). *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, MA.

Lerdahl, F., & Jackendoff, R. (1983b). An overview of hierarchical structure in music. *Music Perception*, *1*(2), 229–252.

Lewis, J. P. (1989). Algorithms for music composition by neural nets: improved CBR paradigms. In *Proceedings of the International Computer Music Conference*, pp. 180–183. Computer Music Association.

Lewis, J. P. (1991). Creation by refinement and the problem of algorithmic music composition. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 212–228. The MIT Press, Cambridge, MA.

Lo, Z., & Bavarian, B. (1991). Improved rate of convergence in Kohonen neural network. In *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2, pp. 201–206.

Lo, Z., Fujita, M., & Bavarian, B. (1991). Analysis of neighborhood interaction in Kohonen neural networks. In *Proceedings of the Fifth International Parallel Processing Symposium*, pp. 246–249.

Marr, D. (1982). *Vision*. W. H. Freeman, New York.

Mozer, M. C. (1989). A focused backpropagation algorithm for temporal pattern recognition. *Complex Systems*, *3*, 349–381.

Mozer, M. C. (1991). Connectionist music composition based on melodic, stylistic, and psychophysical constraints. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 195–211. The MIT Press, Cambridge, MA.

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.

Page, M. P. A. (1994). Modelling the perception of musical sequences with self-organizing neural networks. *Connection Science*, *6*(2&3), 223–246.

Palmer, C., & Holleran, S. (1994). Harmonic, melodic, and frequency height influences in the perception of multivoiced music. *Perception & Psychophysics*, *56*(3), 301–312.

Peretz, I. (1990). Processing of local and global musical information by unilateral brain-damaged patients. *Brain*, *113*, 1185–1205.

Peretz, I., & Babaï, M. (1992). The role of contour and intervals in the recognition of melody parts: evidence from cerebral asymmetries in musicians. *Neuropsychologia*, *30*(3), 277–292.

Peretz, I., & Morais, J. (1987). Analytic processing in the classification of melodies as same or different. *Neuropsychologia*, *25*(4), 645–652.

Peretz, I., & Morais, J. (1988). Determinants of laterality for music: towards an information processing account. In Hugdahl, K. (Ed.), *Handbook of Dichotic Listening: Theory, Methods and Research*, chap. 11, pp. 323–358. John Wiley & Sons, New York.

Povel, D., & Egmond, R. (1993). The function of accompanying chords in the recognition of melodic fragments. *Music Perception*, *11*(2), 101–115.

Prout, E. (1890). *Counterpoint: Strict and Free*. Augener, London.

Prout, E. (1891). *Fugue*. Augener, London.

Prout, E. (1969). *Double Counterpoint and Canon*. Greenwood Press, New York.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986a). A general framework for parallel distributed processing. In Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.), *Parallel Distributed Processing*, Vol. 1, chap. 2, pp. 45–76. The MIT Press, Cambridge, MA.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). Learning internal representations by error propagation. In Rumelhart, D. E., McClelland, J. L., & the PDP Research Group (Eds.), *Parallel Distributed Processing*, Vol. 1, chap. 8, pp. 318–362. The MIT Press, Cambridge, MA.

Rumelhart, D. E., & McClelland, J. L. (1988). On learning the past tenses of english verbs. In McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (Eds.), *Parallel Distributed Processing*, Vol. 2, chap. 18, pp. 216–271. The MIT Press, Cambridge, MA.

Sano, H., & Jenkins, B. K. (1991). A neural network model for pitch perception. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 42–53. The MIT Press, Cambridge, MA.

Scarborough, D. L., Miller, B. O., & Jones, J. A. (1991). Connectionist models for tonal analysis. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 54–63. The MIT Press, Cambridge, MA.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce english text. *Complex Systems*, *1*, 145–168.

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. *Psychological Review*, *76*, 1–15.